HDRUK
Health Data Research UK

Data Utility Framework

November 2020

HDRUK
Health Data Research UK

Overview

HDR UK has a mission to unite the UK's health data to enable discoveries that improve people's lives. This is being approached by bringing the data together through the work of the UK Health Data Research Alliance and Innovation Gateway (Uniting the Data), by making the data more useful for science and innovation through the work of the Health Data Research Hubs and the development of tools and approaches (Improving the Data) and using the data for specific purposes with the potential to transform people's lives (Using the Data).

The work to Unite and Improve the Data is important, but it is only meaningful if it leads to greater quantity and quality of research and the generation of more meaningful insights. There is much discussion about the importance of data curation, and how much should be invested in this area. However, there is little clarity regarding exactly what activities are meant by 'curation' and in order to ensure that resources are effectively targeted, this should be informed by user needs. For example,

- A pharmaceutical company reviewing the effectiveness of a cancer treatment over a 10-year period requires linked primary and secondary data, with high levels of trust in the provenance and data quality management processes and follow-up of at least 10 years.
- A medical device regulator tracking adverse incidents associated with implantable devices requires detailed data on the devices and implantation procedure, linked device registry data with longitudinal patient information and consistency in coded records.

To achieve the ultimate aim of improving people's lives, HDR UK has developed a framework to articulate the potential usefulness of datasets for specific purposes. This will support users of data in the discovery and selection of datasets for their purposes, as well as providing an evidence base for identifying specific areas of activity for improvement to allow for wider use or greater insights.

This has been developed in consultation with users of data and data custodians from a range of sectors, through interviews, surveys and a green paper consultation process[1], working in partnership with MetadataWorks Ltd. The framework was further refined by being tested on approximately 50 datasets held by Health Data Research Hubs. The detail on how the framework was developed will be included in an upcoming academic publication.

Structure

The framework contains five categories, separated across a range of dimensions, each of which are qualitatively evaluated to describe the characteristics of a dataset. Each dimension has a progressive series

[1] Data Utility Green Paper: https://www.hdruk.ac.uk/wp-content/uploads/2020/08/200820-Data-Utility-Green-Paper-Consultation-Draft.pdf , p.4

of criteria, allowing for a rating from ‘Bronze’ to ‘Platinum’ for each, provided the minimum criteria is met. The purpose is not to achieve a ‘Platinum’ rating across all dimensions, but to enable a user to exclude datasets that would not meet a specific threshold based on their needs. The framework enables:

- data custodians to communicate the utility of their dataset, and improvements made in the data set
- users to identify datasets that meet the minimum requirements for their specific purpose
- System leaders and funders to identify where to invest in data quality improvements, and to evaluate what improvements have happened as a result of their investments

There is much more information available about a dataset (metadata) than what is captured in the framework, and the Innovation Gateway contains detailed metadata to allow a user to understand more about the dataset.

Next steps

The framework will be incorporated into the Innovation Gateway, with the aim to evaluate the utility of the majority of datasets on this resource. This would support the ability for users to filter out the datasets that would not meet their needs and identify those that would support it.

Further work will be required to refine the categories and wording, including ‘normalisation’ of the categories to establish an appropriate balance. An updated version will be shared in 2021.

The data utility framework will play a key role in evaluating the impact of the Health Data Research Hubs, an investment from the UK Life Sciences Industrial Strategy.

Appendix: Data Utility Framework v2, November 2020

The Data Utility Evaluation Matrix is a work in progress and will be refined following testing and feedback. Note that datasets which do not achieve Bronze in a category would be classified as “White”.

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 IG approval) and review the impact of each stage of data cleaning.

Category	Dimension	Definition	Bronze	Silver	Gold	Platinum
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 ISO, CQC, 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 IG 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 IG 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

Category	Dimension	Definition	Bronze	Silver	Gold	Platinum
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.

Appendix: Updates to the Data Utility Matrix

Data Documentation

A number of users queried the requirement on maintaining support personnel to assist with any questions on a dataset in order to achieve platinum scoring. We understand that a support team may be on hand in addition to data dictionaries and documentation already in place. As a result, it was decided to maintain the gold criteria within the platinum level, with the addition of a personnel support team.

Allowable Uses

Lots of feedback was received concerning the commercial use of a dataset assigned as a platinum score. It was highlighted that licensing agreements per dataset will determine whether the dataset can be available for commercial use, therefore potentially hindering a lot of datasets from achieving a platinum level.

A ‘filter’ is in the process of being developed on the HDR UK Innovation Gateway to distinguish commercially available datasets from others using the platinum rating as a guide. Therefore, the rating will remain the same and will be used as a means of distinguishing applicable datasets.

Pathway Coverage

There were several suggestions to amend the wording of the platinum score within the ‘Pathway Coverage’ dimension. The term ‘whole pathway of care’ was considered ambiguous in referring to particular care pathways, e.g. social care, primary care, community care etc. In order to maintain the breadth of the framework criteria, the platinum score has been changed to account for more than two tiers of coverage but remains general enough to incorporate multiple health care sectors.

Provenance

As it was indicated that ‘raw data’ is against the policy of ‘de-identification’, we have re-worded the platinum rating to specify the ability to view de-identified as part of Information Governance approval.

Length of Follow-Up

Users put forward that in some cases, an increased length of follow up does not necessarily denote additional value, e.g. COVID-19 datasets will naturally have a short follow-up compared to longitudinal datasets. However, in order for the scorings to simultaneously reflect a particular *use* of a dataset accurately, the time frames have remained the same.

Research Environment

After consultation with the community as part of the Green paper and follow up within the Central team, it was decided to remove the Research Environment dimension until further work is done to align our TRE principles with the dataset needs from users.

Time Lag

Users have flagged that that the very nature of ‘real-time data’ may not indicate a ‘platinum’ equivalent quality of the data contents. Similarly, it was also raised that the time lag is driven by a specification and therefore cannot be improved upon. As both of these points will be governed by the specific prescribed features of a dataset, the score definitions have been maintained.

Timeliness

Timeframes for each scoring level have reduced slightly for additional clarity and to incorporate any access timeframe less than 2 weeks.