

Data Utility Green Paper – Draft for Consultation

17th July 2020



Contents

Data Utility Green Paper – Draft for Consultation	1
Contents	2
Overall vision	3
Evaluating Data Utility	4
Data Utility Evaluation Matrix	4
Data Quality Tools	6
Metadata Improvement	7
Data standards.....	8
Next Steps.....	8
Appendix 1: Definitions	10
Data quality	10
Curation	11
Appendix 2: Data Utility Evaluation Matrix MVP, as at 14 August.....	12
Appendix 3: List of use cases from interviews	15

DRAFT

Overall vision

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.

To achieve the ultimate aim of improving people's lives, the "Improving the Data" strategy area within HDR UK is developing 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.

By March 2021, we aim to have a widely adopted framework for articulating the utility of the datasets discoverable through the Gateway. This would be supported by rich metadata, allowing users to search and filter for datasets that match their particular needs. We will also have collated a growing list of case studies and examples of data improvement for specific features or types of data, to inform future best practice in this area.

For the purposes of this document, the term 'data utility' is used to refer to the usefulness of a dataset for a given purpose, and 'data quality' is used in reference to a more narrow set of characteristics, such as completeness, accuracy and validity. This is because the elements that make a dataset useful, such as whether there is documentation available or whether it can be linked to another dataset, are much broader than the characteristics more commonly described as 'data quality'. These topics are explored further in Appendix 1: Definitions.

Evaluating Data Utility

Data Utility Evaluation Matrix

Effective scalable and transferable communication of the 'usefulness' of dataset requires a generalisable framework for rating and communicating how useful a dataset is *for a given purpose*, as well as the ability to assess and compare datasets from different sources at scale. This approach has led to the Data Utility Evaluation Matrix.

In partnership with MetadataWorks, the matrix was developed using input from users of data on their needs and expectations for data. This includes workshops, interviews and surveys with scientists, innovators and researchers including large Life Science / Pharma industry, Large AI and Technology, SMEs, consultancies, NHS and data custodians, academic researchers and regulators. Between April 2019 and June 2020, HDR UK conducted interviews with representatives from each of these sectors¹ to learn i) what they use health data for ii) what dimensions of data quality matters to them for these uses, and iii) what level of quality is required to meet their purpose. Users of data have consistently fed back that it is difficult to understand in advance whether a given dataset would answer their specific question, leading to lengthy access request processes, only to discover months later that the dataset does not meet their requirements. The significant advances in metadata discovery through the Innovation Gateway, as a result of work across the Alliance and the Hubs, has gone some way to addressing this issue, and the Data Utility Evaluation Matrix would help this further.

The interviews highlighted that different users have different requirements from a dataset for different purposes, 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.

Specific requirements for data vary based on the individual needs, which are summarised through the use cases identified in Appendix 3. Five categories, separated into twenty elements, were identified as being of most interest to users. The framework sets out these elements of data utility, and a qualitative articulation that scores a dataset through a progressive process from Bronze to Platinum.

A detailed report on the project will be released in the coming months.

¹ Industry organisations interviewed: Allergan, ABPI, Bayer, C4X, Covance, Illumina, IQVIA, Qiagen, AZ, BMS, J&J, Lilly, MSD, Roche, Sanofi, UCB, IBM, ABHI, Babylon, Medtronic, Albion, Benevolent AI, Decipher, Ixico, Neem, Owkin, RAIR, Repositiv, McKinsey, Roche, Novartis, AstraZeneca, Google, Microsoft, ZOE, Deloitte

The framework enables:

- data custodians to communicate the utility of their dataset, and improvements made in the data set (Fig 1 below)
- users to identify datasets that meet the minimum requirements for their specific purpose (Fig 2 below)
- 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

The intention is for the framework to be a ‘living’ tool, which will therefore be iteratively modified with user feedback and evaluation of metrics.

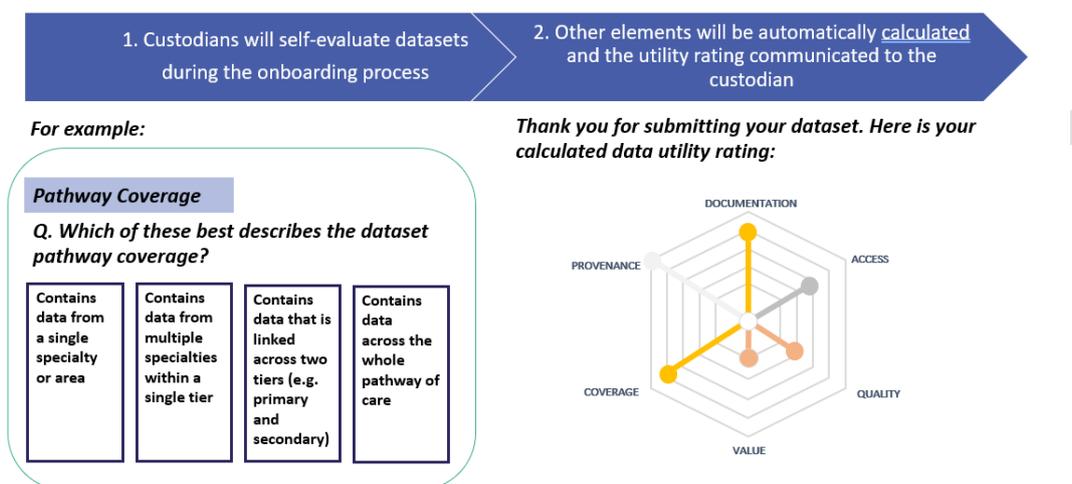


Figure 1: Example process by which a custodian could generate a data utility score on the Gateway

In addition to the ability to self-evaluate for a given dataset, and communicate on the Gateway the uses for which that dataset might be best suited, the adoption of the framework will also support the gathering of information about the specific processes that lead to improvements in the utility scores for particular dataset types of use cases. As the Hubs have a specific mandate to improve data, work that they do to impact the characterisation of a dataset on the matrix can be captured and a basic input/output score identified, which will inform future activities and investment.

Populating the Gateway with the outputs of evaluations would allow users to search, filter or otherwise interact with datasets on the Gateway to help identify the datasets that would be most useful for their purposes. This could be implemented through a decision-support tool or a “wizard”, as shown in Figure 3.

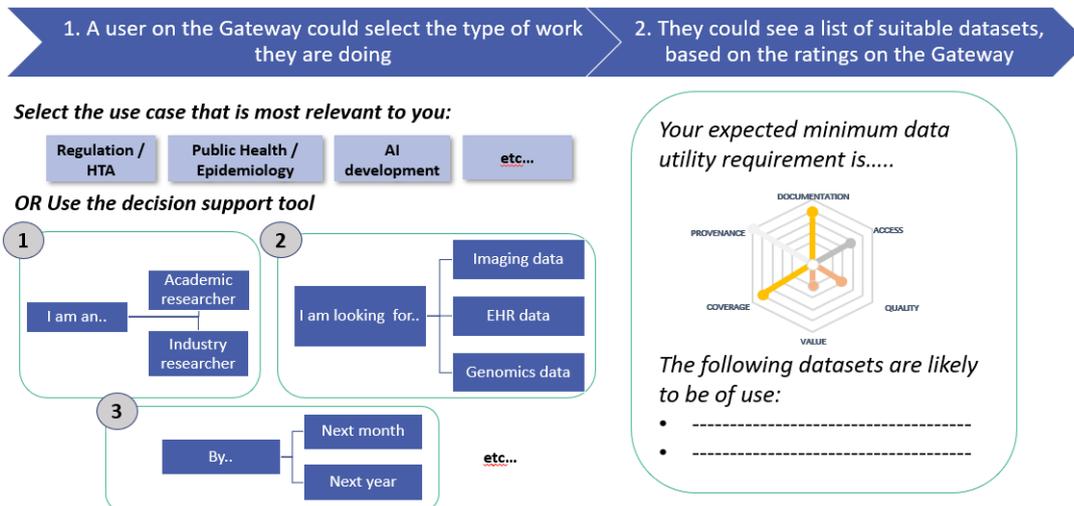


Figure 2: Example process by which a user may identify suitable datasets on the Gateway

Research identified several elements which they would wish to understand in advance of requesting access, but would not necessarily lend itself to a progressive, medallion-based categorisation. For example, the purpose for which a dataset was collected (a clinical trial, routine care, commissioning, amongst others) will help to indicate whether it would be suitable for a particular use-case, but no one use is necessarily ‘better’ than others, and so it would not be suitable for a ranked system. These elements have been added to the latest metadata specification, and will be represented on the Gateway, but not included in the matrix.

Questions for consultation:

- Looking at the current version of the matrix in Appendix 2, are there any changes or adjustments you would make to this?

Data Quality Tools

A subset of the data utility rating is the presentation of “classic” technical data quality dimensions (Completeness, Uniqueness, Accuracy, Validity, Timeliness and Consistency as outlined by the Data Management Authority²). In partnership with Inspirata, HDR UK have tested over 20 open and open-source tools to evaluate specific types of datasets using the listed technical dimensions of data quality. All tools evaluated through this process will be made available through the Gateway, along with the output of the initial evaluation process, for use by data custodians and users across the UK.

A shortlist of tools were evaluated in more detail, including testing against a reference dataset and by specific data custodians from the Alliance. This detailed evaluation will provide a balanced view on the most suitable tools to be used by data custodians, including the Hubs in advance of Milestone 2 (below).

² DAMA UK Working Group: The Six Primary Dimensions for Data Quality Assessment, 2015

This will provide tools for increased reliability and consistency of dataset evaluation and profiling, which will contribute to the evaluation of dataset utility and identification of areas for improvement activity.

A detailed report on the project will be released in the coming months. In addition, the tools will be hosted on the Gateway tools repository, and the project outputs communicated through a webinar scheduled for 24 September.

Questions for consultation:

- How should the outputs of this project (findings from the evaluation, sample outputs, instructions) be more widely disseminated?
- Beyond hosting the metadata for the tools on the Gateway, how else should the recommended tools be made available to custodians?

Metadata Improvement

A significant amount of work in “Improving the Data” to date has focused on metadata development. Much of the early feedback from data users centred on not knowing what data was available for research, and not knowing what it was or how it might be used. For those reasons, initial work focused on collecting detailed metadata for datasets that are held across the Health Data Research Alliance and Hubs. In November 2019, we developed a “Minimum Viable Product” [metadata specification](#) in consultation with users and the HDR UK community. This allowed for the discovery of over 400 datasets across the Alliance on the initial release of the Innovation Gateway, supported by metadata that outlined the description of the datasets, steps for access and much more. Prior to this point, metadata was not available in a consistent format across the UK, with the closest equivalent being Health Data Finder, which contained metadata on 16 datasets held by data custodians in England. Since launching in January, the depth of metadata available on the Gateway has improved, and by May 34% of datasets contained technical metadata (showing the individual fields of a dataset), up from 15% in January.

The initial specification released in November 2019 has been refined based on feedback from users of the Gateway and custodians. This has led to additional fields being included, which support more in-depth search, and constraints imposed on the allowable entries for specific fields. This makes it easier to search, sort and filter the returns when searching on the Gateway. Following a lengthy consultation, which involved input from across the health data community, the updated metadata specification was released in July 2020, and is available as a markdown or JSON schema on [GitHub](#) to allow for wider input and more rapid adoption into local systems. As the Gateway develops the functionality to support federated searching, the specification can be adopted locally by individual metadata catalogues, rather than being included in the HDR UK metadata catalogue.

To achieve wider adoption (supporting federation at a greater scale, including internationally), we are seeking endorsement of this specification by the British Standards Institute, as well as the Office of National Statistics and NHSX.

Further improvements in the metadata involve collecting a more granular level of detail on each dataset, which will provide more insight on how datasets may be linked or compared, and collecting dataset profiling information. The collection of rich metadata will also support the activities of the data utility work, described below.

Questions for consultation:

- Are there any alterations you would suggest for the metadata specification for future versions?
- How might the metadata specification be more widely adopted or used?

Data standards

Working alongside many national bodies and organisations in the UK, HDR UK have an important role in the refinement and adoption of data standards to support research. While not mandating specific standards for adoption, in June 2020, HDR UK issued a series of [principles for data standards](#) that should be used by organisations in this domain. These principles were developed in consultation with over 50 data officers from over 30 organisations.

HDR UK have also played a major role in collection and communication of standards in response to the COVID-19 pandemic. This has initially involved the collection of data elements from 25 current studies, in order to identify common data elements with the aim of informing future study design and onward linkage. The pre-print paper is available [here](#). This work could be expanded into other national efforts, as well as being used as a model for study design in other conditions.

Future development for this work should be for the understanding of how the data principles could be put into practice and the identification of specific standards which may be recommended and the pragmatic aspects to be considered regarding future implementation of such standards.

Next Steps

The Innovation Gateway has already greatly improved the discoverability of datasets, and now provides metadata for 455 health datasets, with 5,000 searches a month. The access management function on the Gateway has already facilitated 58 access requests and is being integrated into data custodians' systems by October 2020. The data utility framework will be fully integrated in to the Gateway by October 2020 and will help custodians make data more useful, users find the right data to meet their needs and system leaders see improvements in data and identify opportunities for further development.

The Health Data Research Hubs are improving data utility at both a local and national level. They have already improved metadata of datasets in the Innovation gateway, created new datasets, curated and linked existing datasets, improved data by adding new or improved data feeds, worked with wider stakeholders to develop near real-time datasets and developing/integrating new platforms including TRE's. The framework will be used to evaluate the utility improvements being achieved by the hubs, to share lessons on which improvement approaches are most effective and to identify data utility most valued by their users.

A key date for reviewing the improvements in the data is March 2021, when the Hubs are required to deliver against Milestone 2. This requires evidence that data has been improved and is being used to generate insights.

The approach should be broadened to include other work across the HDR UK infrastructure, including activities in Phenomics, which can make data more useful through standardised approaches to capturing and communicating conditions, as well as use of AI in the Applied Analytics national priority, which can use tools and algorithms to derive greater insight for researchers.

We will work closely with system leaders to use the framework and insights from users, the Gateway and Hubs to guide decision making around further investments in data utility.

DRAFT

Appendix 1: Definitions

The terms ‘data quality’ and ‘data curation’ are often loosely used when defining or describing how useful a dataset is for a particular purpose and the activities required to somehow improve the data from its original state. These terms are interpreted in different ways by different people, so we propose alternatives below for the purposes of HDR UK discussion and documentation.

Data quality

Data quality has many different definitions, most broadly relating to how well a dataset is fit for a particular purpose. In some cases this is limited to specific “technical” dimensions such as those outlined by DAMA³ or the [global standard for data quality \(ISO8000\)](#). Others use this term more broadly. Through the Design and Dialogue phase of the Digital Innovation Hubs programme, HDR UK engaged with many users of data from the NHS, academia and industry. Many of those consulted made general reference to “low quality data”. However, on further investigation, the substance of the complaint could relate to the dataset itself, the quality of the description or metadata provided, the length of time taken to gain access to the dataset or even the fact that the data desired by the user was not being collected.

As stated, many organisations therefore use a more pragmatic definition of data quality, describing high quality data as being that which is “fit for purpose”. For example, the [ISO9000:2015](#) definition of quality describes data quality as being “the degree to which a set of characteristics of data fulfils requirements”. The challenge posed by this definition is that the specific purpose must be known in advance for the data to be described as being “fit for purpose”.

Alternative term: “data utility”

To circumvent the challenges above, we aim to describe factors relating to the concept of data utility, rather than data quality. This is evidently a much wider term, and unlikely to be confused with the specific definition of technical ‘data quality’ dimensions. Furthermore, it clearly relates the description to the requirements of users. The work in the development of the standard data utility matrix, described below, is an attempt to develop a framework which can be applied across multiple datasets from different centres and for different purposes, but provided the broad scope of purpose is known, the principles should remain sound.

This definition of data utility therefore allows for a view across a broad range of areas. Making data more ‘useful’ could mean removing errors in a dataset, adding data through derived columns or annotations, increasing the depth or scale of the dataset by gathering or linking additional data or simply by improving the use restrictions.

³ DAMA UK Working Group: The Six Primary Dimensions for Data Quality Assessment, 2015

Curation

Data curation is a term often used to refer to managing or improving data. This approach may conjure different ideas to different people, for example from an algorithm that removes specific errors in a dataset through to use of a highly trained member of staff manually comparing electronic health records to written notes or free text. Across the many definitions of 'curation', the overarching theme is that data is actively managed through the whole data life cycle from its collection through to use

Alternative term: "Improving the data"

Given the multiple personal interpretations of data curation, we speak more broadly of "improving the data". This allows for a direct comparison with data utility, where improving the data means making data more useful (the specifics of which may therefore differ by dataset and user etc).

Within this concept there are many possible tasks, with an expected diminishing marginal return for the improvements to be gained, as outlined below:

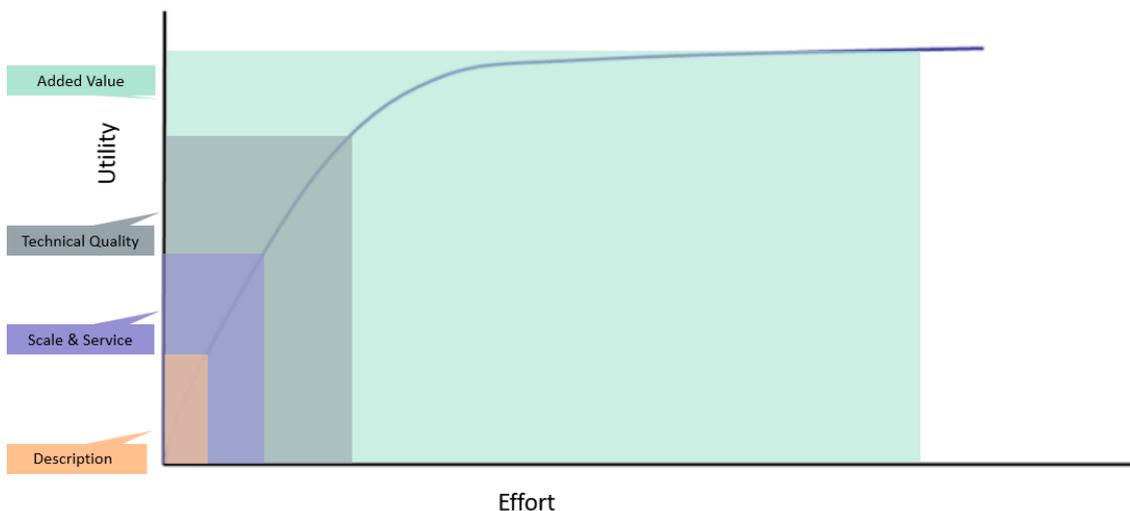


Figure 3: Chart showing diminishing returns for effort expended to increase utility of a dataset

Given this split, it is likely that many different organisations and entities can support the improvement of data, for different user needs.

Given this conceptual model, it is likely that many different organisations and entities can support the improvement of data, for different user needs although the optimal distribution of resource to improve datasets will differ between them.

Appendix 2: Data Utility Evaluation Matrix MVP, as at 14 August

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	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	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 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

Category	Dimension	Definition	Bronze	Silver	Gold	Platinum
Technical Quality	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
	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			
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

Category	Dimension	Definition	Bronze	Silver	Gold	Platinum
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 required
	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.

Appendix 3: List of use cases from interviews

Use Case Category	Example	Data Requirements	Selected Elements in Matrix Profile
Service improvement / pathway redesign	Using EHR extract to develop an orthopaedics OP clinic and staffing model	<ul style="list-style-type: none"> Data should be coded to defined standards to ensure results are comparable Longitudinal (full pathway) required to support planning 	<ul style="list-style-type: none"> Data Dictionary – at least Bronze Data Model – at least Silver Provenance – at least Silver Pathway Coverage – at least Gold
	Integrating data returns from multiple trusts for production of commissioning datasets	<ul style="list-style-type: none"> Consistent data model and terminology fields codified to at least a national standard to enable integration of multiple datasets A clear definition of the content and source of each field 	<ul style="list-style-type: none"> Data Model - at least Gold Data Dictionary – at least Bronze Provenance - at least Gold
Treatment Evaluation	Pharma company reviewing the effectiveness of a cancer treatment over a 10 year period	<ul style="list-style-type: none"> Linked primary & secondary data, including outcomes and labs High levels of trust in the source and accuracy Duration / follow up of at least 10 years Data sourced from real world evidence sources, such a hospital episode statistics, in order to establish the nature of medical conditions in the general context (outside of clinical trials) Available for use by commercial entity 	<ul style="list-style-type: none"> Provenance – at least Gold Data Quality Management Process – at least Gold Pathway Coverage - at least Gold Length of Follow Up – Platinum Allowable Uses – at least Gold
	Assessing the level of effectiveness of medicines to control results from patient's blood tests	<ul style="list-style-type: none"> Requirement to identify and link a patient prescriptions and laboratory results overtime Detailed bridging notes covering the dynamic dimensions of time series data 	<ul style="list-style-type: none"> Pathway Coverage – at least Silver Availability of additional documentation and support - at least Silver
	Administration of 'managed access' regulatory mechanisms for medicinal products	<ul style="list-style-type: none"> A traceable, trusted and defensible source of data – “Regulation grade” Requirement to establish the extent to which the dataset to be considered "generalisable to the wider population" Trusted and robust methods for data collection and verification 	<ul style="list-style-type: none"> Pathway Coverage – at least Silver Provenance - at least Gold Data Quality Management Process - at least Gold

Use Case Category	Example	Data Requirements	Selected Elements in Matrix Profile
Patient Safety	Medical device regulator tracking adverse incidents associated with implantable devices	<ul style="list-style-type: none"> Detailed data of the medical devices (manufacturer, model, preferably a unique device identifier) and the implantation procedure. Linked device registry data with longitudinal patient information Consistency in clinically recorded measures. Coded records, rather than free text. 	<ul style="list-style-type: none"> Data Dictionary – at least Bronze Provenance – at least Silver Pathway Coverage – at least Gold Data Model - at least Gold
Population Health	Academic research on the accuracy of reported alcohol consumption in clinical records	<ul style="list-style-type: none"> Linked datasets to enable comparison of the same patients in different clinical records A detailed description of data elements and their source Low time lag due to grant time frame Clear information governance procedures to enable timely access to required data 	<ul style="list-style-type: none"> Pathway Coverage – at least Silver Provenance – at least Silver Data Dictionary – Platinum Access & Approvals – at least Bronze Timeliness - at least Silver
	Commissioned research of the number of patients with atrial fibrillation that experienced a stroke in last 12 months, the outcome of the incident, and how this compares with trends over time	<ul style="list-style-type: none"> Quickly accessible requirement to meet the timeframe of the research commissioning organisation Requirement to link data at either the London CCG level or at the level of a single GP practise A detailed description of data elements 	<ul style="list-style-type: none"> Timeliness – Platinum Linkages – at least Bronze Data Dictionary – at least Bronze
Operational efficiency for health systems	Developing a triage solution system to automatically review incoming referrals for a hospital, and automatically accept, reject or redirect, based on an NLP machine learning algorithm.	<ul style="list-style-type: none"> Dataset of unstructured referral letter data alongside details of the outcome of the letter (accepted, rejected, redirected). Permission to use the data for innovative solutions Clear expectations of the IG requirements in order to plan and manage commercial risks 	<ul style="list-style-type: none"> Data Enrichment – at least Bronze Allowable Uses – at least Gold Access & Approvals - Platinum
	Using an annotated imaging dataset to refine a machine learning computer aided diagnosis algorithm	<ul style="list-style-type: none"> High quality imaging dataset Rapid access due to short, VC timelines Annotated, well curated dataset Available for use by commercial entity 	<ul style="list-style-type: none"> Allowable Uses – at least Gold Format – at least Silver Timeliness – at least Platinum Data Enrichments – at least Bronze