

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"	<i>Set initial figures for this</i>			
	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.