Health Data Research UK

Data Quality and Standards Strategy

Green Paper for Consultation September 2019

Contents

Executive Summary

Establishing common standards for healthcare data and metadata is a fundamental requirement for uniting the UK's health data to make discoveries that improve people's lives. This Green Paper sets out how this challenging requirement could be achieved as part of the one institute programme that includes developing a network of data owners to improve access to datasets (Alliance), an infrastructure layer providing greater dataset visibility (Gateway) and centres of excellence providing services including data curation (hubs) and research (substantive sites).

Common standards and improved data quality will allow interoperability of data between organisations, enabling collation of datasets and facilitating a step change in scale and efficiency of health data research through collaboration, federated analytics, and development of scalable and reusable research and analytic pipelines and tools.

Health Data Research UK (HDR UK) has an important role in leading the development of best practice. The intention is to provide position papers and recommendations, developed conjunction with the health data community, that would encourage behaviours to improve data quality and usability. HDR UK can mandate adoption through its funded activities such as the hubs, Innovation Gateway and substantive sites. HDR UK will also work with member of UK Health Data Research Alliance, including organisations such as NHSX and its equivalents in the devolved nations to shape system-wide mandatory requirements.

Two enabling mechanisms are proposed:

1. Establishing the role of chief **data officers** throughout the UK Health Data Research Alliance and wider network.
2. Requiring organisations to collect, store and provide their datasets according to defined **open standards** through the Alliance and other HDR UK infrastructure such as DIHs and substantive sites.

We are seeking your input on these and a range of areas set out in the document. At the end of each section you will find the questions we would like you to help us address and to gauge your level of agreement. Please complete the online survey to provide your answers and feedback https://www.surveymonkey.co.uk/r/WNQ8ZDK. The relevant text from this document is repeated in the survey so you can read it and respond directly online.

Our intention is to use the outcomes from this consultation to help shape the strategy and prioritise the work ahead. Together we can realise the benefits of health data research to patients and the nation more readily. It will also support the UK's position as a global leader in the field of health data research.

Overview

Background

The overall aim of HDR UK is to unite the UK's health data to enable discoveries that improve people's lives such that every health and care interaction and research endeavour will be enhanced by access to large scale data and advanced analytics.

In addition, HDR UK aims to support FAIR data principles in order to maximise value from health datasets for secondary use. In order to achieve these aims, governance and platform infrastructure, dataset availability and analytic capability is being developed through the HDR UK Innovation Gateway (the Gateway), UK Heath Data Research Alliance (the Alliance), Hubs and substantive sites (SS).

An essential enabling component to facilitate the function and interaction of these elements is provision of high-quality datasets with common (data and metadata) standards to allow machine readability and semantic interoperability; these areas are within the remit of Data Quality and Standards (DQS) workstream, led by the Chief Clinical Data Officer (CCDO).

Principles

In addition to alignment with the principles encompassed by FAIR (making data findable, accessible, interoperable and reusable), the HDR UK approach to data quality and standards is proposed to be guided by the following principles:

- alignment with existing and proposed national (e.g., NHSX, NHS Digital, PRSB and equivalents in devolved nations) and international, health data standards where possible*
- leverage existing developments in this area by other groups, for example US ONC, GEL etc
- optimise both research and clinical benefits of data use and interoperability by having common standards for research and clinical / operational value
- use existing open standards where possible, including planned alignment with internationally supported standards such as HL7 FHIRv4
- minimise additional resource required to manipulate datasets through HDR UK following the general principle of 'without special effort', via use of open standards and APIs.

**There are already several groups (CQC, NHSD, NHSE, NHSI, NHSx, PRSB etc) with data standards remits (in addition to consideration of international standards and interoperability) and it will be important to ensure alignment and reduce fragmentation of recommendations where possible. Alignment with International standards should also be considered.*

- Do you agree with these principles?
- Is there anything else that you would add?
- Which additional organisations do you think are most important to ensure alignment with?

Scope and aims

Whilst quality and standards regarding all health data relevant to HDR UK is potentially within scope of this strategy, activity will be phased.

The ultimate aim is that all health data* relevant to HDRUK purposes from organisations within the United Kingdom, including Alliance members, organisations associated with the One HDRUK Institute, and NHS trusts and organisations, will be of high quality (see later section) and available for healthcare research and innovation purposes. Our ambition is that:

- All datasets will have both dataset level and variable level open standard metadata and have quality metrics available.
- All data will be available according to well-described open standard formats.
- Use of data will be associated with provenance tracking to provide information regarding its use and value, linking to benefits for NHS patients.
- All associated organisations will have established and aligned data strategies and data quality strategies.

** For purposes of the current document, health data refers to data generated by, or associated with, health care provision and at this stage does not include broader social, environmental, etc data, although it is recognised that these data may also be highly relevant to health.*

- Do you agree with these broad ambitions?
- Is there anything you would add (or remove)?
- Do you agree with the scope of health data as defined above for this strategy?

Phase 1

Initial focus is to support and enable the DIH programme by providing first wave of quality and standards guidance for datasets generated from the Alliance and Hubs and how these are discoverable through the Gateway.

The first milestone is to have an operational offer for researchers and innovators through the Gateway Minimum Viable Product (MVP), with datasets from the Alliance and Hubs discoverable by January 2020.

Phase 2+

This will extend the focus through additional Alliance members, Hubs, specific research study data sets, data from external sources, and increasingly, routine NHS data from many providers, in addition to focus on additional 'novel' data sources such as patient generated health data and data from Internet of Things (IoT)/streaming devices. Duration TBC, suggested Jan 2021

Whilst activity should be prioritised for phase 1, in order to achieve Phase 2 (which provides long term significant benefit), some work packages and strategy work will need to be initiated to run concurrently with Phase 1.

- What do you think should constitute an operational offer for researchers and innovators for January 2020?
- What interim milestones should we set between January 2020 and January 2021?

End states and interactions

Whilst not an achievable goal, we should consider long term optimal scenario for all datatypes in order to inform strategy and direction, namely:

- All providers having 'perfect' data quality
- All using a single open standard with full metadata available via a metadata catalogue and dataset selection tool
- Ability to generate watermarked, DOI linked and metadata-rich datasets mapped to multiple terminologies all using a common open standard

In order to determine the highest value activities through HDR UK to achieve these end-states, it is necessary to understand the current state, following which gap analysis will reveal tasks and outcomes to bridge the gaps.

This initial scoping work will be based on the MVP concept to allow function of the Gateway MVP and will be carried out in conjunction with the Alliance and Hub streams. Once the tasks and outcomes have been determined this will be mapped back to the available resource for prioritisation and project planning. Based on best estimates, it is possible to generate tasks which can add value and will be required, with prioritisation TBC.

Data quality and standards (DQS) activity will therefore require interaction and close alignment with many other aspects of HDR UK activity including Alliance, Hubs, Gateway, public advisory board and communication and Digital Health Insights group.

Proposed workstreams and actions

Four inter-related workstreams are proposed. The initial focus is on phase 1 MVP deliverable, which is to ensure the alliance and hub datasets are discoverable and visible on the gateway:

Workstream#1 Aligning with UK-wide standards setting

1. Develop an overarching '**Data Quality and Standards Oversight Group'** composed of HDR UK CCDO plus senior strategic members from NHSI/NHSE, CQC, NHSD, NHSX, NICE, PRSB plus equivalent bodies in devolved nations and others as appropriate.

   The role of this group will be primarily to provide strategic oversight and 'sense check' regarding alignment of standards and recommendations across organisations for HDR UK, could be virtual in nature, with initial focus on Alliance / DIH programme, but with wider scope for HDR UK as appropriate. The group will primarily advise the HDR UK infrastructure steering group/ Alliance Data Quality and Standards Delivery Sub-Group, and through that, the senior management team. The group will also provide a source of information for 'central' standards.

Phase 1 output by Jan 2020; DQS white paper published
Phase 2 output suggested Jan 2021; revised and updated white paper based on learnings

- Who do you recommend should form part of this group?

Workstream #2 Implementing through the Alliance

2. Develop the UK Health Data Research Alliance Data Quality and Standards Delivery Sub-group, through a **Data Officers Group (DOG)**, with named data officer representatives from all organisations associated with the Alliance, hubs and substantive sites (with input from other organisations as required*).

   This will be the main delivery group that discusses, agrees standards, coordinates, and provides guidance to the other groups, in addition to supporting with the initial scoping work regarding current state of data. This group will also have the tasks of producing Alliance guidance / policy regarding DQS covering:
   - data quality
   - data standards
   - metadata standards
   - data provenance
   - ontology / terminology services and use

** Additional Data officers may be required from HDR UK substantive sites and collaborators such as ONS, NIHR, BHF, PRSB, ODI, RAENG, EUSTANDS4PM, GA4GH, etc TBC)*

The initial task for the DOG will be scoping the current state via a scoping template process. Therefore, the initial focus will be on Alliance members and hubs, but with wider scope for HDR UK as appropriate.

Phase 1 output by Jan 2020; groups established, DQS white paper published
Phase 2 output suggested Jan 2021; TBC activity depending on infrastructure developments

- Do you agree with the list of areas requiring policy and guidance set out above?
- What work to date would you recommend as input to the work of the DOG?
- Are there other collaborators who you think should be engaged?

Workstream#3 Future planning and horizon scanning

3. Develop several smaller task-specific '**special interest groups' (SIGs)**, involving individuals with a particular interest or subject matter expertise in specific topics and feedback to the main DOG and oversight group through the CCDO.

   The SIGs would aim to both provide advice (act as 'standing' virtual groups) and deliver position papers (task and finish groups) for various elements which are of strategic importance for UK health data research, particularly in the medium term, as determined by the DOG and user community, for example:
   - data de-identification standards
   - synthetic data standards
   - phenomics (link with the main human phenome stream)
   - data streaming standards
   - data provenance standards
   - FHIR / SMART on FHIR standards / implementation
   - NLP standards
   - PROMS/PREMS (iCHOM etc) standards
   - Imaging standards (inc whole slide imaging)
   - Genomic standards

Phase 1 output by Jan 2020; groups established and key topics identified
Phase 2 output by Jan 2021; position papers published from all SIGs

- Do you agree with the list of potential SIGs?
- What work to date would you recommend as input to the work of the SIGs?
- Are there particular individuals you would recommend for a SIG?

Workstream#4 Influencing through others

4. Strategic wider work to be considered as ongoing aspects of data quality and standards (not otherwise specified)
    - establish links with NHSE/I regarding incentives to **improve NHS data quality** through initiatives such as CQINS etc
    - establish **links with NHSX regarding emerging NHS standards** for new data types including Internet of things and streaming data
    - establish a **'funders research data group'** to encourage FAIR data principles, including pilot study with some funders and charity organisations, and potential links through to HDR UK approve trusted research environments for charity use
    - establish a data group / focus within the public advisory board, particularly around developing HDR UK guidance for **data standards for patient generated data**
    - establish strong link between the applied analytics work stream and the digital health insights work stream particularly regarding **standards and guidance for deployment of analytic / CDS tools**

Phase 1 output by Jan 2020; DQS white paper published referencing above
Phase 2 output by Jan 2021; TBC

- Do you agree with the list of potential topics for influencing through others?
- Are there any you would add or remove?

Proposed data standards

Terminology and definitions

A wide array of terms and abbreviations are used in this area. This section provides definitions of the various terms as used by HDR UK for the purposes of this document.

Term	Definition
Standard	Technical, functional, or performance-based rule, condition, requirement or specification that stipulates instructions, fields, codes, data, materials, characteristics, or actions for common usage.
Data Standard	Standards intended to provide consistent meaning to data across information systems and organisations which may include representation, format, definition, structure, transmission, manipulation, use, and management.
Data model	Description of the structure in which elements of data are organised and standardised, including how they relate to each other and real-world entities.
Data schema	Description of how data is organised in relation to how a data repository is constructed
Data structure	Collection of data values, relationships and functions that can be applied to the data
Data format	Organisation of data according to pre-set specifications
Terminology	Collection of terms used in a given setting / scenario
Value set	Subset of specific terms for particular use cases
Clinical classification	System for assigning clinical data items to categories
Ontology	Description of entities and how they are subdivided and related
Specification	Detailed description of components required for a specific function/activity
Dataset	Collection of related data elements
Data element	Specific unit of data within a dataset that has precise meaning
Metadata	Set of data providing information about other data, either at dataset level or value level
Data dictionary	Information describing the contents, format, and structure of a specific database
Syntax	Set of rules or structure of statements

- Do you agree with these definitions?
- Are there other terms that should be added?

Our approach to standard setting

There are a large number of existing data schemas and formats for healthcare data. Whilst there are several open data standards and schemas, many are proprietary. Particular data models/standards are often designed for specific purposes and therefore when attempting to determine a unifying standard there are advantages and disadvantages to each. In addition, most research data is not collected and stored according to any published standard, and often has extremely limited or absent metadata.

These factors make sharing and interoperability of both clinical and research datasets either impossible or extremely difficult and time-consuming requiring individual data mapping and curation for any projects.

Specific datatypes which are relatively novel may require different approaches to data quality and standards compared to traditional healthcare data, specifically streaming data / data from IoT devices. This provides an opportunity to agree standards early and avoid the legacy issues that are described above.

HDR UK will encourage the use of common data and metadata standards throughout the Institute, aligning with other national standards for health data where possible and using established open standards. It will do this by embedding them within the Innovation Gateway and any HDR UK contracted services and include as a requirement of any funded activities.

Furthermore, HDR UK will work with partners to align with existing data and metadata standards rather than developing any HDR UK specific standards, wherever possible.

HDRUK recommends aligning with WHO approved terminologies/ontologies including ICD10, SNOMEDCT, LOINC and HL7, in addition to NHS OPCS codes. In addition, several other initiatives which include reference to data standards are ongoing and it is intended, where possible, that the HDR UK strategy and recommendations aligns where possible. These include, but are not limited to:

- GA4GH – global alliance for genomics and health https://www.ga4gh.org
- EU-STANDS4PM - standards for in silico approaches in personalised medicine https://www.eu-stands4pm.eu/about (WP1 data standards and sources). Links with UK Genomics Informatics Strategy
- European innovation through health data https://www.i-hd.eu/

Standards are ultimately designed to advance interoperability and data sharing. We subscribe to the concept recently proposed by the US ONC of sharing health information *without special effort.* https://www.federalregister.gov/documents/2019/03/04/2019-02224/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification.

Without special effort refers to the recipient of the data/API users and the APIs must be:

- **Standardized** – same technical API capabilities using modern computing standards such as RESTful interfaces, XML/JSON etc.
- **Transparent** – technical documentation necessary to interact with the APIs freely and publicly accessible

- Do you agree with the outline approach above, including proposed alignment with WHO approved terminologies and ontologies?
- Are there any other data standards related initiatives that should be added to the list above?
- Do you agree with the principle and key elements of US ONC 'without special effort'?

Standards

Data standards

We would like to your views on the following recommendations regarding data standards for HDR UK research purposes:

1. We do NOT propose to mandate a named standard for export
2. The export must be accompanied by appropriate information, such as a data dictionary or export support file, for the exported information to assist the receiver in processing the dataset without loss of information or its meaning to the extent reasonably practicable.
3. The export format should be made publicly available
4. The data must be made available in an open format as above. The export format need not be the same format used internally by the data owner and proprietary data models do not need to be made public.
5. We propose to adopt the HL7® Fast Healthcare Interoperability Resources (FHIR®) standard as a foundational standard*. This aligns with ONC regulations for health IT providers in USA, NHS Digital and healthcare technology vendors.
6. We recommend use of FHIR Release 4 where possible. This has several key improvements, including certain foundational aspects in the standard and "FHIR resources" designated as "normative". Release 4 has additional implementation guidance that explicitly specifies how to handle batch exports via FHIR more efficiently.
7. The Alliance Data Officers' Group (DOG) should develop and propose an initial set of core FHIR resources for health research as a possible 'HDRUK Core Dataset'. This may be aligned to the NHS FHIR profiles or US core API specifications, (E.g. composed of FHIR resources such as AllergyIntolerance; CarePlan; Condition; Device; DiagnosticReport; Goal; Medication; MedicationOrder; MedicationStatement; Observation; Patient; and Procedure, Provenance and DocumentReference resources). This could result in a FHIR implementation reference group.
8. We propose to adopt the OpenID Connect Core 1.0 standard paired with OAuth 2.0 implementation for user authentication through RESTapis with industry developed security best practice guidelines for OAuth 2.0 implementations, including use of access tokens and refresh tokens for API use. This will be developed and aligned with current NHSD/NHSx guidance around web standards and will support SMART on FHIR development.

Please state your level of agreement with the individual recommendations and add comments if you disagree or have other suggestions.

Adopting the FHIR standard alone is insufficient to provide the level of consistent implementation that will be necessary for "without special effort". FHIR profiles to describe either an individual FHIR resource, or an entire implementation specification consisting of multiple FHIR resources for specific use cases should be documented appropriately.

In addition, within the FHIR information model, a range of terminologies may be referenced, mapped and/or used including SNOMED CT, ICD10, DM&D, LOINC. These should also be appropriately referenced in the documentation. The appropriate and consistent use of ontology/terminology services is an area to be developed in due course in conjunction with other organisations active in this space such as NHSD/NHSx and Ontoserver project.

Other data models/standards are available, such as openEHR, OMOP, etc, and research datasets may also be associated with non-standard formats but mapped to terminologies such as SNOMED CT or LOINC, all of which may enable mapping to FHIR-aligned standards. In circumstances in which the organisation and data owner is unable to provide data in a FHIR-aligned format**, data can be provided in other formats (E.g. OMOP) with the proviso that there would be an expectation that a data model/schema, including formats and syntax in addition to terminologies used and data dictionary, can also be provided. This is relatively straightforward providing open standards are used from which appropriate mapping, interpretation and semantic interoperability can be derived.

Notes regarding FHIR

* FHIR was designed primarily as data communication specification rather than for clinical data storage / persistence. However, from a data model perspective, the FHIR model broadly follows an Entity-Attribute-Value (EAV) pattern. There is no specific 'right' way to store data in the persistence layer for FHIR. Such data could be stored directly in a datastore using JSON format or in a specific SQL or NoSQL database for example. However, one major advantage of the FHIR is a well-described and ready-to-use informational model that is good enough for most purposes. We therefore recommend starting with the FHIR data model, and to support FHIR, there may be a need for transformation from an existing to FHIR and vice-versa. Such transformation may be a relatively trivial process if the local model is conceptually aligned to FHIR, whereas use of normalized relational databases for FHIR resources may result in large numbers of tables. Modern databases may allow a hybrid approach to efficiently store resources using other features for search and transformation.

There is a misconception that FHIR provides a single industry standard 'data format'. However, implementations may differ, and the capabilities of specific APIs may differ, etc. Similarly, two organisations may implement the FHIR API but with differing specifications and data elements / resources. Finally, the use of FHIR extensions, which may be required for defining data for specific use cases, may further reduce immediate interoperability.

Nevertheless, alignment with open, freely available standards and specifications such as FHIR begin to address many issues regarding data interoperability and it is the intention of HDR UK to use the expertise of those working with FHIR and other standards and specification to develop best practice through SIGs and the DOG.

**For ease of use throughout the document we use the 'FHIR' notation. For the purposes of this documentation this could mean either full FHIR compliance, through a FHIR API, providing data in JSON/XML FHIR format, or, simply storage/provision of data in a 'FHIR-compliant' format. It is recognised that most organisations cannot currently provide data through a full FHIR API, and this may not be appropriate, where the ability to do this would require significant investment. Therefore FHIR-aligned in this context means that the data is not necessarily delivered through a FHIR API, but may be delivered in other standard database file formats in which the FHIR data model is followed and the data elements or variables maintain general FHIR naming conventions, value formats, terminologies, etc., to maximise semantic interoperability. https://www.hl7.org/fhir/

Non-standard data

A significant proportion of research data may be non-standard in nature and therefore may not have an existing FHIR (or other open standard) descriptions. In such cases the general FHIR information model/schema should be used, the details of which can be provided with the data. For example, using FHIR extension methodology.

HDR UK will discuss and engage with other standards bodies around the appropriate curation of extensions and profiles in order to prevent multiple forking of standards.

- Do you agree with this approach to non-standard data?
- If no, why not?

Metadata standards

Datasets will be required to be associated with both dataset level metadata and data level metadata. In phase 1 (Gateway MVP), the requirement will be for dataset level metadata only.

Precise requirements for metadata standards will be guided by the Metadata catalogue and gateway requirements, but it is envisaged that industry metadata standards will be adopted (for example Dublin Core / ISO 15836 / DataCite. http://dublincore.org/(DCMI)). The gateway would be intended to be able to adjust and read metadata in machine learnable format (XML).

Attribute level metadata would ideally be aligned with FHIRv4, but any appropriate open standard is acceptable (see above).

- Do you agree with this approach to metadata standards?
- Please add any suggestions or alternative standards.

Other standards

The predominant focus of this document is around metadata and data standards but HDR UK will consider alignment with other standards throughout the organisation, for example:

- ISO 27001 Trusted Research Environment information security management system (ISMS).
- ISO 9001 Quality Management
- ISO 8000 Data Quality and Enterprise Master Data
- ISO 11073 Personal health data standards
- ISO 12300 Mapping between terminologies
- https://www.iso.org/committee/54960/x/catalogue/

- Are there other standards that you think should be included in the list above with relation to health data for research?

Data quality measures

Whilst there are some relatively well accepted measures of 'data quality ', these tend to be crude or generic. Data quality as a concept is difficult to assess objectively since the quality requirement depends on the use / purpose. For example, data for population scale planning may be of 'acceptable quality' at low resolution and with numerous missing values, whereas the same data may be of limited value and therefore 'low quality' for a particular disease related longitudinal study.

Nevertheless, the Data Officers' Group will collaborate to develop reasonable measures of health data quality, which will be extended with time and for specific use cases/scenarios.

Initially, it is proposed to develop a minimum requirement for metrics applicable to datasets using 'generic' data quality measures derived from established organisations such as DAMA (https://dama.org/content/what-data-quality) in relation to the main quality dimensions:

- Completeness
- Consistency
- Integrity
- Validity
- Uniqueness
- Accessibility
- Timeliness
- Reasonability
- Accuracy

This preliminary set of Data Quality elements will be built upon existing resources. This includes the NHS Digital Data Quality Strategy and Data Quality Maturity Index https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/data-quality. The initial focus will be on coverage, completeness, validity and default values, extending to consistency, timeliness and integrity.

During Phase 2, the intention is to develop more sophisticated and health specific measures of data quality through the Alliance, hubs and gateway. It is anticipated that data quality tools for dataset profiling will be increasingly included within the gateway, which may also provide additional information such as metadata or creating / updating data dictionaries. Such gateway tools are likely to represent both objective metrics-based criteria (such as proportion of missing values etc) and subjective 'quality' data feedback from users.

If issues are apparent regarding data quality it will not be the HDR UK policy to attempt or require remediation of such data, this will be at the discretion of the data processors and controllers. The focus will be on making the data quality information available to users and custodians.

- Do you agree with the approach outlined above? If no, please provide comments or suggestions?
- Please provide a brief description of your current approach to data quality measures and highlight any tools, resources or methods that you feel could have broader applicability
- Do you have a view whether data quality should be evaluated through self-assessment or third party?
- What group is best placed to arbitrate on what is 'good' data?

Additional Questions / areas for comment

- Should we develop / support standards for analytics?
- What should be the internal process for agreeing standards through the DOG where there is disagreement?

References

- https://www.healthit.gov/sites/default/files/page/2019-02/HITACNPRMPresentation.pdf
- https://www.healthit.gov/topic/laws-regulation-and-policy/notice-proposed-rulemaking-improve-interoperability-health
- https://www.hl7.org/fhir/
- https://www.ga4gh.org/