HDRUK
Health Data Research UK

PRINCIPLES FOR DATA STANDARDS

June 2020

HDRUK
Health Data Research UK

Contents

HDRUK
Health Data Research UK

Overview

Health Data Research UK's mission is to unite the UK's health data to enable discoveries that improve people's lives, so that every health and care interaction and research endeavour will be enhanced by access to large scale data and advanced analytics. Establishing common standards for healthcare data and metadata is a fundamental requirement for this mission.

Health Data Research UK (HDR UK) has an important role in leading the development of best practice. The intention is to encourage behaviours to improve data usefulness and usability through the provision of clear guidance and recommendations, developed in conjunction with the health data community, as well as patients and the public as recipients of the benefits. This best practice can be implemented through the different elements of the HDR UK infrastructure, including the Health Data Research Hubs, the UK Health Data Research Alliance (the 'UK Alliance'), the Health Data Research Innovation Gateway (the 'Gateway') and research sites. In addition to the elements of the infrastructure where HDR UK has closer control, it will also work with organisations such as NHSX and its equivalents in the devolved nations to shape system-wide mandatory requirements.

There are many existing data schemas and formats for healthcare data, many of which are proprietary. In addition, most research data are not collected and stored according to any published standard, and often have extremely limited or absent metadata. These factors make access to, and interoperability of, both clinical and research datasets either impossible or extremely difficult and time-consuming, requiring individual data mapping and curation. In addition, several open data standards, specifications and schemas for healthcare data are also available (including HL7 FHIR, OMOP, Sentinel). Particular data models/standards are often designed for specific purposes and therefore there are advantages and disadvantages when attempting to determine a unifying standard.

HDR UK proposes that organisations collect, store and provide their datasets according to defined open standards through the UK Alliance and other HDR UK infrastructure including the Hubs and research sites. HDR UK will encourage the use of common data and metadata standards throughout the Institute. These will build on and be aligned with existing standards for health data where possible and use established open standards. They will be embedded within the Gateway and specified as requirements within HDR UK contracted services and activities managed by members of the HDR UK community. It is recognised that specific standards may be required for particular projects.

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 the 'Improving Health Data' workstream, with the standards and utility elements led by the Chief Clinical Data Officer (CCDO).

Initial principles have been developed in consultation with data officers across HDR UK's community (the Data Officers Group). Feedback was received from almost 50 individuals across more than 30 organisations. The feedback has been used to help shape the strategy and prioritise activities to realise the benefits of health data research to patients and to UK society. It is anticipated that having a common direction for

health data standards through HDR UK will also support the UK's position as a global leader in the field of health data research. The document is therefore presented as initial principles which will be developed further as the community grows and matures.

Whilst maintaining the principles noted in this document, HDR UK intends to be pragmatic and respect the feedback from implementers, to take seriously any challenges that implementers raise in relation to the use of suggested standards, such that the HDR UK position paper is revised accordingly with user feedback.

Implementation

HDR UK will work with partners to liaise and align with existing data and metadata standards where appropriate rather than developing any HDR UK specific standards. HDR UK 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 that HDR UK strategy and recommendations align with these where possible. These include, but are not limited to:

- **Global Alliance for Genomics and Health (GA4GH)**
    - www.ga4gh.org
    - For example: htsget API, refget API, Crypt4GH, CRAM, DURI, DUO, GA4GH passports, phenopackets (and phenopackets on FHIR)
- **EU-STANDS4PM** - standards for in silico approaches in personalised medicine
    - www.eu-stands4pm.eu/about
    - Links with UK Genomics Informatics Strategy
- **European Innovation through Health Data**
    - www.i-hd.eu/
    - OHDSI / EHDEN
- Other relevant groups as appropriate, including Sentinel, PCORnet and others

Principles for Data Standards

These principles are encouraged for organisations that are participating in any part of HDR UK's activities, including contracted services or activities managed by members of the HDR UK community. Further details on the principles are given in the following section.

0. As described in the Health Data Research UK's Principles for Participation, data should be **Findable, Accessible, Interoperable and Reusable (FAIR)**
1. This work is to be **minimally interventionist**, and to only prescribe specific actions or specific standards where this is deemed necessary. Where principles alone will suffice (when multiple standards would meet the requirements), no specific standards will be mandated
2. Standards that are used should be **explicitly described**, including the descriptions of any export which should include the model/schema, syntax and data dictionary or reference. This should include provenance tracking where possible
3. **Open** standards should be adopted wherever possible, minimising the proliferation of proprietary data standards
4. Organisations should aim to maintain a **consistent, internal approach** to data standards, explicitly referencing their approach to standards in their data strategy
5. Data should be able to be used according to the principle of **without special effort** as a result of the standard used
6. Standards adopted should be **aligned with existing and provisional standards proposed by national and international bodies** where possible, recognising that the remit and aims of HDR UK and other bodies may overlap but differ
7. Ideally, standards should be **common for both research and clinical or operational uses**, in order to optimise both research and clinical benefits of data, recognising that the primary focus of HDR UK is research use of health data
8. Organisations forming part of the HDR UK network should have **established and aligned data strategies**, including how these improve the usefulness of data
9. Benefits of standards should be widely disseminated through **communication and educational** events, both to researchers and the public

These standards will be followed up with implementation guidance later in 2020. It is likely that, for several reasons (including to align with UK NHSX/NHS Digital interoperability guidance and ONC regulations for health IT providers in USA, and healthcare technology vendors) HDR UK will broadly support formal adoption of the HL7 FHIR[1] standard where appropriate, and the associated implementation specifications.

- We propose to adopt the HL7® Fast Healthcare Interoperability Resources (FHIR®) standard as a foundational interoperability standard where appropriate in alignment with NHSX/NHS Digital
- We suggest use of the appropriate stable release, currently FHIR Release 4, where possible, which 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.[2]

[1] We note that 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 the majority of purposes. We therefore recommend generally 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. However, 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' since 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.

[2] This is also in alignment with the 2019 announcement from the US NIH recommending FHIR for research data use;

https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-122.html,
https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-127.html,
https://grants.nih.gov/grants/guide/notice-files/NOT-HS-19-020.html

HDRUK
Health Data Research UK

Further Details of Principles

Relating to Principle 2:

"Standards that are used should be **explicitly described**, including the descriptions of any export which should include the model/schema, syntax and data dictionary or reference. This should include provenance tracking where possible."

As much detail about the expected standards should be provided in advance, to all users. This should be openly available and discoverable to all, via the Health Data Research Innovation Gateway, in line with the metadata specification. The metadata specification for the Gateway is based on existing industry standards (for example: Dublin Core / ISO 15836 / DataCite. http://dublincore.org/(DCMI)). The Gateway will be able to adjust and read metadata in a machine-readable format.

The Gateway would be intended to be able to adjust and read metadata in machine learnable format (XML). 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, but the data must be made available in an open format as above.

We do not intend to mandate a named standard for export but data providers must provide 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. The export format should be made publicly available.

Relating to Principle 5:

"Data should be able to be used according to the principle of 'without special effort' as a result of the standard used."

In line with US ONC, health information should be shared in a way that minimises additional effort by the recipient and data/API users: www.federalregister.gov/documents/2019/03/04/2019-02224/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification.

For example, the APIs must be:

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

Relating to Principle 6:

"Standards adopted should be **aligned with existing and provisional standards proposed by national and international bodies** where possible, recognising that the remit and aims of HDR UK and other bodies may overlap but differ"

It is recognised that a significant proportion of research data may be non-standard in nature and therefore may not have an existing FHIR, OMOP or other open standard descriptions. In such cases the information model/schema and data dictionary used should be provided with the data.

HDR UK should discuss and engage with other standards bodies around the appropriate curation of extensions and profiles in order to prevent multiple forking of standards (E.g. NHSX/NHS Digital).

HDRUK
Health Data Research UK

Appendix 1: Terminology and definitions

For the purposes of this document, health data refers to data generated by, or associated with, health care provision. At this stage, it does not include broader data such as social or environmental data, although it is recognised that these data may also be highly relevant to health and may subsequently become within scope. This is an area for ongoing discussion with the current position based on the 2020 consultation findings with the HDR UK community. This will expand to include additional 'novel' data sources such as patient generated health data and data from Internet of Things (IoT)/streaming devices.

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 preset specifications
Clinical terminology	Collection of terms used in a given clinical 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 including history and changes and context
Syntax	Set of rules or structure of statements
Information standard	Rules by which information is described and recorded
Reference data	Known dataset that defines permissible values to be used or for comparative profiling
API	Application programming interface (communication protocol between different software elements)
Data provenance	Record of the origins of data including derivations or transformations from the original data, which can be used to form assessments about its quality, reliability or trustworthiness
Information model	Representation of concepts and relationships for a particular context
Interoperability	Ability to function with systems other than the index system
Data mapping	Describing the relationship between data elements in different data models
Data controller	Controls data usage and has data protection responsibility
Data processor	Uses or processes the data on behalf of the data controller

HDRUK
Health Data Research UK

Appendix 2: Next Steps

This paper will be taken forward using the following next steps:

1. Review as part of engagement through the Alliance Data Officers Group (DOG) to decide on **key standards and approach to implementation**. Additional members could be included from other groups as required.[3]
2. Develop **Special Interest Groups (SIGs)** as need, to work on areas of specific interest on behalf of the DOG. Initial consultation with the community has suggested the following as possible SIGs, from which specific priority SIGs will be established:
    - Data Linkage standards
    - Data de-identification standards
    - Synthetic data standards
    - Phenomics standards
    - Data streaming standards
    - Data provenance standards
    - NLP standards
    - PROMS/PREMS standards
    - Imaging standards (Inc. pathology whole slide imaging)
    - Genomic standards (links with GA4GH etc)
    - Knowledge graph standards
    - Terminology use standards
    - FHIR / SMART on FHIR standards / implementation guidance
    - Data Use standards (including legal requirements such as GDPR, non-UK data)
    - 'Clinical data for research use' improvement
    - Data Security standards
    - Data visualisation /presentation standards
3. **Influence wider community** to input into and adopt the principles. This includes establishing links with key organisations including:
    - establish links with NHS England/NHS Improvement 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

[3] Including, but not limited to Office of National Statistics, National Institute for Health Research, British Heart Foundation, Professional Records Standards Body, Open Data Institute, OHDSI, RAENG, EUSTANDS4PM, Global Alliance for Genomics and Health, Association of British Pharmaceutical Industry, Medicines and Healthcare Products Regulation Agency, Medtech companies, research data alliance, FCI, British Medical Association, Royal Colleges, Public Health England, data coordination board, Genomics England, EBI, SNOMED, commercial partners, Interopen, CDISC, National Institutes of Health, NHS trusts, Healthcare Quality Improvement Partnership, HL7, Local Health and Care Record Exemplars, devolved nation representatives and patients and families

- 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
- establish group to engage with, and feedback to, NHS trusts and frontline clinicians regarding data quality (possibly through Professional Records Standards Body)
- establish links with international groups working on health data standards

4. It may be necessary to convene a national group to develop and 'sense check' proposed standards, facilitated by HDR UK. This should include senior strategic members from NHS England and Improvement, Care Quality Commission (CQC), NHS Digital, NHSX, National Institute for Care Excellence (NICE) and equivalent bodies in devolved nations, as well as others as appropriate.

HDRUK
Health Data Research UK

Appendix 3: Notes regarding FHIR Specification

It is recognised that adopting the FHIR standard alone is insufficient to provide the level of consistent implementation that will be necessary for "without special effort" (Principle 5) since in FHIR additional constraints on base FHIR resources for specific use cases can be developed through FHIR profiles. These could describe either an individual FHIR resource, or an entire implementation specification consisting of multiple FHIR resources and should be documented appropriately.

In addition, within the FHIR information model, a range of terminologies may be referenced/mapped/used including SNOMED CT, ICD10, DM&D, RxNorm, which should 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 bodies active in this space such as NHSD/NHSX and Ontoserver project).

We suggest that HDR UK consider developing an initial set of core FHIR resources for health research as a possible 'HDR UK Core Dataset', which may be, for example, aligned to the core NHSX FHIR profiles and/or US core API specifications as appropriate. This will be defined by the DOG and could be expanded further in a FHIR implementation reference group, in close relationship with existing groups such as NHSX and bodies such as Interopen.

We also propose to adopt authentication standards such as OpenID Connect and OAuth 2.0 implementation for user authentication through REST APIs with industry developed security best practice guidelines for implementations, including use of access tokens and refresh tokens for API use. This will be developed and aligned with current NHS D and NHSX guidance around web standards and will support 'SMART on FHIR' development.

It should be noted of course that other data models/standards are available, such as openEHR, OMOP, etc, and research datasets may currently be associated with many non-standard formats, some mapped to terminologies such as SNOMED CT or LOINC. All of these may enable mapping to FHIR or other standard data models / specifications. In circumstances in which the organisation and data owner is unable to provide data in a FHIR API or FHIR-aligned format[4], or where this is not appropriate depending on use case, data should be provided in other established standard formats with the expectation that a data model/schema, including file format, syntax, in addition to terminologies used / data dictionary, can also be

[4] 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 the majority of organisations cannot currently provide data through a full FHIR API, and this may not be appropriate, with the ability to do this would require significant investment. Therefore FHIR-aligned in this context means that the data is not delivered through a FHIR API, but rather may be delivered in other standard database file formats but in which the broad FHIR data model is followed and the data elements / variables maintain general FHIR naming conventions, value formats, terminologies, etc to maximise semantic interoperability. https://www.hl7.org/fhir/

provided. This is trivial providing open standards are used from which appropriate mapping, interpretation and semantic interoperability can be derived. Established standards should be used for specific data types such as clinical documentation (e.g. XDS-IHE, CDA) or imaging (DICOM).

HDR UK should review and update the position with regard to specific suggested standards for data models based on year 1 feedback from the Hub and Alliance members. For example, OHDSI OMOP CDM is widely used, especially in the US, with many existing data engineering and analysis tools available, and may be a suitable bulk data persistence format. The feasibility of widespread HDR UK use of models such as FHIR and OMOP should be explored through the Data Officers Group.