HDRUK
Health Data Research UK

HDR UK – Recommendations for Data Standards in Health Data Research

Interim Paper June 2021

HDRUK
Health Data Research UK

Contents

Executive Summary

Establishing common standards for healthcare data and metadata is a fundamental requirement to enable health data research to improve people's lives. Health Data Research UK's (HDR UK) Principles for Data Standards (June 2020) set out an approach to adopting data standards for health data research concerning structured electronic health records data only. This paper on Recommendations for Data Standards builds on these principles and encourages improvements in data usefulness and usability with the primary focus to benefit research through the provision of clear guidance and recommendations, developed in conjunction with the health data community, as well as patients and the public. This document does not cover ontology standards, Web Ontology Language (OWL) and web standards.

User and custodian surveys were conducted with academic researchers, charities, data custodians, healthcare providers, life science companies and AI & Technology companies to understand the current use of health data standards and opportunities for greater alignment. The surveys identified that:

- The majority (64%) of health data users said they have basic or no data standards expertise, with much greater stated expertise among the industry users compared with the academic researchers.
- 85% of users were in support of a core set of data standards to enable health data research
- Both users and custodians highlighted the importance of open standards and clinical terminologies
- Currently data users are using a wide range of data standards, with greatest alignment around Clinical Data Interchange Standards Consortium (CDISC), Observational Medical Outcomes Partnership (OMOP)
- Data custodians also currently support a wide range of data standards, with OMOP and HL7 FHIR the most frequently supported

Recommendation: For organisations considering establishing standards such as data models and messaging standards for research purposes, HDR UK is suggesting the use of OMOP and/or HL7 FHIRv4 as standards/specifications for research which can be adopted most widely by both users and custodians for a range of purposes if an organisation is in the position and has access to the financial resources to do so.

However, for organisations already using different standards or models HDR UK recognises that there are resourcing and cost implications for each organisation and potential information loss associated with transition to OMOP and HL7 FHIRv4. Ultimately, the appropriate standards should be selected to support the specific use cases and capabilities.

Metadata Specifications and Data Utility: For metadata specifications and data utility, we recommend using HDR UK's metadata specification and the Data Utility Framework. The metadata specification makes it easy for researchers to search, sort and filter the datasets when using the Health Data Research Innovation Gateway to access data. The Data Utility Framework evaluates the usefulness of the data for a given purpose and provides the ability to assess and compare datasets from different sources at scale.

These recommendations are for the health data user and custodian community. Following this Interim Paper, the HDR UK White Paper for Data Standards will be published in Autumn 2021.

Context

HDR UK is the UK's national institute for health data science with a mission 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. Our focus is data to benefit research directly in the first instance as opposed to clinical care, which may in turn lead to benefits at the frontline of clinical care.

In June 2020 HDR UK released the Principles for Data Standards paper which set out a series of principles for organisations considering the adoption of new standards for electronic health record data only, and not imaging, genomics, ontology standards, OWL and web standards.

The initial principles were developed in consultation with data officers across HDR UK's community (the Data Officers Group) in June 2020 and with feedback was received from almost 50 individuals across more than 30 organisations.

This interim paper builds on the existing principles, as well as feedback provided by both custodians and users of health data that was conducted through structured surveys. Consultations took place April-June 2021 with input requested in the relevant sections. As part of these consultations, HDR UK would particularly value feedback from organisations on further lessons or challenges raised while implementing the suggested standards. Following this consultation, an updated "white paper" will be published in Autumn 2021.

HDR UK will encourage adoption of the recommendations 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. We also welcome the adoption by partner organisations within the UK, for example NHSX and its equivalents in the devolved nations to shape system-wide requirements.

The paper is part of a suite of HDR UK resources developed to enable health data research, including the Data Utility Framework which provides a common standard for measuring data utility of health datasets, the metadata specification and the Health Data Research Innovation Gateway. The "Gateway" provides a common portal to discover health datasets (including information on the standards used by those datasets) and will support the work developing Trusted Research Environments and the move to federated analytics.

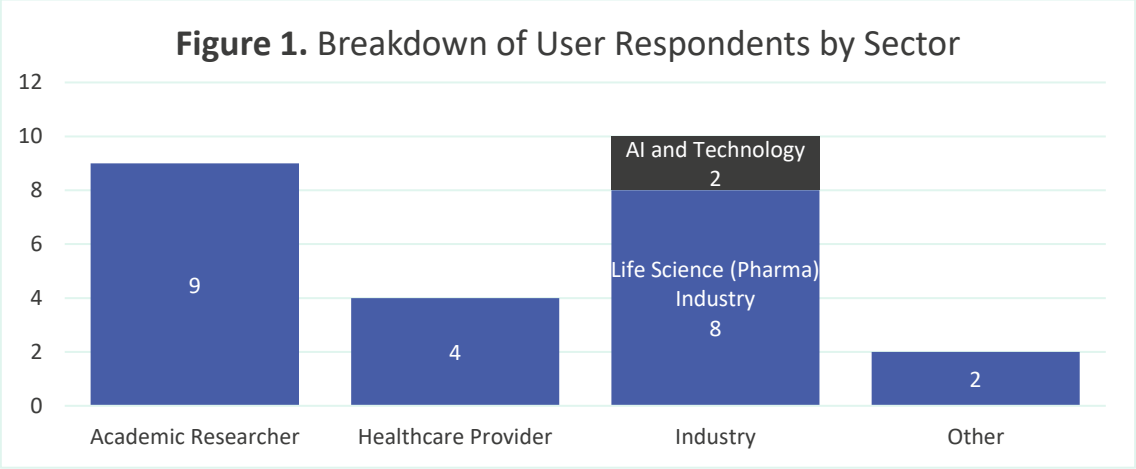
HDR UK will also be including details of the international landscape of health data sharing initiatives in the white paper, literature and legislation of data standards; case studies from the Health Data Research Hubs; and the Trusted Research Environment White Paper.

Health data user and custodian survey findings

HDR UK conducted a data standards survey with data custodians to understand the data models and standards supported in their organisations (July 2020). A subsequent survey with health data users was conducted to ask them about the data standards they would like to see or that are required for them to conduct their research (January 2021).

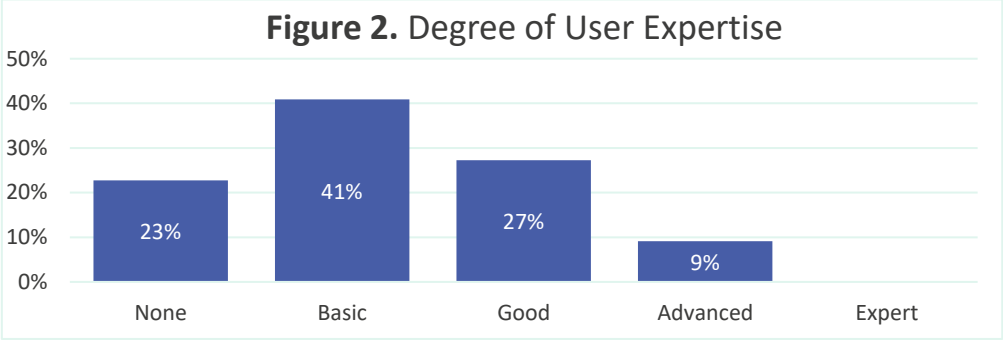
Together the data custodian and user survey provide an overview of the current landscape of data standards in the UK, and indication of the current degree of alignment and opportunities for further alignment, and what users and custodians need to do this.

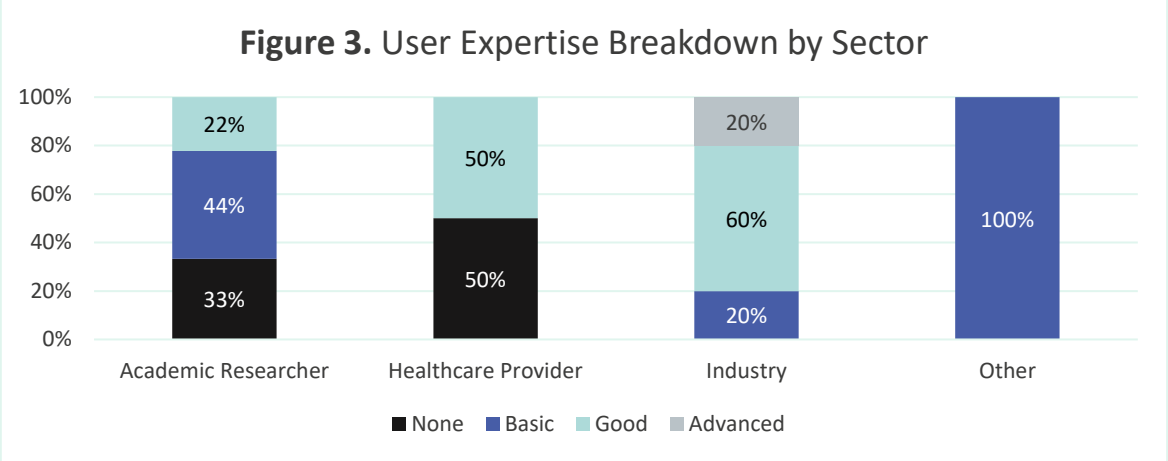
With thanks to our partners, the Association of the British Pharmaceutical Industry (ABPI) for their valuable engagement and for circulating the surveys, thirty-eight users and data custodians responded to the data standards surveys. Twenty-four of these respondents were users, and the user sectors represented in the survey are displayed in Figure 1 (N.B. some users did not specify their sector and are therefore excluded in the breakdown):



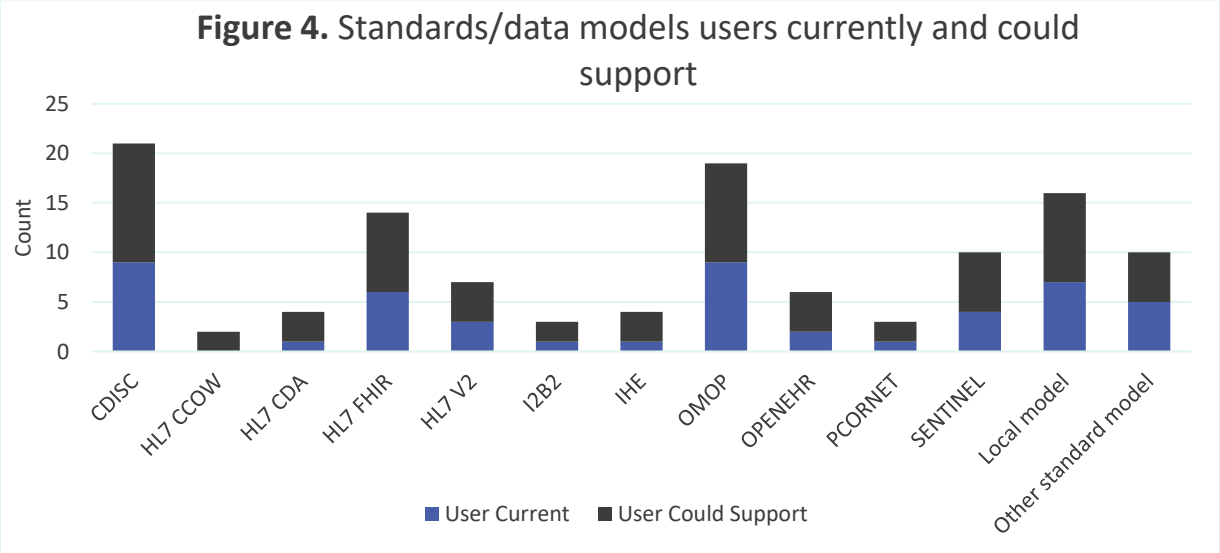
Findings

For the user survey, respondents were asked about their degree of expertise and capability regarding data standards/models. 64% of respondents to the user survey stated they had basic or no degree of data standards expertise with no respondents claiming expert knowledge (Figure 2). Of these respondents, 80% of Industry users stated that they had good or advanced expertise in data standards, while 77% of Academic Researchers had basic or no expertise (Figure 3).



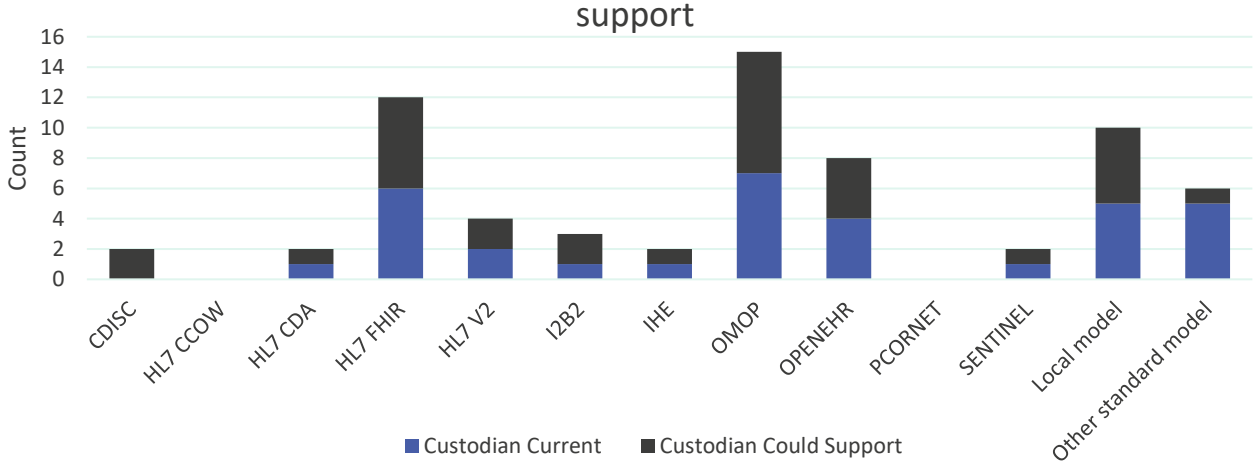


In the user survey, respondents were asked which standards, specifications and data models they currently support and potentially could support (Figure 4):



In the custodian survey, respondents were asked the same question regarding standards, specifications and data models (Figure 5):

Figure 5. Standards/data models custodians currently or could support



Amongst users, the most common four standards users can currently or could support are CDISC (31%), OMOP (29%), local models (24%) and HL7 FHIR (21%). Amongst data custodians, the most frequent standards are OMOP (38%), HL7 FHIR (30%), local models (25%) and OPENEHR (20%). OMOP and HL7 FHIR are standards which are most likely to be supported by both users and custodians. These data demonstrate some areas of discordancy between custodians and users; for example, CDISC is the most common standard for users but is only currently supported by a small minority (5%) of data custodians. Furthermore, the majority (67%) of users who support CDISC are those in the industry/pharma sector only.

Users were asked whether HDR UK's use of a core set of standards may benefit health data research in terms of volume, speed and quality.

- **85% of users** were in **support of alignment around a core set of standards** with the following considerations:
    - Users should be given resources/training and provided with examples/implementation guides
    - No one standard should be mandated since the "wrong" choice of standard for a particular use case could lose information important for research
    - Core standards will help with reproducibility, consistency in findings and transparency

In response to the question requesting any further data standards needs that HDR UK should be aware of, users highlighted the specific requirement for training around use of specific standards with custodians stating the usefulness of recommended standards/terminologies.

- User responses:
  - Users felt that training is required, or they were unsure of what data standards are
  - Need to involve frontline practitioners in the process, including those extracting data from clinical systems
  - General support for more commonly used standards such as OMOP as the general-purpose standard, however recognising that improvements should be made for specific purposes
  - Standards should be considered in relation to development and use of generic data analysis/visualisation tools

- Custodian responses:
  - Recommended core required standards and data models to support would be helpful
  - Use of open standards and associated clinical terminologies is very important
  - Custodian requirements for harmonisation and standardisation are likely to be similar and it would therefore be useful to learn from other community experiences

Users were asked how easy it would be to support additional standards compared to current practice:

Figure 6. Difficulty for users to support further standards

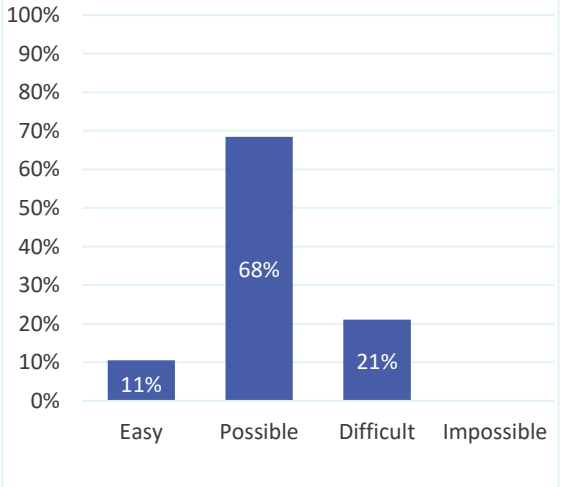
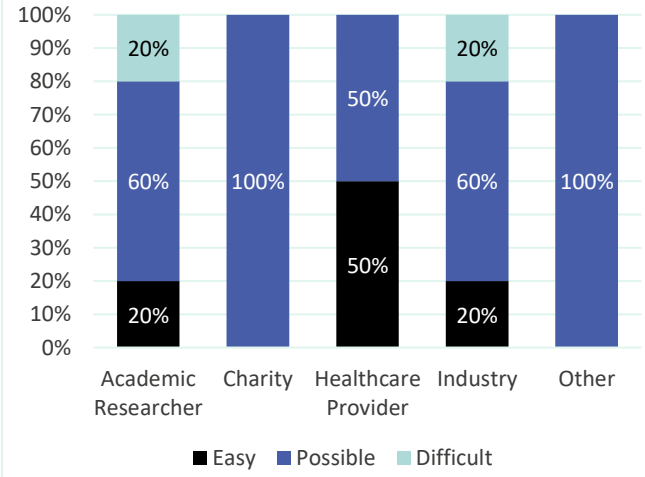


Figure 7. Breakdown of user difficulty to support further standards by sector



Of these users, the majority (79%) stated that it would be possible to support additional standards. Only 20% (from academia and industry) indicated that it would be difficult. However, when asked about specific requirements to support any such additional data standards:

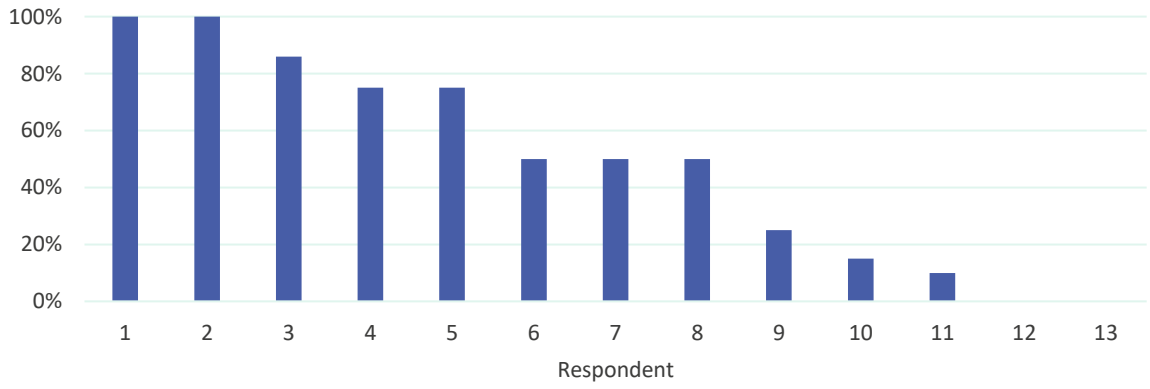
- **50% of users did not provide a response** to this question with an additional **25% users** said they **did not know** what would be needed

- Of those who responded, the following information was provided:

- Clarity regarding the reason why the additional standard is better than the existing standard
- Ensuring regulatory/external acceptance of data in such formats as well as group sharing
- Training analysts and engineers to work with such models
- Clarity regarding what is lost in mapping to different models
- Need for a well-established and preferably open-source technology base

Custodians were also asked approximately what percentage of their datasets are being mapped to the standards you currently adopt. 62% of data custodians are currently mapping ≤ 50% of all datasets to the standards they have available, and therefore many datasets are not adopting open published data standards (see Figure 8 below).

Figure 8. Percentage of datasets mapped to standards currently adopted by custodians



Metadata Specification

The first version of the metadata specification was developed to support the launch of HDR UK Innovation Gateway minimum viable product in January 2020. Version 2 of the metadata specification includes additional fields and constraints on the allowable entries for specific fields. This makes it easier for researchers to search, sort and filter the datasets when searching on the Gateway. Further improvements in the metadata involve collecting a more granular level of detail on each dataset, which provides further insight on how datasets may be linked or compared along with additional dataset profiling information. The collection of rich metadata will also support the activities of the data utility work.

The latest version of the specification for datasets onboarded onto the Gateway can be found the following HDR UK GitHub repository:

- https://github.com/HDRUK/schemata

The HDR UK dataset schema is available in the following YAML and JSON formats:

- https://hdruk.github.io/schemata/schema/dataset/latest/dataset.schema.yaml
- https://hdruk.github.io/schemata/schema/dataset/latest/dataset.schema.json

Data Utility

The term 'data utility' refers to the usefulness of a dataset for a given purpose. Effective scalable and transferable communication of the 'usefulness' of datasets requires a generalisable framework for rating and communicating how useful a dataset is, as well as the ability to assess and compare datasets from different sources at scale. Users of data have consistently fed back the difficultly to understand in advance whether a given dataset would answer their specific question, leading to lengthy access request processes, only to discover that the dataset does not meet their requirements. The significant advances in metadata discovery through the Health Data Research Innovation Gateway and the Data Utility Framework have helped to address this issue.

The framework contains five categories, separated across a range of dimensions, each of which are qualitatively evaluated to describe the characteristics of a dataset. Each dimension has a progressive series of criteria, allowing for a rating from 'Bronze' to 'Platinum' for each, provided the minimum criteria is met. The purpose is not to achieve a 'Platinum' rating across all dimensions, but to enable a user to exclude datasets that would not meet a specific threshold based on their needs. The framework enables:

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

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

Recommendations

HDR UK encourages the use of well-described open data standards within an organisation. For those custodians who have not decided on a specific model to adopt, HDR UK recommends the OMOP data model and HL7 FHIRv4 messaging standards. HDR UK recognises that specific use cases or users may require/prefer data presented in using other models and formats. There are conversion capabilities and OMOP supports i2b2 translation where appropriate. However, while OMOP and HL7 FHIRv4 facilitate interoperability, neither is suitable for all types of data. The choice of standard should therefore be selected

based on specific needs and capabilities. Although data structures are important, the community is encouraged to collaborate on approaches for mapping between standards because there is merit in the drive towards harmonisation on the data elements. HDR UK recognises the need for harmonisation and is therefore developing an MVP Common Data Elements recommendations tool to support the standardisation of elements across different studies.

HDR UK are not mandating standards and all recommendations should be interpreted alongside resource limitations and implications.

Recommendation 1: Consideration of HL7® Fast Healthcare Interoperability Resources (FHIR®) for data transit and associated APIs

HDR UK recommends the formal adoption of the HL7 FHIR[1] standard where appropriate, and the associated implementation specifications. This rationale for support is that this 1) leverages use of FHIR in the US NIH Strategic Plan for Data Science, and 2) aligns with UK NHSX/NHS Digital interoperability guidance, and 3) is included in Office for National Coordinator for Health Information (ONC) regulations for health IT providers in USA and healthcare technology vendors. We suggest use of the appropriate stable release, currently 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.[2]

[1] We note that FHIR was designed primarily as data communication specification rather than for clinical data storage or persistence. However, from a data model perspective, the FHIR model broadly follows an Entity-Attribute-Value (EAV) pattern. There is no specific 'correct' 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 specification that is good enough for the majority of purposes. We therefore recommend generally starting with the FHIR data specification, 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 numerous 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 also differ, etc. Similarly, two organisations may implement the FHIR API but with differing specifications and data elements or 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 specifications 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

The recommendation for the adoption of HL7 FHIR is supported by findings from the Data Standards surveys shared in this paper.

Recommendation 2: Consideration of OMOP as the standard data model for observation and health record data

HDR UK recommends the implementation of the OMOP data model standard for Electronic Health Record (EHR) and similar observational data in the absence of other specific requirements which may preclude this. OMOP is already reasonably widely adopted by users and custodians. The data standards surveys detailed in this paper reveal that OMOP is the most supported data standard by data custodians (38%) and second most supported by users (29%). Furthermore, common standards can be successfully be converted to other standards such as in randomised controlled clinical trials (RCTs) for example.

Note: If you are using these common architectural approaches, this allows for interoperability through the collaboration between HL7 and OHDSI with FHIR and OMOP by developing a single common data process (https://www.ohdsi.org/ohdsi-hl7-collaboration/).

Recommendation 3: Metadata Specification and Data Utility

The Health Data Research Innovation Gateway and the HDR UK metadata specification can be used to ensure interoperability to gain access to data that all meet the same minimum requirements for dataset metadata. The HDR UK metadata specification defines the minimum metadata needed for the onboarding process because it outlines the requirement for high level information that describes the datasets that are available for research and innovation across members of the UK Health Data Research Alliance and the health data research hubs. Using the specification makes it easier for researchers to search, sort and filter the datasets based on metadata rather than the data itself.

The Data Utility Framework evaluates the usefulness of the data for a given purpose and provides the ability to assess and compare datasets from different sources at scale. Therefore, we recommend using the metadata specification and Data Utility Framework for organisations to make their data more discoverable and useable.

Principles for Data Standards

The initial Principles for Data Standards were published in June 2020, following several rounds of input and consultation. We actively encourage organisations to adopt these principles when participating in any HDR UK's activities, including contracted services or activities managed by members of the HDR UK community. We also encourage their use more widely and to be treated as a broad guiding principle. Further details on the principles are provided in the following section.

1. As described in the Health Data Research UK's Principles for Participation, data should be **Findable, Accessible, Interoperable and Reusable (FAIR)**
2. 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
3. Standards that are used should be **explicitly described**[3], 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. However, a library of descriptions is for the research use of data, not for clinical care.
4. **Open** standards should be adopted wherever possible, minimising the proliferation of proprietary data standards[4]. Common standards are recommended. It is recognised that for some purposes it may be appropriate to use other specific or proprietary standards
5. Organisations should aim to maintain a **consistent, internal approach** to data standards, explicitly referencing their approach to standards in their data strategy
6. Data should be able to be used according to the principle of **without special effort** as a result of the standard used. For example, one should be able to utilise standard analytical tools to support rapid analysis of the data
7. Standards adopted should be **aligned with existing and provisional standards proposed by national and international bodies**[5] where possible, recognising that the remit and aims of HDR UK and other bodies may overlap but differ. Future international landscaping from HDR UK of common data standards will demonstrate interoperability and strengthen recommendations.
8. Ideally, standards should be **common for both research and clinical or operational uses** to optimise both research and clinical benefits of data, recognising that the primary focus of HDR UK is the research use of health data and the reasoning for using a standard should be explicitly described
9. Organisations forming part of the HDR UK network should have **established and aligned data strategies**, including how these improve the usefulness of data
10. Benefits of standards should be widely disseminated through **communication and educational**[6] events, both to researchers and the public. There should be transparency and sharing of experience of ETL processes to data standards.

[3] This is supported in the Data Standards surveys. 75% of user respondents did not answer or know what would be required to support additional standards.

[4] Custodians highlighted the importance of open standards and clinical terminologies.

[5] Users suggested that ensuring regulatory acceptance of data in such formats is required for them to support additional standards.

[6] Comments from the User survey include the need for training, resources, examples/implementation guides around data standards and the involvement of front-line practitioners.

Details of Principles

Principle 3:

"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. 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. If information is lost, the exact information lost must be understood for effective mapping to other standards. The export format should be made publicly available.

Principle 6:

"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, and where possible APIs should be open.
- Secure – adopt standards for user authentication through REST APIs with industry developed security guidelines for implementations, using access and refresh tokens.

Granularity of data should be appropriate to the need and principles adopted accordingly.

Principle 7:

"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 to prevent multiple strands of standards, for example NHSX/NHS Digital.

Notes on 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 and 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 as well as supporting GA4GH.

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 and some are mapped to other standard 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[7], 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 and syntax, in addition to terminologies used/data dictionary, can also be provided. This is trivial providing open standards are used from which appropriate mapping,

[7] 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/

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 regarding 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.

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 specific 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: Preliminary International Landscaping

Later this year HDR UK will be publishing a paper detailing the international landscaping of health data sharing initiatives, literature and legislation of data standards. HDR UK have already identified and drawn upon the following UK-wide recommendations and documentations in regards health data standards:

- NHS Digital/NHSX
    - Hold a list of Data Coordination Board (DCB) and Information Standards Board (ISB) approved standards (https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections)
    - Published Information Standards Notices (ISNs) from Jan 2017 (https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/information-standards-notices)
    - Digital Technology Assessment Criteria (DTAC) (https://www.nhsx.nhs.uk/key-tools-and-info/digital-technology-assessment-criteria-dtac/)
    - NHS England's Open policy (https://www.england.nhs.uk/digitaltechnology/connecteddigitalsystems/interoperability/open-api/)
- National Institute for Health and Care Excellence (NICE)
    - The latest quality standards from the NICE guidance and advice list (https://www.nice.org.uk/guidance/published?type=qs)
- UK GOV
    - Interoperability and open standards https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology (Section 10)
    - Open Standards Principle (https://www.gov.uk/government/publications/open-standards-principles/open-standards-principles)
- British Standards Institution (BSI)
    - www.bsigroup.com/en-GB/healthcare
    - BSOL for Healthcare

HDR UK encourages data custodians to align with WHO approved terminologies/ontologies including ICD10, ICD11, ICF, ICHI, SNOMEDCT, LOINC, DICOM and HL7, in addition to NHS OPCS codes, with reference to how such standards relate to use of OMOP and HL7 FHIRv4.

We have also identified the following initiatives/information regarding international data standards for which HDR UK recommendations can align where possible:

- European Commission
    - EU interoperability & standardisation (https://ec.europa.eu/digital-single-market/en/news/eu-activities-field-ehealth-interoperability-and-standardisation-overview)

- European Institute for Innovation through Health Data (i-HD)
    - Involvement in a DigitalHealthEurope project promoting interoperability standards
    - EHR2EDC (https://www.i-hd.eu/rd-and-collaborative-projects/ehr2edc/)
- European Bioinformatics Institute (EMBL-EBI)
    - www.ebi.ac.uk
    - ENA / EGA / EVA
    - OHDSI / EHDEN
- International Organization for Standardization (ISO)
    - https://www.iso.org/committee/54960/x/catalogue/
    - Health informatics standards ISO/TC 215
- W3C
    - www.w3.org
    - RDF for Semantic Interoperability group collaboration with W3C Healthcare and Life Sciences group
- US - National Institutes of Health (NIH) National Library of Medicine (NLM)
    - SARS-CoV-2 and COVID-19 data standards resources (CPT, LOINC, RxNorm, SNOMED CT, VSAC)
    - Downloadable content on vocabulary standards and mappings (UMLS, SNOMED CT, Mapping, RxNorm)
    - Implementation resources for standards and mappings, terminology tool and UMLS Learning Resources
    - NLM partnered with government agencies (ONC, CMS, FDA, VA, etc), HL7, SNOMED International, The Regenstrief Institute
    - LHNCBC Health Information Standards and Discovery (https://lhncbc.nlm.nih.gov/LHC-research/health-information.html)
- US Food and Drug Administration (FDA)
    - www.fda.gov
    - Data Standards Catalogue (https://www.fda.gov/media/85137/download) for submission to CBER, CDER and CDRH.
    - NDC API
- Australian Institute of Health and Welfare (AIHW)
    - Health sector standards in the metadata online registry (METeOR) repository (https://meteor.aihw.gov.au/content/index.phtml/itemId/181245)
    - Metadata standards approval from the National Health Data and Information Standards Committee (NHDISC)

HDR UK plan to develop the international landscaping work further by highlighting the alignment with policies outside the UK.