UK Health Data Research Alliance

Recommendations for Data Standards in Health Data Research

White Paper

November 2021

UK Health Data Research Alliance

Contents

UK Health Data Research Alliance

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 builds on these principles and proposes recommendations for data standards. It encourages improvements in data usefulness and usability with the primary focus to benefit research through the provision of clear guidance and recommendations and has been developed with input from the health data community, as well as patients and the public. This document does not cover ontology standards and web standards.

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

- There was a greater level of stated expertise in data standards in industry compared to academia. The majority (around two thirds) 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
- Almost all (around 90%) 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 using open standards and clinical terminologies
- Currently data users are using a wide range of data standards and data models, with greatest data model alignment around Clinical Data Interchange Standards Consortium (CDISC), Observational Medical Outcomes Partnership (OMOP) Common Data Models (CDM)
- Data custodians also currently support a wide range of data standards and data models, with OMOP CDM and HL7 FHIR interoperability specifications the most frequently supported

Recommendations: For organisations and companies considering establishing research data standards, such as data models and messaging standards, HDR UK suggests the initial consideration for use of OMOP CDM and HL7 FHIRv4 as first models/specifications. These could 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.

1. Consideration of HL7® Fast Healthcare Interoperability Resources (FHIR®) for data transit and associated APIs
2. Consideration of OMOP Common Data Model as the standard data model for observational health record data

However, it is recognised that specific use cases may have specific requirements and for organisations already using different standards or models, HDR UK recognises the practical resourcing, cost and potential information loss implications associated with transition to OMOP and HL7 FHIRv4. Ultimately, the appropriate standards should be selected to support the specific use cases and capabilities.

UK Health Data Research Alliance

3. Metadata Specifications and Data Utility: To help researchers find and understand the usefulness of a dataset for research, we recommend using the Health Data Research Alliance common metadata specification and the Data Utility Framework.

The use of structured metadata makes it easy for researchers to discover data and to search, sort and filter the datasets when using the Health Data Research Innovation Gateway to access data. The Data Utility Framework, which is a tool rather than a data standard, evaluates the usefulness of the data for a given purpose and provides the ability to assess and compare datasets from different sources at scale, including standards and data models used. It provides a means by which a user can find data that is likely to meet their needs.

International positioning for data standards: A high level review of health data standards used across the G7 countries and Australia, with broadly comparable health care systems to the UK, is provided to allow alignment with HDR UK for International use where possible.

For information on definitions and the main data standards and models, please refer to Appendix 1 and Appendix 2.

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 health research directly in the first instance as opposed to clinical care, which may in turn lead to benefits at the frontline of clinical care.

HDR UK convene the UK Health Data Research Alliance, an alliance of leading health, care and research organisations united to establish best practice around the ethical use of UK health data for research and innovation at scale. In June 2020, as part of work across the Alliance, 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 feedback was received from almost 50 individuals across more than 30 organisations.

This white 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 that took place in 2021 and discussions across the Alliance's Data Officers' Group. As part of these consultations, HDR UK particularly valued feedback from organisations on further lessons or challenges raised while implementing the suggested standards.

HDR UK recognise and encourage the need to use data standards for consistency and harmonisation to bring data together for comparison which enables discoveries to improve people's lives faster. The use of data standards promotes harmonisation for advanced analytics of large-scale data and HDR UK would like

UK Health Data Research Alliance

to encourage adoption of the recommendations through the UK Health Data Research Alliance, and in particular the Health Data Research Hubs. We also welcome the adoption by partner organisations within the UK, for example national bodies to shape system-wide requirements.

The paper is part of a suite of 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.

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 responses provide an overview of the current positioning 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, forty-nine users and data custodians responded to the data standards surveys.

Summary of Findings

User and custodian surveys were conducted with academic researchers, charities, data custodians, healthcare providers, life science companies and AI and 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) Common Data Model (CDM).
- Data custodians also currently support a wide range of data standards, with OMOP and HL7 FHIR the most frequently supported.

UK Health Data Research Alliance

Results

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 only basic or no degree of data standards expertise with no respondents claiming expert knowledge (Figure 1). In contrast, 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 2).

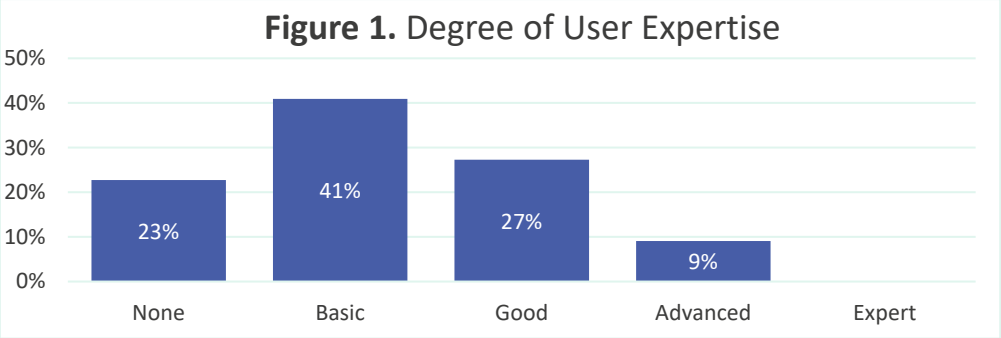


Figure 1. Degree of User Expertise

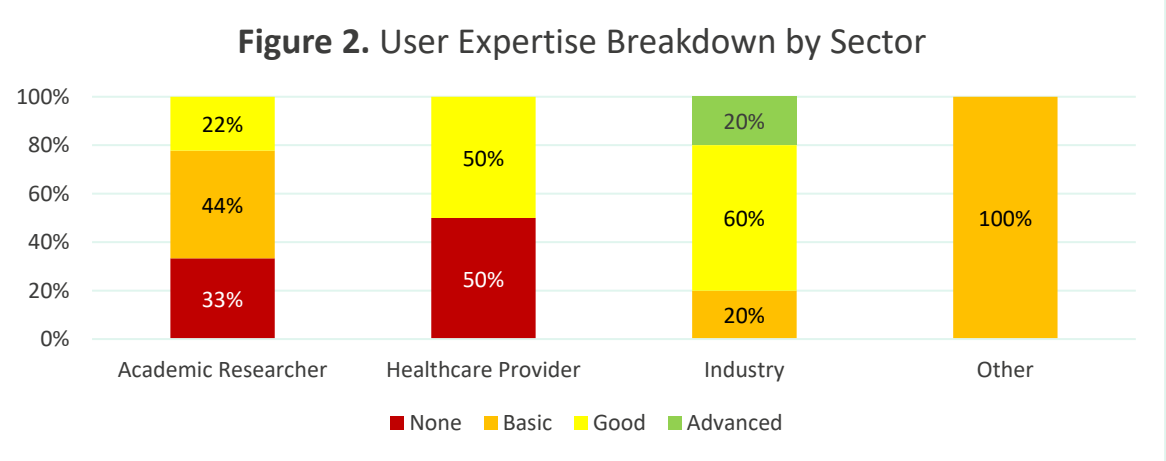
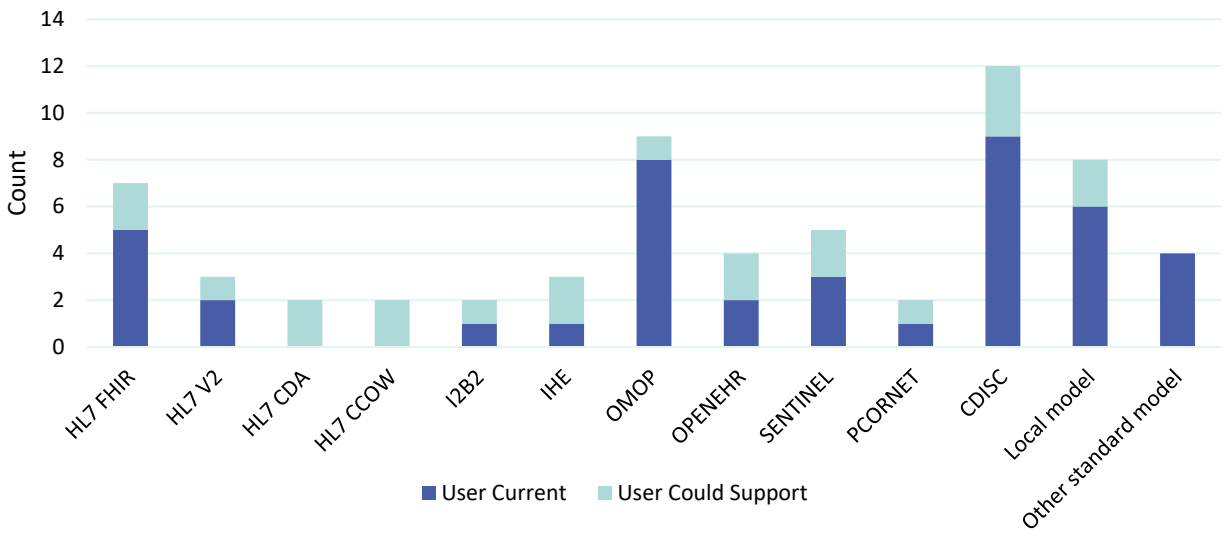


Figure 2. User Expertise Breakdown by Sector

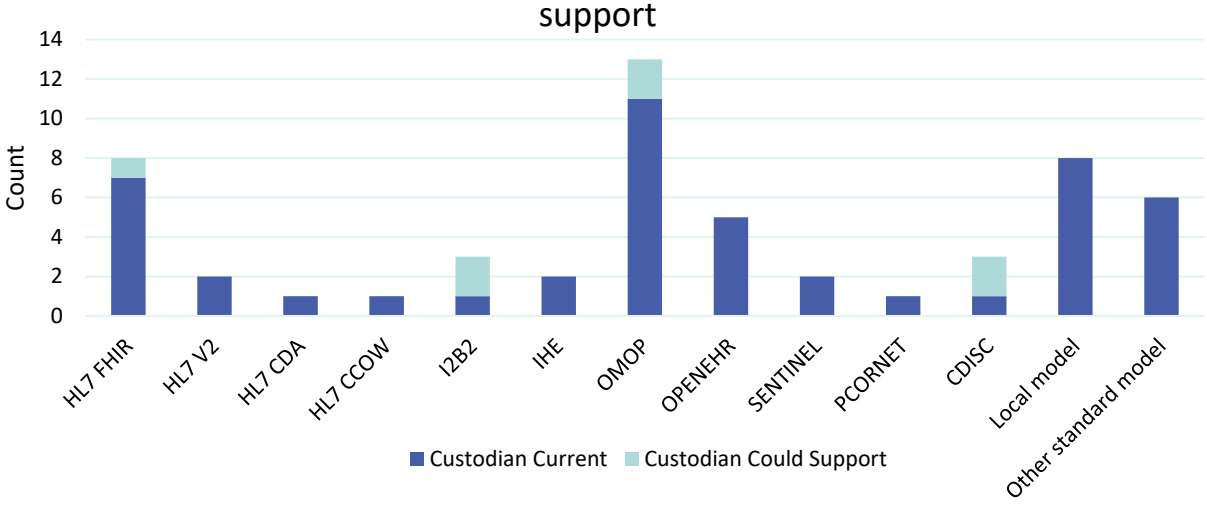
In the user survey, respondents were asked which standards, specifications and data models they currently support and potentially could support (Figure 3):

Figure 3. Standards/data models users currently and could support



In the custodian survey, respondents were asked the same question regarding standards, specifications and data models (Figure 4):

Figure 4. Standards/data models custodians currently or could support



Amongst users, the most common standards users currently or could support are CDISC (19%), OMOP (14%), local models (13%) and HL7 FHIR (11%). Amongst data custodians, the most frequent standards supported are OMOP (24%), HL7 FHIR (15%), local models (15%) and other standard models (11%). OMOP and HL7 FHIR are data model and interoperability standards which are most likely to be supported by both users and custodians but are far from universally supported or used by the community at present. These data further demonstrate discordancy between custodians and users; for example, CDISC is the most

common standard for these users (which may be biased due to strong pharma / industry representation) but is only currently supported by a small minority (5%) of data custodians. (CDISC is the Clinical Data Interchange Standards Consortium that develop standards in collaboration with experts in pharmaceutical organisations and aims to improve standards for pharmaceutical organisations internationally).

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 single standard should be mandated since the "wrong" choice of standard for a particular use case could lose information important for research
    - Use of several 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 modifications may be required 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
    - Attention needs to be given to data quality and how we can systematically measure this. Standardised data quality measures would be very important

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

Figure 5. Difficulty for users to support further standards

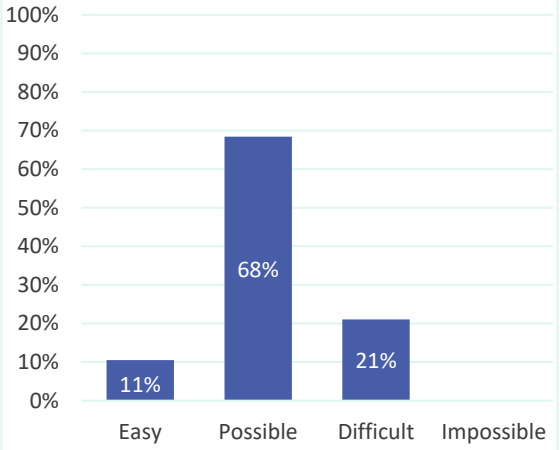
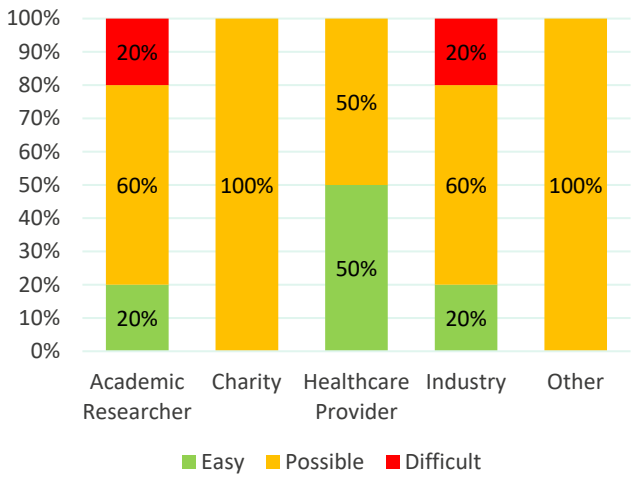


Figure 6. Breakdown of user difficulty to support further standards by sector



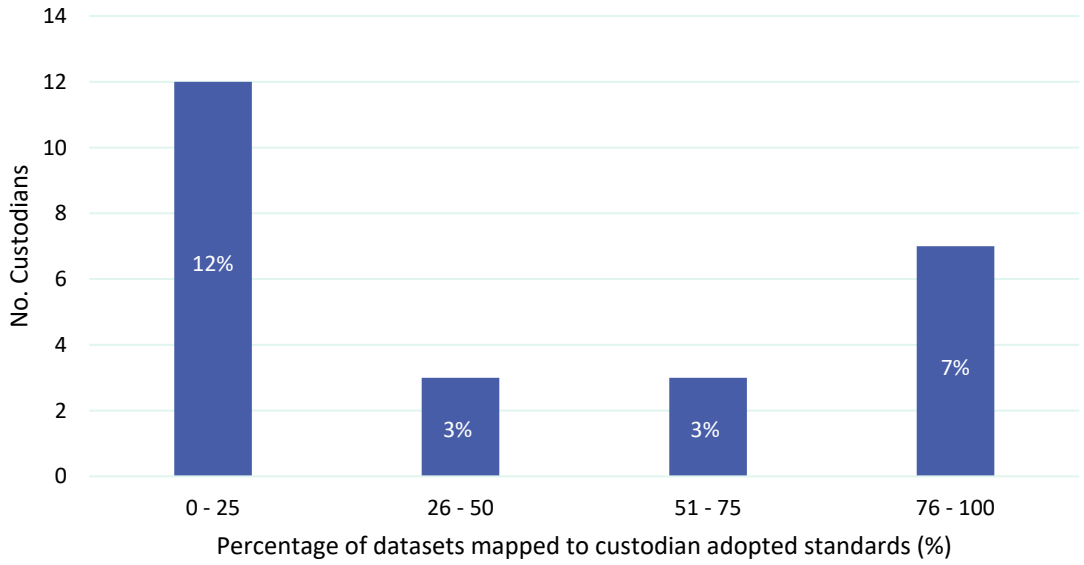
Of these users, the majority (79%) stated that it would be possible to support additional standards. Only 20% (from academia and industry) indicated that it would be difficult. However, when asked about specific requirements to support any such additional data standards:

- **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 mapped to the standards currently supported. Most datasets for the majority of custodians are not currently mapped to open published data standards (see Figure 7 below).

Figure 7. Distribution 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 a 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

UK Health Data Research Alliance

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. An animation of the data utility evaluation is available here.

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 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 rather 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 adopted a specific data standard/model, **HDR UK broadly recommends consideration of the OMOP data model and HL7 FHIRv4 messaging specification standard**. HDR UK recognises that specific use cases or users may require/prefer data presented in using other models and formats. While OMOP and HL7 FHIRv4 facilitate dataset interoperability, neither is suitable for all types of data and purposes hence 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 value in harmonisation of data elements and models, especially in relation to developments such as federated analysis. HDR UK is developing an MVP Common Data Elements tool to support the standardisation of data elements across different studies.

HDR UK is not mandating use of specific standards and all recommendations should be interpreted alongside resource limitations and implications.

UK Health Data Research Alliance

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]

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

HL7 FHIR is a common standard for data exchange in health care. Fast Healthcare Interoperability Resources (FHIR) combines the best features of the different versions of HL7 but focuses on easy implementation and uses web standards such as XML, JSON, HTTP, OAuth. FHIR solves the challenge of growing variability of diverse databases in which more fields are added over time, which in turn often relies on alternative is relying on custom extensions. FHIR defines a simple framework for extending the existing resources, in which all systems can read all resources, but applications can add more control and meaning using profiles. Each resource carries a human-readable text representation using html which is particularly important for complex clinical information where many systems take a simple textual/document-based approach.

---

[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

UK Health Data Research Alliance

Recommendation 2: Consideration of OMOP as the standard data model for observational 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.

Electronic Health Records that capture routine clinical practice information when care is taking place and billing databases are two examples of disparate databases. It is time-consuming for researchers to harmonize disparate databases before analyses before analyses can be carried out across all datasets. Converting data to a common data model is a solution that allows the same data specification to be shared across disparate observational datasets which all differ in purpose and designs.

Note: With such 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 (e.g., 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.

UK Health Data Research Alliance

International Positioning

Scoping of recommended standards across G7 countries and Australia, HL7 FHIR is used for Health information exchange and messaging. OMOP is used for collaboration, especially amongst the OHSDI APAC community, often used for mapping electronic medical records as part of medical insight and research data hubs, and/or for mapping and collaboration of datasets. OMOP and FHIR appear reasonable choices for recommended standards whilst aligning with international developments, but no common ubiquitous standards are currently in widespread use.

> "A truly connected, interoperable and sustainable Common European Health Data Space is a precondition to unlock the potential of health data in the EU." (Digitaleurope.org)

The COVID-19 pandemic has highlighted value of the ability to share and validate health data internationally to support the development of health policy, delivery of care, and regulatory and surveillance activities. At the G7 Summit in May 2021, there was a call for a more systematic approach to data capture, standards, sharing and analysis to better prepare for pandemics, support clinical trials, data regulation and to support accelerated deployment of vaccines in the global crisis (Digitalhealth.et). At the G7 Health Ministers Meeting in June 2021 it was agreed to improve preparedness for and response to health emergencies, while recognising a need for a standards-based and minimum data sets for international use (DHSC Policy Paper, June 2021).

To produce an overview of current international use of data standards and an assessment of data model harmonisation between countries an internet search of papers and public information was performed to include G7 countries and Australia (with broadly similar healthcare systems and health data research outputs to the UK).

UK Health Data Research Alliance

Figure 8. Countries investigated as part of the HDR UK international landscaping.

In general, data standards in clinical settings receive greater attention than the use of standards to manage data in clinical research. Most of the standards highlighted below are used for clinical practice, with particular data models for research purposes. We focus on potential alignments between nations and opportunities for strategic co-development of data standards for use in the UK.

The proposed recommended standards, OMOP and HL7 FHIR, are in use internationally across G7 countries and Australia, but there is no common consensus internationally and many terminologies used require national modifications. An overview of the standards in use will now be described according to categories (please note the UK is not included in the information below).

Terminology covers the classification of diseases, codes used by medical services. Standards for terminology are used for both *clinical data for research purposes.*

From the information available, countries have their own standards or variants of standards used in other countries. However, the following are most widely used internationally.

- ICD10/11
- SNOMED-CT

Messaging/interoperability refers to the communication of clinical messages of EHR for research and clinical purposes.

- HL7 FHIR is the main data exchange standard

Data models can be used for alignment and supporting analysis of research and clinical trial data between organisations and allow conversion of routine clinical data to specific data elements for research purposes. The most commonly used models are:

- CDISC
- OMOP

Privacy and security standards to protect and use medical information, the equivalent being GDPR and Data Protection legislation in the UK. In most of the selected countries privacy and security is simply covered by legal Acts and Bills on personal data and protection. Therefore, most countries have individual and specific legal frameworks leading to their own customised standards.

Broadly, HDR UK therefore encourages data custodians to align with the most commonly used international and WHO approved terminologies and ontologies including ICD10, ICD11, SNOMED CT, LOINC, DICOM, CDIS, OMOP and HL7 FHIR with reference to how such standards may be related.

UK Health Data Research Alliance

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 of 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

---

[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.

UK Health Data Research Alliance

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.

Details of Specific 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.

---

[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. The need for communication and educational resources is shown by the majority of health data users saying they have basic or no data standards expertise.

UK Health Data Research Alliance

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.

UK Health Data Research Alliance

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 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, 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

[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/

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.

UK Health Data Research Alliance

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 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
API	Application programming interface (communication protocol between different software elements)
Clinical classification	System for assigning clinical data items to categories
Clinical terminology	Collection of terms used in a specific clinical setting / scenario
Data controller	Controls data usage and has data protection responsibility
Data dictionary	Information describing the contents, format, and structure of a specific database including history and changes and context
Data element	Specific unit of data within a dataset that has precise meaning
Data format	Organisation of data according to preset specifications
Data mapping	Describing the relationship between data elements in different data models
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 processor	Uses or processes the data on behalf of the data controller
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
Data schema	Description of how data is organised in relation to how a data repository is constructed
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 structure	Collection of data values, relationships and functions that can be applied to the data
Dataset	Collection of related data elements
Information model	Representation of concepts and relationships for a particular context
Information standard	Rules by which information is described and recorded
Interoperability	Ability to function with systems other than the index system
Metadata	Set of data providing information about other data, either at dataset level or value level

UK Health Data Research Alliance

Ontology	Description of entities and how they are subdivided and related
Reference data	Known dataset that defines permissible values to be used or for comparative profiling
Specification	Detailed description of components required for a specific function/activity
Standard	Technical, functional, or performance-based rule, condition, requirement or specification that stipulates instructions, fields, codes, data, materials, characteristics, or actions for common usage.
Syntax	Set of rules or structure of statements
Value set	Subset of specific terms for particular use cases

UK Health Data Research Alliance

Appendix 2: Data Standard/Model Definitions

Data Standard/Model	Definition
CDISC	Clinical Data Interchange Standards Consortium (CDISC) that develop standards in collaboration with experts in pharmaceutical organisations and aims to improve standards for pharmaceutical organisations internationally.
HL7 CCOW	HL7 Clinical Context Object Workgroup (CCOW) defines standards enabling the visual integration of healthcare applications.
HL7 CDA	HL7 Clinical Document Architecture (CDA) is a document markup standard that specifies the structure and semantics of "clinical documents" for the purpose of exchange between healthcare providers and patients.
HL7 FHIR	Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) is a standard for health care data exchange, published by HL7.
HL7 V2	HL7 Version 2 is an application protocol for electronic data exchange in healthcare.
i2b2	Informatics for Integrating Biology & the Bedside (i2b2) is a self-service "cohort discovery tool" that allows you to explore and query clinical data that has been deidentified and aggregated.
ICD-10	International Classification of Diseases (ICD) tenth revision (10) is the world's standard tool to capture mortality and morbidity data.
OMOP CDM	Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) allows for the systematic analysis of disparate observational databases to transform data in the databases into a common format.
openEHR	openEHR is the name of a technology for e-health, consisting of open specifications, clinical models and software that can be used to create standards, and build information and interoperability solutions for healthcare.
RxNorm	RxNorm is normalised naming system for generic and branded drugs. It is also a tool for supporting semantic interoperation between drug terminologies and pharmacy knowledge base systems produced by the National Library of Medicine (NLM).
Sentinel	The United States Food and Drug Administration (FDA) Sentinel Common Data Model (SCDM) is a standard data structure which implements standard queries across the Sentinel Distributed Database (SDD).
SNOMED CT	Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) is a structured clinical vocabulary for use in an electronic health record and is most comprehensive clinical terminology in use around the world.

UK Health Data Research Alliance

Appendix 3: Preliminary International Positioning

Prior to the work on international landscaping health data standards and sharing initiatives, HDR UK identified and drawn upon the following UK-wide recommendations and documentations in regards health data standards. These are as follow:

- 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

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)

UK Health Data Research Alliance

- 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)Heading 1