EOSC-Life: Building a digital space for the life sciences

D4.4 – Report on data standards for observational and interventional studies, and interoperability between healthcare and research data

WP4 – Policies, specifications and tools for the management of data for biological and medical research
Lead Beneficiary: ECRIN, BBMRI
WP leader: Jacques Demotes and Michaela Mayrhofer
Contributing partner(s): ECRIN, EATRIS, ERINHA, Lygature, VHIR

Authors of this deliverable**: Steve Canham, Christian Ohmann, Jan-Willem Boiten**
With contributions from: **Maria Panagiotopoulou** (ECRIN)**, Nigel Hughes** (EHDEN/Janssen)**, Romain David** (ERINHA)**, Alex Sanchez Pla** (VHIR)**, Lauren Maxwell** (ReCoDID/Heidelberg University Hospital)**, Jozef Aerts** (XML4Pharma)**, Rhonda Facile** (CDISC)**, Nicolas Griffon** (APHP)**, Gary Saunders** (EATRIS)**, Kees van Bochove** (EHDEN/The Hyve)**, Jonathan Ewbank** (ERINHA)

Contractual delivery date: **31 August 2021**
Actual delivery date: **30 December 2021**
H2020-INFRAEOSC-2018-2

Grant agreement no. 824087
Horizon 2020
Type of action: RIA

Table of Contents

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Executive Summary

The scope of the report is discussed, and clarified as an examination of data standards and the interoperability of data both within and between healthcare data (including that used for observational research) and interventional research data (largely from clinical trials).

The data standards in use within healthcare data are examined from the perspectives of outcome measures, syntactic and transport standards, semantic standards and metadata standards. The same is then done for standards within interventional research.

The practicalities of interoperability, in particular with respect to making healthcare data more like interventional research data, are explored – again under the headings of outcome measures, syntactic and transport standards, semantic standards and metadata standards.

The major finding is of a current semantic incompatibility between healthcare and interventional research data standards – i.e. they tend to use different vocabularies and concepts.

A conclusion summarises the report's findings and discusses possible actions to improve data interoperability in the short, medium and long term.

Project Objectives

This deliverable has contributed to the following objectives:

a. Establish EOSC-Life by publishing FAIR life science data resources for cloud use

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Detailed Report on the Deliverable

1. Introduction and Terminology

1.1 Studies and Data

The title provided for the report differentiates both ‘observational and interventional studies’, and ‘healthcare and research data’. The following definitions and discussion, taken from the literature when possible, are intended to clarify how these potentially overlapping categories have been interpreted.

Interventional studies: Interventional research designs are described by Thiese [1], as “those where the researcher intervenes at some point throughout the study”. By far the most common type of interventional study within clinical research is the **clinical trial**, based on a protocol that assigns participants to one of two or more pre-specified interventions. As an experiment on humans, a clinical trial requires both ethical approval and the informed consent of all participants.

Observational studies: Observational research covers such a wide range of different methods that in some ways it is easier to define it in the negative, as ‘non-interventional research’. Thiese takes much the same approach: “Observational studies, …, are those where the investigator is not acting upon study participants, but instead observing natural relationships between factors and outcomes” [1].

Observational research includes cross-sectional studies, case-control studies, retrospective and prospective cohort studies, testing and screening evaluations, case studies and case series. Their scope can range from a single individual to an entire population. Depending on the nature of the study and the jurisdiction in which it is carried out, observational research may or may not be based on a protocol, and may or may not require the consent of participants.

A strict interpretation of the Thiese definitions makes survey based research, or the questionnaire components of observational research designs, difficult to categorise. A questionnaire is certainly an intervention, but it is the same intervention for all participants and is designed only to collect data, not to affect the treatment or influence the experience of the study participant in any way. In the context of this report, surveys and questionnaires are included within observational research, albeit of a slightly more pro-active variety than a ‘purely’ observational study.

Real world data (RWD): This is defined here as data primarily designed to support clinical and managerial decision making and record keeping within normal health care practice, and is generated by that practice on a routine basis. It includes the data in electronic health records (EHRs), in both primary and secondary care, and in patient registries and claims databases. It also includes the increasing volumes of data generated by wearable devices, in both domestic and healthcare settings.

Unfortunately, conceptions of ‘real world data’ seem to vary considerably – one study of the literature, coupled with interviews, produced 38 different definitions, split into 4 categories [2]. As that paper makes clear, and as is shown in figure 1, a spectrum of health-related data exists. At one extreme is the highly controlled and specialised output of a classical randomised clinical trial

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

(RCT), at the other is the entirely non-controlled and ‘routine’ data within electronic health records (EHRs). In between, other research types, such as a pragmatic clinical trial, may make use of data points that are routinely available, as well as research specific ones, while a survey may be used to supplement the routine data points of an observational study.

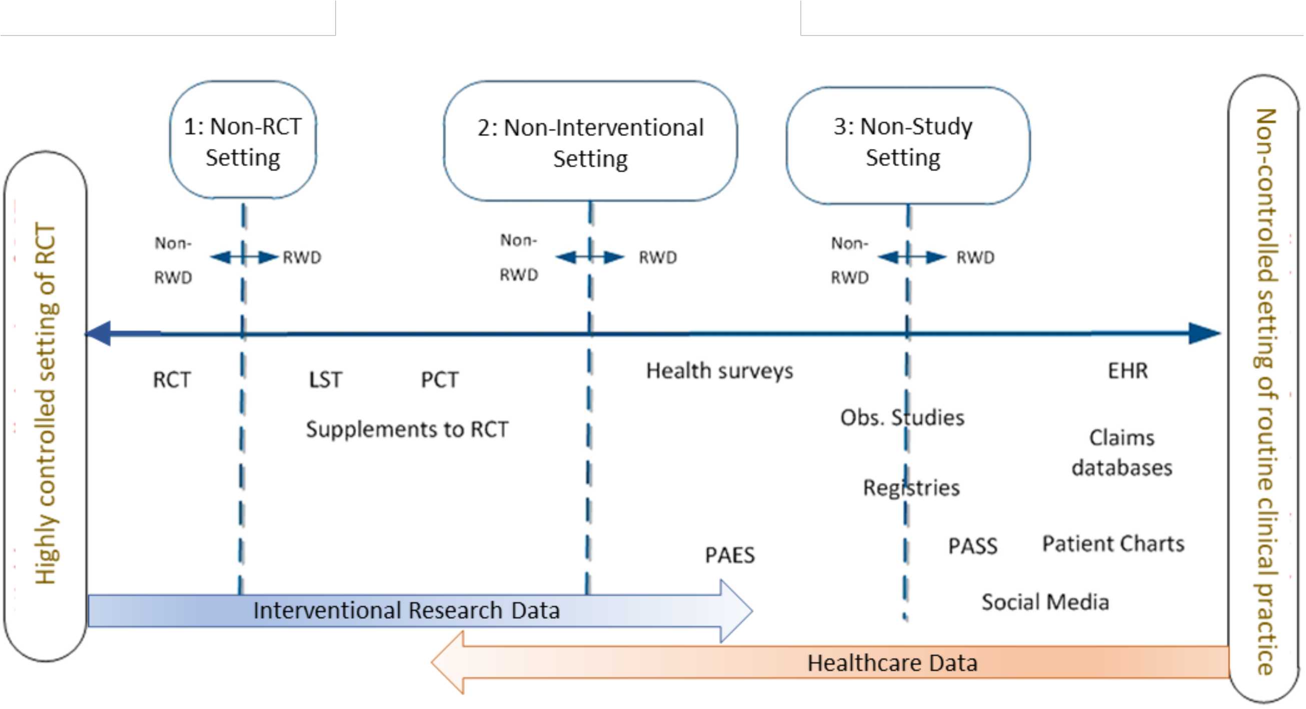


Figure 1: Different types of data sources and the resultant data. Adapted from Makady et al, 2017 [2]. ***Legend: RWD** = Real World Data; **RCT** = Randomised controlled trial; **LST** = Large simple trial; **PCT** = Pragmatic clinical trial; **PAES** = Post authorisation efficacy study; **PASS** = Post-authorisation safety studies; **Obs. Studies** = Observational studies; **EHR** = Electronic health record.*

The review found that different people put the distinction between RWD and non-RWD at different points along this continuum (for example the definition given above puts it at point 3, but others draw the distinction at points 1 and 2). There seems to be a general confusion between the terms ‘observational data’ and ‘real world data’, with people often conflating or overlapping the two categories, whether or not any research activity is involved, but with different boundary cases defined for each.

It is suggested that the most useful distinction for the purpose of this report is less about the type of activity that generates the data, and more about how and where that data is *defined*. In particular:

- The data items in interventional research are – in most cases – defined by the experimenters themselves. They are intended to answer a small and specific set of questions, are subject to regular quality checks, and within any one study they are highly consistent and controlled.
- The data items in real world data are defined by the ‘context’ – they have arisen to meet the needs of medical record keeping and decision making, and associated management and

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

resource decisions. They are not linked to any particular question, though will probably reflect the specialism where they are derived. They are not normally monitored or corrected in any systematic way, and hence often contain inconsistent or missing data.

- The data items in observational research are (for the most part) the same as those used in RWD. In some observational research the data may be supplemented by research specific questions, e.g. in questionnaires, and it may be more complete than in RWD, but fundamentally it consists of data items defined within the healthcare context, assembled to meet a research need.

On that basis the following terms are suggested to most usefully differentiate the two main forms of data:

A) **Interventional research data** (or as a synonym, c**linical trial data**): data defined by researchers to use within interventional research studies, chiefly clinical trials. The term ‘research data’ is not specific enough to be clear.

B) **Healthcare data** (or as a synonym, **RWD**): data created within the healthcare context and used within that context (as RWD) but *also* exploited for observational research. The term ‘observational data’ is not used, as it is potentially confusing.

These different types of data are indicated in figure 1 by the blue and orange arrows. The arrows overlap, because some types of research make use of both data types. Usefully, this dichotomy of data types also matches the major ‘ecosystems' of data standards, one of which is found almost exclusively within interventional research, and the other almost exclusively within healthcare and observational research.

Real World Evidence (RWE): is simply the use of RWD, after an appropriate analysis, in support of a particular scientific or medical hypothesis, or a regulatory decision.

Metadata: is used in this report in the broad sense defined by ISO11179 (as reported in [2]). That is, rather than the traditional definition (simply ‘data about data’), it is ‘data about a digital object’. A digital object is anything that is an independent electronic entity, normally a file stored within a computer’s storage system. In the case of clinical research this certainly includes datasets, but it could also be a document, such as a protocol, statistical analysis plan, journal article, coding manual, etc. This allows a distinction between:

> **Discovery Metadata**: Applies to any digital object, and provides information about the object – its subject matter, authors, dates, version, location, access process etc. This information is critical for supporting the findability and accessibility of the object, providing the ‘FA’ in making the object ‘FAIR’.
>
> **Descriptive Metadata:** Applies mostly to data, and is the detailed description of each data item (name, type, definition, etc.), that provides a full understanding of the data itself. This information is critical in enabling the interoperability and re-usability of the data, the ‘IR’ in making the object ‘FAIR’.

1.2 Interoperability

Interoperability: is defined here as in the original description of the FAIR principles [3]:

> ‘The ability of data or tools from non-cooperating resources to integrate or work together with minimal effort’.

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

That ability, however, can operate at a variety of different levels, as described within EOSC's interoperability framework [4], which distinguishes technical, syntactic, semantic, organisational and legal interoperability. In this document the emphasis is on:

> **Syntactic interoperability:** When systems use data *structured* in the same way, along with common (or easily transformed) data formats and communication protocols, thereby permitting a relatively straightforward transfer of data from one system to another.
>
> **Semantic interoperability:** When the precise *meaning* of exchanged data and information is preserved and understood. Semantic interoperability requires a shared vocabulary and ontology within any particular area, or at least ontologies and vocabularies that can be accurately mapped to each other.

1.3 Data Standards

Data standards are rules and conventions that can be applied when designing data systems, to promote consistency within and similarity between datasets. They make it easier and quicker to design a dataset, and understand those designed by others, and make it much easier to compare or aggregate data from different sources. The use of data standards is therefore fundamental to making data interoperable.

Five different types of data standards are recognised within this report, and are used to structure discussions in later chapters.

Standards for outcomes measures: No amount of re-structuring and data manipulation will make data interoperable if the data have been used to measure different things. Decisions on data content are therefore the most fundamental type of standard that can be applied. Within both interventional research and healthcare, much data will simply describe the experience of the participant or patient: for example, their presenting symptoms, the assessments undertaken, their test results and the treatments they received. But **outcome measures** are also (or should be) a key component of the data – the data elements that are selected to act as indicators of effectiveness, efficiency and safety.

Whether assessing actions taken in a healthcare or a research context, using the same outcome measures, in effect using the same definitions of 'success', is key to being able to compare or aggregate data from different sources. Accordingly, initiatives to promote common outcome measures are discussed in later sections.

Standards for data structures and syntax: If data is structured in the same way, for example split into the same tables, and organised in the same way, with variables having the same names and definitions, it is obviously much easier to compare or aggregate that data.

The major data standards 'systems' devote much effort to this structural or syntactic aspect of data interoperability, though given the scale and complexity of health and biomedical data, and its tendency to evolve, developing comprehensive systems for consistently organising all of that data is an extremely challenging task. Systems are therefore usually a mix of fixed and user-defined elements, with the process for creating and documenting the user-defined components tightly controlled.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Standards for data transport: Transport standards are concerned with how data is structured when it is transferred from one system to another. They are therefore also concerned with data structure and syntax, except that the context is data transfer rather than data storage, and the data is often transferred as discrete 'packets' of information rather than as a comprehensive dataset.

In the discussions in later sections, data transport standards are discussed along with the standards for structuring data at rest, as the main determinants of syntactic interoperability.

Standards for data semantics: In any dataset many variables are *categorised or coded*, using a constrained list of terms for each data item, a **controlled terminology** or CT. The difficulty is that in many cases the data item can be categorised or coded in one of several ways, using different CTs – e.g. a diagnosis can be expressed using ICD 10 or 11, MedDRA, or SNOMED, in each of those cases at one of various points within a hierarchical system. CTs may also be constructed locally, e.g. different ways of categorising reasons for a participant's withdrawal from a trial, or a patient's discharge from secondary care, or selected from different sources, e.g. different systems for staging tumours, and even when they are relatively straightforward CTs can be coded differently, e.g. 'sex' as male/unknown/female, or M/U/F, or 0, 1, 2, or +1, 0, -1.

Different CTs represent a significant barrier to interoperability because a CT is much more than a simple list of terms or codes – it represents how a particular idea is *conceptualised* within the system (e.g. drugs classified by chemical structure, by action on the body, or by the illnesses they are used to treat), and / or how granular the data will be (e.g. adverse events classified using MedDRA or the Common Toxicity Criteria grades). CTs also vary considerably in their sophistication – some are simple lists, some are hierarchies, and some are full ontologies, with data about the inter-relationships between entities also included. 'Mapping' between CTs is therefore often difficult and inexact, and using different CTs can have a profound effect on the semantic interoperability of datasets.

Standards for metadata: Data needs associated descriptive metadata if its contents are to be understood - a detailed item by item catalogue that, ideally, gives each item's definition as well as its name, code, type, possible values etc. That metadata should be machine readable as well as understandable by humans, to allow it to be quickly searched and datasets relevant to any particular task more easily discovered, which demands a consistent structure. Standards for metadata, when they exist, are therefore also discussed in the later sections.

1.4 Scope of the Report

The second phrase of the report title ('Data standards for observational and interventional studies, *and interoperability between healthcare and research data'*) could be read as implying that *within* healthcare, or *within* interventional research, data was already organised in a uniform way, with the sole remaining issue being the interface between the two systems. As the report makes clear, this is very far from the case, and in fact there are three areas where data interoperability could and should be improved, largely by the increased use of data standards:

- Within healthcare, to better support both observational research and the further development of self-monitoring 'learning' health systems.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

- Within interventional research, in particular to make it easier to aggregate and compare data from non-commercial research with that from the pharmaceutical industry
- Between healthcare and interventional research data, to support the growing use of healthcare data within activities traditionally supported by interventional research, including regulatory decision making.

With regards to the last point, the direction of data flow is important. 'Interoperability' normally implies that data flow in both directions is equally likely or equally important, that – in this context – there is as much interest in transferring research-derived data to the clinical area, to be integrated with RWD, as there is in using the real world 'context defined' data to replace or complement the data from clinical trials. This is not the case.

While the *conclusions* drawn from interventional research are obviously fed back into healthcare, it is relatively unusual for the data itself to be fed back, at least while the research is going on. If data is returned to the clinical area it tends to be in the form of individual data points rather than datasets. For example, scores on the Hamilton Depression Scale that indicate suicidal ideation can lead to alerts being sent to a trial participant's physician, while genetic studies may contribute pharmacogenetic data that can inform prescribing decisions, (if that information is not already obtained on a routine basis). Because only isolated data points are involved interoperability is not really an issue.

Conversely, as figure 2 illustrates and the next chapter makes clear, considerable time, effort and money have gone into investigating and demonstrating how RWD datasets can be used within research, both within observational research and surveillance programs, and as a supplement to interventional research data, including within regulatory decision making.

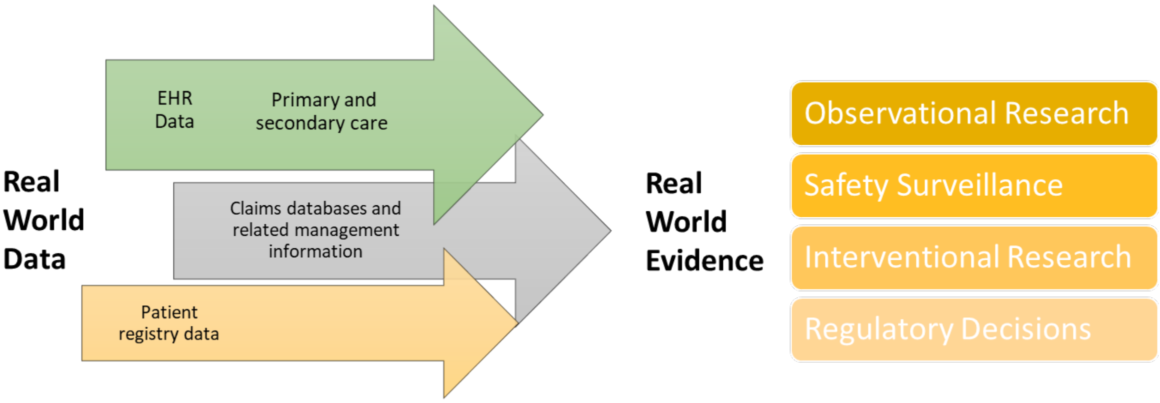


Figure 2: The main data flows in practice, the transformation of RWD to RWE.

There are a large number of issues that can affect the *utility* of real world data used for research, for example the completeness and quality of the data and metadata, the provision of an adequate ethical and legal framework for data transfer and/or reuse (e.g. inclusion of broad consent for future use), and adequate de-identification measures being in place to protect privacy. Although these issues are mentioned, for the most part this report is focused on *technical* interoperability and data management.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

The focus is also on real world data as directly obtained from healthcare institutions, i.e. as entered into hospital or primary care electronic health records (EHRs), and then, in some cases, transferred to patient registries, or claims databases. A number of countries have, or are developing, national repositories of de-identified healthcare data, often linked to research and / or socio-economic data, that can be made available for research purposes – for instance for hypothesis generation, or for epidemiological research (e.g. [5 - 7]).

These are important resources, but each will have their own data sources, formats, access procedures and capabilities, and, in many cases, will already have pre-processed the data obtained from healthcare records. They merit their own report (Deliverable 4.5: Public database inventorying the national health databases and registries and describing their access procedures for reuse for research purposes), and are therefore not covered here.

Given all of the points discussed in this introduction, the scope of the present report, reflected in the title of the chapters that follow, can be summarised as:

- The use of RWD in research.
- Data standards and interoperability within healthcare data and observational research.
- Data standards and interoperability within interventional research data.
- Interoperability issues when integrating healthcare data with interventional research data.

2. The Use of Real World Data in Research

2.1 Promises

A 2006 paper from the eClinical Forum estimated that, at that time, 20-25% of US hospitals used EHR, with the corresponding figure in different European countries ranging from 20 to 90%. Use of electronic remote data collection in clinical trials (eRDC) was estimated at 27-30%.

It was clear even then, however, despite the obvious differences between the two types of data, that having routine medical data in an easily transmissible electronic form had huge potential value for clinical research, while making participation in that research less burdensome for healthcare staff:

> “The vision is for shared systems and processes that would allow the use of patient electronic medical data for clinical research in a way that meets data protection, regulatory, and ethical research requirements and thereby minimizes the challenges of clinical research for healthcare professionals.” [8]

Fifteen years later, the intrinsic advantages of EHR systems, together with various governmental initiatives, such as the ‘Meaningful Use’ programme in the US, have helped to push EHR adoption to 80+% levels in most developed countries. In the US, for instance, in 2019, EHR use stood at 93% in smaller rural hospitals and 99% in large hospitals [9]. At the same time, anecdotal evidence suggests that eRDC usage in clinical trials is now almost 100% for commercially sponsored research, and between 80-90% in the non-commercial sector.

It is true that ‘EHR use’ can mean different things in different contexts, with some healthcare facilities still using paper alongside electronic records. It is also true that electronic systems

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

themselves may contain large amounts of unstructured text. Nevertheless, a high proportion of both healthcare and research data is now available, or potentially available, in electronic form, (though empirical data on the exact proportion is hard to find). The potential for using healthcare data as a direct research resource, rather than in its traditional role as 'source documentation', has increased accordingly.

In addition, EHR systems have become more sophisticated, and can include a wide range of data. For example, they may contain prescribing data, test orders as well as test results, pathology data, nursing care plans, images and associated reports, family health histories and in some cases genomic sequencing and expression data, in addition to the more traditional demographic, diagnostic and observational information [10].

Randomised control trials (RCTs) have long been, and remain, the evidential gold standard for medical research, but obtaining the data is costly and labour-intensive. Well-defined inclusion and exclusion criteria and a rigid protocol give trials greater internal validity, but they can, for the same reason, make it more difficult to recruit study participants, especially in research for rare conditions. Traditional RCTs have also been criticised as having poor external validity, as being too divorced from real-world clinical practice, and too restrictive in the selection of participants [11, 12].

The enormous volumes of healthcare data offer the promise of circumventing or mitigating some of these issues. The data is already collected, from millions of people that – by definition – truly represent the populations of interest. Finding ways to harvest the potential scientific value of all that data has long been recognised as both a challenge and an opportunity [13 - 15]. The sheer scale of the data available has tempted many to explore the use of 'big data analytics' to develop insights into healthcare and treatment, but the focus here is on more specific usage, more directly linked to traditional observational and interventional research.

2.2 Previous work and initiatives

Researchers have tried to exploit real world healthcare data in a wide variety of ways. Table 1, which is adapted and extended from [16], lists different forms of RWD use, with concrete examples given for each application. While the use of RWD for observational research, and safety and other forms of post-marketing surveillance, has been labelled as 'established', its role in interventional research is generally viewed as much more 'experimental' [16]. In many cases, as Table 1 demonstrates, in interventional research the RWD acts as an adjunct to more traditionally collected trial data, providing support for an RCT rather than replacing it.

The main RWD usage not included in Table 1, but which has attracted considerable interest in recent years, has been its potential role in supporting regulatory decisions. An example of this, from 2019, was the FDA extending the indications for the drug Palbociclib (Ibrance), in combination with endocrine therapy, to male breast cancer. They stated that the decision was

> "based upon data from post-marketing reports and electronic health records showing that the safety profile for men treated with Ibrance is consistent with the safety profile in women treated with Ibrance" [33].

Guidelines about the use of RWD in regulatory submissions have been offered in recent years from both the FDA [34, 35] and the EMA, though the latter has often couched this in terms of 'big

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Application	Example studies (full references in References section)
Observational studies, including epidemiological work, investigating the natural history of disease, associated risk factors, drug prescription patterns, etc.	Jeon et al. 2015. The Association of Statin Use after Cancer Diagnosis with Survival in Pancreatic Cancer Patients: A SEER-Medicare Analysis [17] Vashisht et al. 2018. Association of Hemoglobin A1c Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin. Analysis from the OHDSI Initiative [18] Suchard et al. 2019. Comprehensive Comparative Effectiveness and Safety of First-Line Antihypertensive Drug Classes: A Systematic, Multinational, Large-Scale Analysis [19]
Safety surveillance	Castro et al. 2013. QT Interval and Antidepressant Use: A Cross Sectional Study of Electronic Health Records [20] Vickers-Smith et al. 2020. Gabapentin Drug Misuse Signals: A Pharmacovigilance Assessment Using the FDA Adverse Event Reporting System [21]
Hypothesis generation	Onukwugha E. 2017. Visualising data for hypothesis generation using large volume claims data. [22]
Trial feasibility assessments	Visweswaran et al. 2018. Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network [23]
Patient recruitment	Quint et al. 2018 Recruitment of Patients with Chronic Obstructive Pulmonary Disease (COPD) from the Clinical Practice Research Datalink (CPRD) for Research [24]
Providing control data for single arm trials	Gökbuget et al. 2016. Blinatumomab vs historical standard therapy of adult relapsed/ refractory acute lymphoblastic leukemia. [25]
Evaluation in health technology assessment (and funding support decisions)	Makady et al. 2017. Policies for Use of Real-World Data in Health Technology Assessment (HTA): A Comparative Study of Six HTA Agencies. [26] Bell et al. The Use of Real World Data for the Estimation of Treatment Effects in NICE Decision Making, 2016 [27]
Providing evidence for pragmatic trials	Marquis-Gravel et al. 2020. Rationale and Design of the Aspirin Dosing—A Patient-Centric Trial Assessing Benefits and Long-Term Effectiveness (ADAPTABLE) Trial [28]
Providing long term follow up data	Davies et al. 2018. Long Term Extension of a Randomised Controlled Trial of Probiotics Using Electronic Health Records. [29]
Direct import of EHR data to eCRFs	Erlinge et al. 2016. Bivalirudin versus Heparin in Non-ST and ST-Segment Elevation Myocardial Infarction—a Registry-Based Randomized Clinical Trial in the SWEDEHEART [30]
Comparative effectiveness trials	Albertson et al. 2017. The Salford Lung Study: A Pioneering Comparative Effectiveness Approach to COPD and Asthma in Clinical Trials' [31]
Checking the representativeness of study populations	Lee et al. 2012. Representativeness of the Dabigatran, Apixaban and Rivaroxaban Clinical Trial Populations to Real-World Atrial Fibrillation Patients in the United Kingdom: A Cross-Sectional Analysis Using the General Practice Research Database'. [32]

Table 1: Applications of RWD in Clinical Research

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

data' [36]. A recent review of the different approaches to RWD, by regulatory authorities in the US, Europe and China, is provided by [37].

In Europe the EMA included '*Promote use of high-quality real-world data (RWD) in decision-making*' in a list of strategic goals first published in 2016. It then undertook an extensive 3-year 'strategic reflection' exercise, involving a wide variety of stakeholders, and asked for their rating of the goals in terms of their importance for delivering significant change. The top 5 are listed in table 2 – 'promoting the use of RWD' was ranked as number 2 [38].

Order	Goal #	Strategic goal
1	9	Foster innovation in clinical trials
2	18	**Promote use of high-quality real-world data (RWD) in decision making**
3	17	Reinforce patient relevance in evidence generation
4	15	Contribute to HTA's preparedness and downstream decision making for innovative medicines
5	1	Support developments in precision medicine, biomarkers and 'omics

Table 2: EMA stakeholder ranking of strategic goals, to deliver significant change [from 38]

At the same time, the goal seen as most significant involved clinical trials, and this seems to echo the EMA's own thinking on the role of RWD – as an adjunct or complement to clinical trial data. In the same document (EMA Regulatory Science to 2025, Strategic Reflection) they state:

> "Real world data is currently used predominantly in the post-authorisation phase but there are opportunities for further application throughout the medicines lifecycle to help address some of the limitations of clinical trials. The Agency recognises the fundamental importance of clinical trials in the establishment of a products benefit risk, however, there is potential for benefit of using RWD to generate complementary evidence across the product life cycle.
>
> It will be important to agree amongst stakeholders where RWD may add value into the assessment process. Given the often heterogeneous nature of the data sources, further work is also needed on the analytical and epidemiological methodologies needed to deliver robust evidence. There are additional needs to ensure security of the data, …." [38, p 36].

Using RWD in both observational and interventional research has also been the subject of considerable methodological research. Several large scale projects, funded by, amongst others, the EU and the IMI, have tried to develop tools and infrastructure to promote the use of RWD, in clinical research in general, in pharmacovigilance, within 'learning health systems', in drug discovery and development, and in regulatory decision making.

Table 3 lists some of the major projects, and includes links to the relevant project websites (where they still exist).Along with the projects listed in Table 3, there also continues to be a steady stream of methodological research work, for example examining the feasibility of using RWD in different ways [39, 40], discussing design and reporting issues [41, 42], or generally reviewing the potential of RWD [43 - 46].

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Name	Description and link
Sentinel (2008 – Present)	Sentinel System – "Sentinel is the FDA's national electronic system which has transformed the way researchers monitor the safety of FDA-regulated medical products, including drugs, vaccines, biologics, and medical devices." FDA https://www.fda.gov/safety/fdas-sentinel-initiative, 5 year strategy, 2019 – 2023: https://www.fda.gov/media/120333/download
Transform (2010 – 2015) Translational Research and Patient Safety in Europe	EU FP7 project, budget €9.7 million, aims were "…to develop a rapid learning healthcare system driven by advanced computational infrastructure that can improve both patient safety and the conduct and volume of clinical research in Europe."
Open PHACTS (2011 – 2016) Now the Open PHACTS Foundation	IMI project, total budget €20 million, intended "to deliver and sustain an 'open pharmacological space' using and enhancing state-of-the-art semantic web standards and technologies." https://www.imi.europa.eu/projects-results/project-factsheets/open-phacts, https://www.openphactsfoundation.org/
EHR4CR, (2011 to 2016) Electronic Health Record Systems for Clinical Research	IMI project, total budget of €16+ million. Aims were "to improve the design of patient-centric trials by developing a platform that provides access to existing patient electronic health record systems (EHRs)." https://www.imi.europa.eu/projects-results/project-factsheets/ehr4cr Custodix Insite platform established in 2016 as a commercial follow on to the project. InSite itself acquired by TriNetX in 2019 (https://trinetx.com/)
SALUS (2012 – 2015) interoperability framework	EDU FP7 project, to develop a "Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies" https://www.allcryptowhitepapers.com/wp-content/uploads/2018/05/SALUS.pdf
GetReal (2013-2017) Incorporating real-life clinical data into drug development	IMI project, total budget €17 million. Aim was to develop new tools and resources for incorporating real-life data into drug development. https://www.imi.europa.eu/projects-results/project-factsheets/getreal
GetReal Initiative (2018 – 2021)	IMI follow on project (budget €3 million) to drive the adoption of tools developed in GetReal, to increase the quality of real-world evidence (RWE) generation in medicines development and regulatory / HTA processes. https://www.getreal-initiative.eu/. Now the not-for-profit GetReal Institute (https://www.getreal-institute.org/)

Name	Description and link
The Argonaut Project (2014 – present)	HL7 project, with EHR companies, "A private sector initiative to advance industry adoption of modern, open interoperability standards in EHRs", mostly US based, https://hl7.org/implement/standards/fhir/2015Jan/argonauts.html
OHDSI (2014 – present) Observational Health Data Sciences and Informatics	Initially US based, "a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics" Owns and manages the OMOP common data model and a large suite of related tools. Other listed 'Areas of Focus' include safety surveillance, comparative effectiveness research, personalised risk prediction, data characterisation and quality improvement. https://www.ohdsi.org/, Book of OHDSI https://ohdsi.github.io/TheBookOfOhdsi/
EHR2EDC (2018 – 2019)	EIT project. Demonstrated methods and technologies for achieving automatically and secure transfer of EHR data to an Electronic Data Capture (EDC) system, for a study investigator to review and save. Extended by one year with the EHR2EDC Champion Programme (2020) https://eithealth.eu/project/ehr2edc/
EHDEN, (2018 – 2024) European Health Data and Evidence	IMI project, total budget €29 million. Aims to construct a "trusted open science community built for health data research via a European federated network". Strongly promoting OMOP adoption https://www.ehden.eu/
DARWIN EU (2021 onwards) Data Analysis and Real World Interrogation Network	EMA project, designed to create "a coordination centre to provide timely and reliable evidence on the use, safety and effectiveness of medicines for human use, including vaccines, from real world healthcare databases across the European Union". https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu

Table 3: Projects and infrastructures using RWD in Research and Pharmacovigilance

2.3 Problems

With all of this interest and input, one could be forgiven for believing that by now RWD would be playing a substantial role in clinical research, but in fact, as the EMA's statement implies, the approach remains far from the mainstream. Using RWD is often still described as, or is the subject of, a research project rather than normal practice.

The reason, of course, is that there are substantial and well recognised difficulties in using RWD, which have prevented the widespread use of this data in interventional research. Some of these relate to legal and ethical issues – for example the continuing lack of legal clarity (at least in Europe) around the secondary use of sensitive data, collected in this case primarily to support healthcare and treatment activities, outside of any research context, and the need to apply robust privacy protection measures to the data without diluting its scientific value. Some are methodological, for instance the lack of randomisation and blinding in most RWD collection, which may allow unconscious bias to reduce study validity. There is also the fundamental issue that the core purpose of RWD is to support care and treatment, and the relevance of the available data to a particular research question may therefore be incomplete or indirect. Many of the problems, however, are a function of the data itself. The main data related issues are summarised below:

a) *Data heterogeneity:* Different healthcare systems structure, code and categorise their data differently. This of course is the prime reason why data standards need to be applied in healthcare. Even if only one RWD source is used in a study, much effort is likely to be needed to simply understand the data. If multiple sources are used, either concurrently or over time, then without data standards a great deal of additional work is created as each needs to be mapped to the required research data structure individually.

b) *High levels of unstructured data:* EHR records are notorious for including many sections of free text, e.g. for patient history, assessment, treatment plans, goals, etc. Free text is much easier for doctors, nurses and others to use when inputting data, but very difficult to use, at least by a machine, as a data source. A large amount of research has been carried out investigating the use of Natural Language Processing (NLP) techniques for extracting this data (reviews can be found at [47], [48] and [49]) but in general this is a research effort, with no widely used approach available to the non-NLP specialist.

c) *Institution centric:* In general, EHR data and claims databases are institution specific – they record health data in a particular context and do not generally transmit it to or receive it from other systems. Even in a relatively 'joined up' system, like the UK's National Health Service, a discharge letter from a hospital, though it may be transmitted electronically, has to be read and interpreted by staff before its data is entered manually into the primary care EHR [50]. In a more fragmented system, records from one system might not be transferred at all to another, leading to gaps in the available data.

   While patient-centric health records have been developed in various parts of Europe, and are currently also being developed across the EU as a whole, these are usually only intended to contain a core subset of data (e.g. in the UK, allergies, current medication, significant medical history, and currently also COVID-19 history, [51]) to enable emergency treatment when the individual is away from their normal healthcare services.

d) *Data quality:* RWD data is not checked, at least not in any systematic way, and data quality may be poor. Inconsistencies in the data will arise because – without a protocol to timetable assessments, to exclude confounding co-morbidities, and to prescribe specific treatments and assessments – the details of the treatments received and assessments made will vary between individuals even if they have received the same diagnosis and are being treated in the same place. The data may also be stored without an associated audit trail, which theoretically makes it non-compliant with GCP requirements, and practically makes it more difficult to manage errors and inconsistencies.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

> The data inconsistency can be further magnified by variations and gaps in the data entry process, carried out by busy staff without the time to check or chase data, and by the fact that the data will often be entered in different departments and by different people, perhaps with a different understanding of the data items that are being requested, or the terms or units that should be used.
>
> Awareness of these issues has led to proposals for assessing data quality (DQ), e.g. for consistency, completeness, plausibility etc. A recent review drew the conclusion, however, that, in at least the US research network that was studied, "The practice of DQ assessment is still limited in scope. Future work is warranted to generate understandable, executable, and reusable DQ measures" [52].

An account of data (and other) issues encountered when using RWD, in this case for COVID-19 research, which also includes some suggestions for assessing RWD quality and veracity, is provided by Kohane et al [53]. This is particularly relevant given that two early, widely discussed but now retracted papers on COVID-19 ([54], [55]) were based on RWD analysis. The conclusion from the Kohane paper includes the following statements:

> "… We need to be open and transparent about the inherent limitations of the data and the analyses. We should also acknowledge alternative interpretations of the results. … Extra caution is also needed in how we draw causal inferences from EHR data, especially given the noisiness and incompleteness of the data in addition to several sources of bias, …"

This seems to be a fair representation of the current 'state of play' for the use of RWD. While claims have long been made about the potential of this data, and experimental work has demonstrated its value across a variety of use cases, the numbers of studies using this approach, compared to the many tens of thousands of trials run every year is relatively small (about 50,000 new studies were added to trial registries in 2021 [56]). In some cases, the role of RWD has been to help establish, validate, or complement the evidence from randomised controlled trials. In others it has been used more independently, to "draw causal inference" or justify regulatory change. But the more important the role of RWD, in any particular study or decision, the more important it is to be aware of the possible weaknesses in this type of data and to check the provenance, veracity and quality of the source material. This is not to deny the potential value of this type of data, but it does – especially if the use of RWD becomes more routine – underline the need to use it with a degree of caution.

3. Data Standards and Interoperability in Observational Research and Healthcare

3.1 Outcome measures

The difficulty with discussing healthcare 'outcomes' is that they mean different things to different people. For example, they may refer to:

- Economic or organisational 'key performance indicators' (patient drug cost per stay, levels of bed occupancy, average length of hospital stay).

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

- Relatively crude measures of health status (mortality and morbidity data, admission rates), aggregated at different levels.
- Health technology assessment (HTA), defined as "the clinical and cost-effectiveness, and broader impact of healthcare treatments and tests, …" [57].
- Detailed disease specific outcome measures as defined and provided by the medical staff and the healthcare system (e.g. impact of treatment on PSA scores, blood lipid levels, tumour relapse rates, etc.).
- Detailed expressions of satisfaction / discomfort and general quality of life, as provided by the patients themselves, usually through questionnaires.

Each outcome 'type' will have its own community of users and, often, associated researchers, as well as its own techniques and data. Outcomes may also be measured at a variety of levels (disease specific, departmental, organisational, regional, national, international etc.). The data generated within each type of outcome assessment is likely to be similar, but there appear to be few deliberate attempts to standardise it in any formal sense.

An exception is provided by the work of ICHOM, the International Consortium for Health Outcomes Measurement [58]. Originating in the US but now operating globally, ICHOM develops and publishes "standard sets of outcome measures that matter most to patients", with each standard set – there are currently 39 – developed by a consortium of medical experts and patient representatives. Table 4 is a list of some of the patient centric outcomes developed by ICHOM [59]. The use of quality of life questionnaires is a recurring feature, but questionnaires take money and staff to administer, collect and process, and so are not always a feasible option in routine practice.

Area	Patient Centric Outcome Measures
Respiratory diseases	Dyspnoea, worsening disease, HR QoL, symptom control
Rheumatoid arthritis	Pain, Fatigue, Activity limitation, emotional and physical health impact, impact on work/home life
Diabetes	Psychological well-being, diabetes distress, depression
Atrial fibrillation	Ability to work, exercise tolerance, symptom severity, HR QoL
Heart failure	Symptom control, activities of daily living, independence, psychosocial health
Lung cancer	HR QoL, fatigue and vitality, Pain, cough, shortness of breath, performance status
Breast cancer	HR QoL, arthralgia, neuropathy, vasomotor symptoms, fatigue, pain, depression, arm and breast symptoms, body image
Chronic kidney disease	Fatigue, pain, physical function, HR QoL
Inflammatory bowel diseases	Change in bowel symptoms, pain and discomfort, normal activities, energy and fatigue, weight
Major depressive disorder	Physical functioning, work functioning, social functioning, symptoms of depression, symptoms of anxiety

Table 4: Example patient centric outcomes developed by ICHOM. HR QoL = Health related quality of life (quoted in [59])

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

It is difficult to estimate the levels of use of ICHOM standards. The ICHOM website has a global implementation page that has a world map with 290 'pins' (as of December 2021), but these are standard sets, not hospitals – the number of implementing organisations is much lower. Furthermore, in some cases the implementation is described as 'in progress', but so is the web page as a whole. This problem is not restricted to ICHOM – who do at least attempt to display their usage data. Any discussion about standards is handicapped by the lack of empirical data about their level of use. The websites of the standards developers often claim widespread usage, but there are rarely empirical data available to support these claims.

Whilst most of the ICHOM usage is described as being 'in routine practice' attempts have also been made to either apply the standards to patient registries – for example within France in registries for cataract patients [60], and globally for a federation of registries for prostate cancer patients [61] – or to compare registry outcome data with the ICHOM standards – for example with a set of national diabetes registries [62].

Patient registries, defined by the EMA as "organised systems that use observational methods to collect uniform data on a population defined by a particular disease, condition or exposure, and that is followed over time" would seem an obvious additional source of real world outcome data, but traditionally have been set up independently of each other with little standardisation of the data between them.

The EMA set up an initiative in 2015 to promote harmonisation of registry operation and data consistency, so that the data could be more easily included in the benefit-risk evaluations of medicines [63]. A report in 2019 made specific proposals, including common core data elements, harmonised data elements, core patient reported outcome (PRO) measures, and systems to assess data completeness and accuracy [64], but there seem to be no published reports describing the impact of these proposals. Within rare disease registries there is also a project to make data more consistent and more 'FAIR', [65] but it too seems to be at a relatively early stage of development.

Some countries have set up 'clinical quality registries', with a particular emphasis on collecting outcome data and monitoring comparative effectiveness, potentially of both treatments and organisations. A definition is provided below, taken from a paper describing one of the clinical quality registries in Australia and New Zealand [66].

> "A clinical quality registry is a systematic, standardised, structured and continuous collection of a pre-specified minimum data set of health, process and outcomes data for people with particular health characteristics. By organising longitudinal, observational data from multiple participatory sites into a single central repository, clinical quality registries enable large-scale real-world register studies with greater statistical power, external validity and inferential reliability."

In some ways such registries represent an elaboration of the 'minimal datasets' that are collected centrally about care episodes in many countries, and which are also used to monitor comparative effectiveness. Whether their advent will drive a more standardised development and adoption of healthcare outcome measures remains to be seen. At the moment, while there are certainly lots of initiatives recognising the importance and potential value of this type of data, serious attempts to standardise it – in registries, healthcare or elsewhere - appear rare.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

3.2 Syntactic standards

Whilst there are, as described below, emerging syntactic standards for data in healthcare, and underlying semantic standards that appear to be growing in use, there appear to be relatively few syntactic data standards within health-related *observational research*. There are the STROBE (Strengthening the reporting of observational studies in epidemiology) standards for publishing observational studies [67], but these provide a list of report contents, rather than dictating anything about how the data should be organised in a technical sense.

The WHO published Recommended Surveillance Standards in 1999 [68], but again this is a relatively high-level document providing diagnostic definitions, a recommended surveillance strategy and a short list of minimum data elements for a range of conditions, more often concerned with aggregate data rather than individual cases. In 2018 Fairchild et al. were still describing substantial challenges to interoperability of epidemiological data and calling for data standards to combat them [69]. The need for rapid, coordinated evidence generation for the COVID-19 pandemic, where the bulk of data collected has been observational, has turned the current lack of interoperability into a priority for global action [70].

Previous pandemics – specifically the 2009 H1N1pdm09 influenza outbreak – have led to some attempts to address this problem. The International Severe Acute Respiratory and emerging Infections Consortium (ISARIC [71]) developed a Clinical Characterisation Protocol (CCP) 'for any severe or potentially severe acute infection of public health interest'. The CCP, later endorsed by the WHO [72, 73], is a pre-approved protocol that can support rapid data collection in a standardised way. At the outset of the COVID-19 pandemic, in January 2020, it was used by the Infectious Disease Data Observatory (IDDO [74]) and ISARIC to develop COVID-19 specific eCRFs, which were activated immediately. The rapid provision of these standardized eCRFs has allowed the collection of interoperable observational data from over 500,000 participants in at least 50 countries, and is an example of what can be achieved with sufficient preparation.

Another example of standards being applied to observational data are the E2B (R3) reporting guidelines for individual case safety reports, the mechanism for reporting severe adverse reactions to medication, which are highly structured and published as an ISO standard as well as being interpreted by regulatory authorities [e.g. 75]. Unfortunately these guidelines, and the ISARIC CCP, appear to be exceptions. General observational research data – from cohort studies, case studies, retrospective surveys, etc. – whilst including a natural overlap of basic data points (e.g. gender, age, country) does not seem to have any widely used systems for standardising the syntax or structure of the data.

Up to about 10 years ago the same could probably also be said of healthcare data in general, but some syntactic data standards are now beginning to emerge, despite the huge heterogeneity that remains within the source record systems – especially the electronic health record (EHR) systems in both hospitals and primary care.

Attempts have been made in some countries to introduce standards 'at source', within the EHR systems themselves. In the UK this led to the NHS Care Records Service project, running from 2002 until 2011 and intended to provide a national, standardised, health record system. Unfortunately, despite an investment of over £13 billion, the project achieved almost nothing [76]. In the US, the 'Meaningful use' programme, introduced after the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, distributed $30 billion to

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

physicians and hospitals as incentives for them to install EHR systems. Because the bar was set relatively low in terms of system specification, however, the programme – while it increased EHR usage – had varying impacts on healthcare practice and did not greatly contribute to data standardisation [77, 78]. In response to this the initiative was renamed in April 2018 as the ’Promoting Interoperability’ programme [79] with government funds contingent on using EHRs that implement the US ‘Core Data for Interoperability’ set of data items [80]. It was reported in 2021 that version 2 of the Core Data for Interoperability would represent a significant expansion of the data items [81].

In contrast to these governmental ‘top down’ attempts to standardise healthcare data, the OpenEHR project represents a bottom up, open source approach, in a landscape dominated by commercial systems. Begun in 2003, it offers both a clinical modelling system (i.e. an ontology) and a suite of software components that can be used to build EHR systems, with both model and software formally and publicly specified [82]. It is not the only open source EHR system (globally, 15 are listed in a comparison of the 5 most accessed on the web [83]) but it is unusual in providing a formal ontology for clinical care, and it is the open-source system most used in Europe.

The number of deployments of OpenEHR are listed on their website and are tabulated by country in Table 5. As can be seen, there is a strong bias towards North West Europe. The difficulty is that, even if this number is slowly increasing (e.g. the Welsh Health Service announced plans to introduce OpenEHR in 2021 [84]) the total of 67 remains very small on a global scale. Even within Europe, the 60 or so installations – not all of which are necessarily full EHR systems – are a very small proportion of the total EHR systems installed, probably of the order of 1%.

Country	Number of Deployments
Netherlands	18
Sweden	10
Germany	9
United Kingdom	9
Norway	6
Australia	4
Slovenia	3
Switzerland	2
Brazil	1
Finland	1
Italy	1
Malta	1
Philippines	1
Russia	1

Table 5: OpenEHR Deployments, per country (from the OpenEHR website, accessed 30/10/2021)

If attempts to standardise data within the EHR systems themselves have so far been underwhelming, a more promising approach has been to standardise the data as exported /

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

imported, or to use extract – transform – load (ETL) processes to pull data out of the EHR system and into a separate 'bolt-on' database, transforming it to a more consistent structure – a so-called **Common Data Model** (CDM) – in the process.

One CDM based approach is provided by the 'Informatics for integrating biology at the bedside' project, more commonly known as i2b2 [85]. The i2b2 system consists of a front-end client application and a collection of back-end services referred to as a 'hive', which are linked to the EHR and / or other source data systems. At a minimum, a hive must have Project Management, Ontology, and Data Repository components, but will often have other modules to carry out other aspects of overall data workflow. Data is loaded into the hive's data repository from source systems via ETL [86].

An i2b2 data repository is based on an extremely flexible EAV (Entity-Attribute-Value) architecture [87]. The same approach is used by most of the database management systems used for clinical trial data – and for the same reason: the EAV approach maximises the flexibility of data storage. In an EAV structure, the entities being described are themselves part of the data, instead of being 'hard-wired' into the system as fixed column codes.

i2b2 hives are not intrinsically compatible with each other, but can be made so by ensuring their ontology modules have the same lists of possible data items and allowed values, and the same definitions for both – in effect by ensuring they have loaded a common data model. I2b2 provides a framework called the Shared Health Research Information Network (SHRINE) to support the management of different but compatible hives.

An example of its use is the US Accrual to Clinical Trials (ACT) network, which uses EHR data, as extracted and aggregated through compatible i2b2 hives, from 41 sites (in 2020), to identify the best hospitals for multi-centre studies, with the aim of managing and improving trial accrual [88, 89]. The CDM used is one developed by PCORNet (the National Patient Centred Clinical Research Network) in the US.

To *enforce* data standards on EHR data, rather than simply enabling them, requires a more opinionated system than i2b2, one that stipulates a single common data model in detail. While several CDM based systems exist for healthcare data the most important, particularly in Europe, is **OMOP** – the CDM from the Observational Medical Outcomes Partnership [90], managed by **OHDSI** (pronounced Odyssey), the Observational Health Data Sciences and Informatics program, "a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics" [91].

OMOP is once again a 'bolt-on' system, added to a source system as a set of ETL processes and a target database on a linked server, along with query facilities. OMOP is well documented, and users can also take advantage of a large suite of tools developed and supported by OHDSI, to help set up and run an OMOP system. These include:

- White Rabbit: For creating source data (EHR) inventories
- Rabbit in a Hat: For mapping source tables to CDM structure
- Usagi: For mapping source terms to CDM standardised vocabularies
- Achilles: For ETL Verification – review database profiles
- Atlas: For querying and cohort identification within the CDM
- Athena: For searching and loading standardised vocabularies
- Hades: A collection of open source R packages.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Further details and the full list of tools can be found at [92].

The destination system for the OMOP ETL exercise, the OMOP database itself, has a reassuringly traditional look – fixed tables with fixed fields, inter-related by foreign keys. The tables in the latest version (6.0) are shown in figure 3 and are taken from [96]. The OMOP CDM is 'person-centric', meaning that all clinical event tables (in the left hand column of figure 3) are linked to a central 'Person' table. Together with a date or start date for each event, this allows for a longitudinal record of all the healthcare relevant events linked to an individual. The exceptions from this rule are the standardized health system data tables, which are linked directly to events of the various domains.

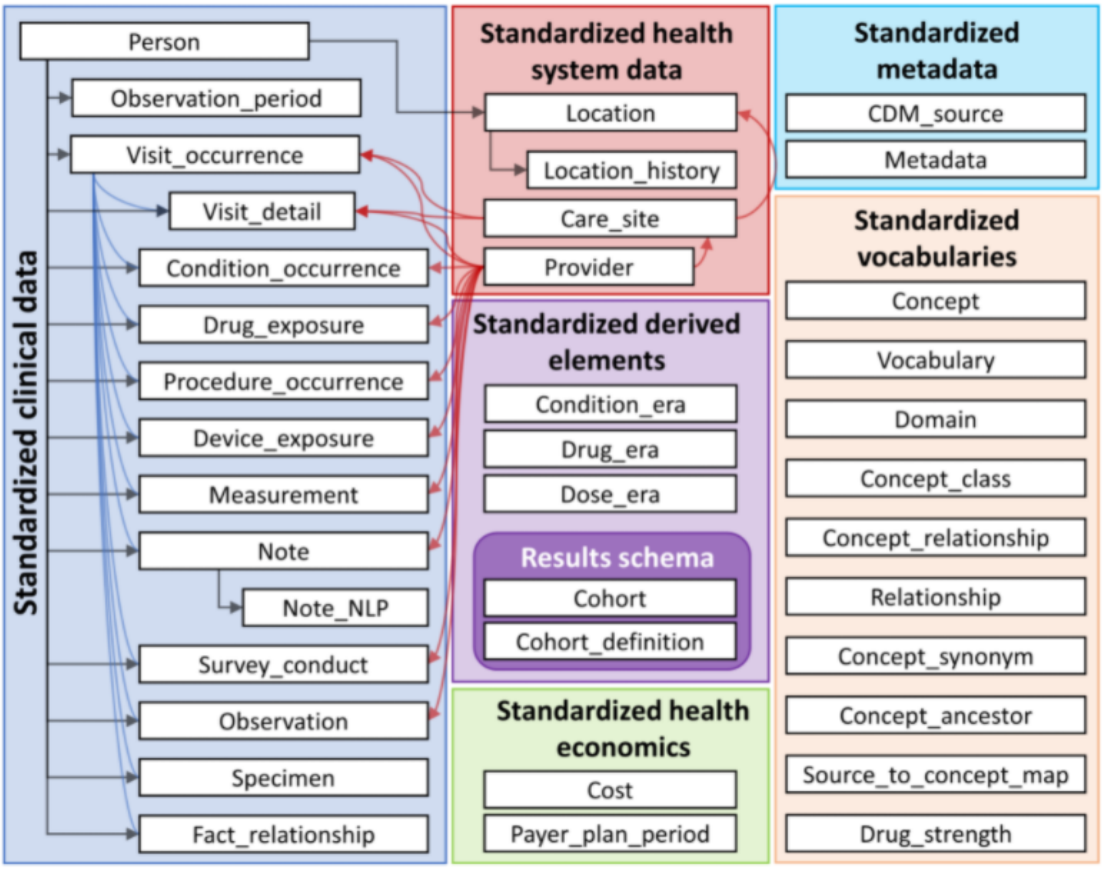


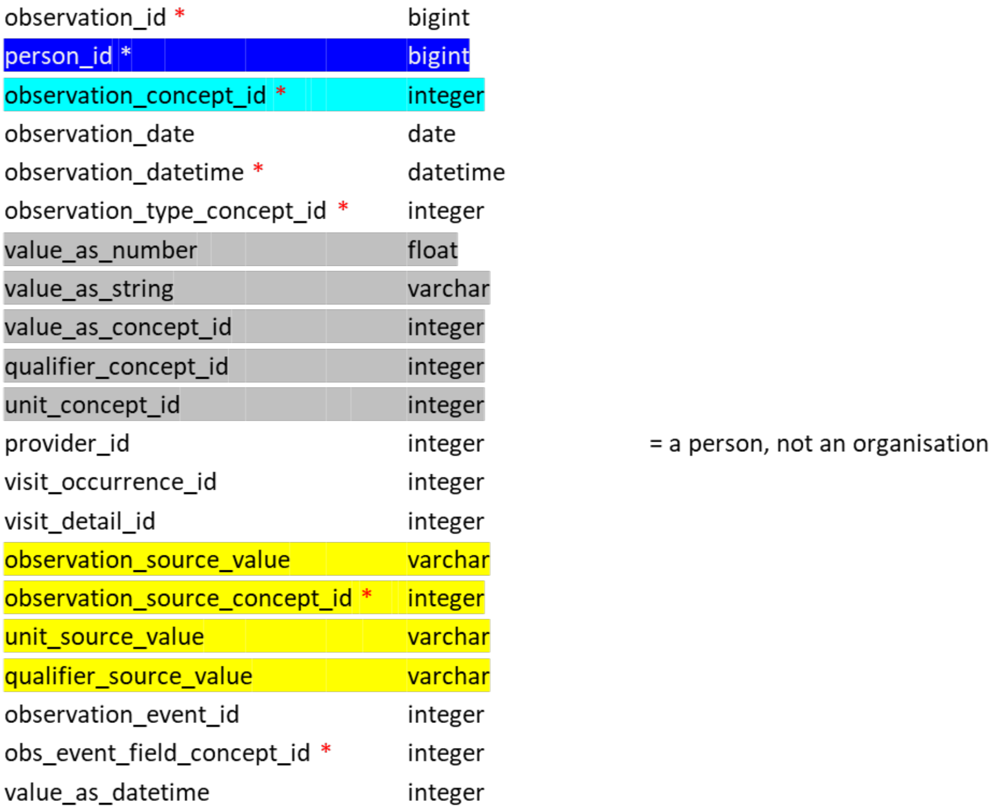
Figure 3: OMOP CDM tables (v6.0)

The OMOP system has detailed and comprehensive documentation, largely provided by a website (and e-book) called the 'Book of OHDSI'. The detailed specification of the Observation table is provided, as an example, as Appendix 3, and demonstrates the detailed instructions and guidance that are available with the system (some of which, e.g. the meaning of CONCEPT_ID, demand familiarity with how the system as a whole works). For greater clarity, the fields in the same table are listed as figure 4 (required fields are starred).

Figure 4 illustrates two common features of OMOP tables:

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

a) As shown by the fields highlighted in yellow, the table includes fields for the data as found in the original source record, in addition to the data as transformed into OMOP data structures and codes. This is true of most OMOP tables and allows much easier validation of, and greater confidence in, the transformation process.



observation_id *	bigint	
person_id *	bigint	
observation_concept_id *	integer	
observation_date	date	
observation_datetime *	datetime	
observation_type_concept_id *	integer	
value_as_number	float	
value_as_string	varchar	
value_as_concept_id	integer	
qualifier_concept_id	integer	
unit_concept_id	integer	
provider_id	integer	= a person, not an organisation
visit_occurrence_id	integer	
visit_detail_id	integer	
observation_source_value	varchar	
observation_source_concept_id *	integer	
unit_source_value	varchar	
qualifier_source_value	varchar	
observation_event_id	integer	
obs_event_field_concept_id *	integer	
value_as_datetime	integer	

Figure 4: Fields in the OMOP Observation table (v6.0).

b) As shown by the other coloured fields, this table is using an EAV structure. The entity is represented by the person id, in dark blue, the attribute by the observation_concept_id in light blue, and the value by one or more of the grey value fields. This is the case for several of the OMOP tables. In other words, despite OMOP looking like a traditional relational design at first glance, the data is structured using what is fundamentally an EAV approach, but split into specific domains.

One of the reasons for OMOP's increasing importance is the strong promotion it has received from the IMI in the context of the EHDEN project. Adding an OMOP system represents a lot of work for a hospital IT department, and buying in specialist contractors to do the task may therefore be necessary.

The EHDEN project, which includes amongst its aims the standardisation of at least 100 million patient records across Europe using OMOP, recognises this, and has used several calls to invite 'data partners' (data holders such as hospitals) to apply for funding to map their health data to the OMOP CDM [93]. So far (late 2021), 98 partners have joined the programme, in 23 countries. At the same time, EHDEN has developed a process to certify SMEs (47 in 19 countries at October 2021) to carry out OMOP installations [94]. Interestingly, the SMEs are not only working with EHDEN data partners, but also starting to add OMOP ETL systems to their own independent customer base. Furthermore, OMOP is not limited to hospital data. A web page describing the mapping of a biobank's data to OMOP (at UCL in London) is available at [95].

For all these reasons it appears that OMOP is slowly becoming the major system of syntactic standards for healthcare data, particularly in Europe. It is, however, difficult to obtain exact figures of working, successful, installations, for OMOP, i2b2, or any other system. Anecdotal evidence, and estimates of the total numbers of healthcare providers in Europe, would suggest that – despite the considerable momentum of EHDEN – only a few percent of healthcare providers have bolt-on ETL systems that can standardise their data. The great bulk of healthcare data therefore remains far from standardised.

In terms of data transfer standards, by far the most important system is **HL7 FHIR.** HL7 first appeared in 1989, as version 2, and quickly established itself, especially in the US, as a standard for moving health data between systems. After HL7 version 3 and then HL7-CDA, which like version 2 were all document based exchange formats for clinical data, HL7 FHIR was released in 2015, with the latest version (v4) released in December 2018. HL7 FHIR allows users to access EHR data based upon a REST API, giving a more flexible, granular data interface, plus the ability to add additional queries based on previous results. The data has to be transformed into FHIR compatible data packages by a bolt-on FHIR server that also interprets the API requests [97].

HL7 is not a data standard, in the sense that data is not normally stored 'as HL7'. It can be transformed into a database structure, and sometimes is, but there is no standardised way of doing this. When data are exchanged using HL7 they are organised into different 'resources' – self contained, structured records that represent healthcare entities such as a patient, observation, care plan, adverse event, medication administration, etc. – about 140 are defined in total [98]. Resources are modified by using 'profiles,' which can constrain or extend the default data elements and attributes of the resource, as defined by HL7. If two organisations can both interpret the same HL7 profiles, they can exchange information, however that information is stored in their local databases. In general, profiles will be shared within a network – whether that is organisational, regional or national – to enable data exchange. The UK's NHS for example, which uses HL7 FHIR, maintains a public listing of all of its HL7 profiles on the HL7 UK FHIR Reference server [99].

HL7 FHIR has been well received and its use appears to be growing rapidly, though it may be some years before it supplants all of the installed older versions (even version 2, which has remained in use in many places because of its relative simplicity). Apart from the NHS in the UK, FHIR users include Medicare / Medicaid in the US, the Brazilian National Health Data network, and the German Medical Informatics Initiative. Being able to 'speak' HL7 FHIR is therefore an increasingly important requirement for any healthcare record system.

3.3 Semantic standards

OMOP's value as a standard system stems from the fact that it highly opinionated - it enforces standardisation. This may seem odd given that many data tables in OMOP include a flexible EAV

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

approach to data storage. The answer lies in what OMOP calls its Standardized Vocabularies, the controlled terminologies that are available for use, because OMOP has very definite views about which controlled terminology should be used for each type of data. This is the strength of OMOP (for instance compared to i2b2) – it has the courage to say exactly how, semantically, it wants the data to be represented. This clear specification of CTs gives the data in an OMOP system the consistency that creates a true data standard, though of course in some situations this could be seen as a problematic inflexibility.

Table 6 provides a summary table, showing the terminology system to be used for each of the major variable groupings. When two or more coding systems are listed in table 6 for a particular domain this is not implying a choice, only acknowledging multiple origins for the set of preferred codes. Thus, although both SNOMED and LOINC are listed as the controlled vocabularies for Measurement, any specific measurement should be coded using the system as stipulated in the OMOP documentation, which will be *either* SNOMED *or* LOINC, but not both. The user (the ETL designer) does not get a choice.

Domain	for standard concepts
Condition	SNOMED, ICDO3
Procedure	SNOMED, CPT4, HCPCS, ICD10PCS, ICD9proc, OPCS4
Measurement	SNOMED, LOINC
Drug	RxNorm, RxNorm Extension, CVX
Device	SNOMED
Observation	SNOMED
Visit	CMS Place of Service, ABMT, NUCC

Table 6. Controlled vocabularies within OMOP (simplified from [98])

In reality, the use of controlled vocabularies is a little more nuanced than as described here – for instance MedDRA is not a preferred vocabulary but it does exist within OMOP, as a so-called 'classification concept', which means it can be used for querying data. Concepts in OMOP can also have a hierarchical relationship with one another [100].

The OMOP CT list shows a marked preference for **SNOMED**, the Systematic Nomenclature for Medicine [101] which, despite the name, is a full ontology rather than just a terminology system, i.e. it stores relationships between entities as well as their codes and names. SNOMED is a comprehensive but precise system that spans the whole range of events, activities and entities involved in healthcare and research. Because of this it is increasingly widely used in healthcare systems, and increasingly integrated into EHRs, as a general machine readable vocabulary for health, illness and related activities – for example its use is now mandated as the 'structured cli'Dnical vocabulary' in the UK's NHS, in both primary and secondary care. The growing use of

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

SNOMED in the healthcare environment is one of the reasons for its importance within OMOP (which, as part of a 'virtual circle', in turn makes it more attractive within the clinical area).

Although the system's roots are in the US, SNOMED is now explicitly an international organisation with global ambitions. Membership is on a country by country basis, with 40 countries currently listed, with membership fees weighted according to the country's GDP. Organisations in a member country can access SNOMED resources for free, (though they normally need to be associated with the national manager of the system, e.g. in the UK be part of or linked to the NHS).

Organisations not in a member country have to pay to use the system, just under two thousand US dollars a year if in a richer country. Exemptions exist, including for development and some research projects, and there is a free, simplified version of a subset of SNOMED, known as the Global Patient Set or GPS [102]. But the licensing model can make things complicated for large international projects. In Europe, for example, most countries are SNOMED members, but France, Italy and Poland are not. Furthermore, even if a country is a SNOMED member, few have emulated the UK and mandated its use everywhere – in general adoption appears to be on a hospital by hospital, or project by project, basis. In addition, SNOMED exists in various different national 'flavours', which may reduce interoperability.

In time, SNOMED may become more widely used in healthcare systems, perhaps even a *de facto* standard. There are also multiple projects looking at how SNOMED can be mapped to other existing CTs, including ICD, LOINC and DICOM [103]. SNOMED is thus one of the most powerful and comprehensive CTs, and widely used in healthcare, though as discussed later its use in interventional research, at least for now, remains limited.

The other major system used within OMOP is **ICD**, the International Statistical Classification of Diseases and Related Health Problems, (to give the system its full title), which is maintained by the WHO [104]. They claim that the "ICD defines the universe of diseases, disorders, injuries and other related health conditions, listed in a comprehensive, hierarchical fashion that allows for: Easy storage, retrieval and analysis of health information … Sharing and comparing health information between hospitals, regions, settings and countries …(and) Data comparisons in the same location across different time periods".

The focus is therefore on morbidity and mortality data expressed in terms of diagnoses and related factors, and ICD codes are used for this purpose in healthcare as well as in epidemiological data. In non-commercial clinical research, if diagnoses have been coded at all, they also often make use of the ICD classification.

One difficulty is that ICD exists in slightly different versions in different countries, to allow for 'local adaption' of the system, as well as sometimes having different major versions in operation at the same time. This can recreate exactly the interoperability problems any standard terminology is supposed to eliminate, and may require a mapping to be established between different versions of the same standard – in one published example, between the French version of ICD 10 (Classification Internationale des Maladies, 10e version, CIM-10) and the US version (International Classification of Diseases, 10th Revision, Clinical Modification, ICD-10-CM) [105].

Finally, in terms of drug coding, OMOP favours **RxNorm** rather than the more widely known (in Europe) ATC and WHODrug systems. RxNorm originated in the US but seems to be finding wider

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

use - it “provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software” [106].

Despite its extensive list of Standardised Vocabularies, efforts to use OMOP outside of the routine care context have sometimes found the CT schemes wanting. Attempts to use the system with specialised oncology [107] and biospecimen data [108] both required the terminology systems to be extended, as did efforts to use OMOP with sleep research data [109], and to aggregate French [110] and German [111] EHR records. Conversely efforts to use OMOP with data from the large US ‘All of Us’ longitudinal study, [112] and with nursing research data [113] were reported as largely successful – perhaps suggesting an anglophone bias within the system.

3.4 Metadata standards

There does not appear to be any widely used or recognised metadata standard within healthcare data in general, or observational research data in particular, though HL7 profiles are a form of metadata for the data transferred from a FHIR API.

Within OMOP, the ‘White-Rabbit’ tool is used to provide a metadata map of the *source* database(s). The OMOP system that is created after ETL will have the standard table structure of the current version of OMOP but, because the system includes an EAV element, that is only part of the story. It is also necessary to use a further tool, ‘Usagi’, to do ‘concept mapping’ from the data elements in the source to the concepts within the standardized vocabularies. The mapping files produced (which normally need manual processing after the Usagi tool has been used) can then serve as a record of the EAV elements.

This is clearly far from ideal. The development of a simple machine readable metadata schema for data items, that could be applied to all health data datasets, plus the tooling to apply and read it, would be a relatively simple but important step forward.

4. Data Standards and Interoperability in Interventional Research

4.1 Outcome measures

Within clinical trials, outcome measures are pre-specified by the researchers as the ‘end-points’ of the study, the variables used to categorise the result of the interventions on any particular individual.

In early phase research such outcomes are likely to be detailed physiological or specialist laboratory measures, used as proxies for clinical measures of efficacy because of the usually short time frame of the trial. In later and longer phase III trials, more ‘realistic’ measures based on clinical assessment and routine tests are more likely, measures which can overlap with those used in real world practice. For comparative effectiveness trials, when the comparison is between two or more ‘usual treatments’, this overlap can – and probably should – be complete, allowing either research derived outcomes to be introduced into the routine assessment of a service’s efficacy, or clinical outcome measures to be assimilated into clinical trials.

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

Within clinical research, the main organisation promoting the standardisation of outcome measures is the COMET initiative (Core outcome measures in effectiveness trials) [114]. COMET does not generate outcome measures itself, though it does provide tools to help with their creation, but it does make available a searchable database of Core Outcome Sets, as developed by many different research groups.

A Core Outcome Set or COS is defined by COMET as *“an agreed standardised set of outcomes that should be measured and reported, as a minimum, in all clinical trials in specific areas of health or health care."* Outcomes in any particular study are therefore not limited to the relevant core outcome set(s), but they should always *include* the COS so that results – at least for these core data items – can be compared across studies.

The COMET database was reported, during the last annual review [115], as having 370 published COS studies (just over 400 by August 2021). That total includes most of the standard sets developed by ICHOM. The database also includes links to a similar number of other related studies, for example systematic reviews of the outcomes used in trials in different disease areas.

COMET is based in the UK but is global in scope, and appears to be the only major resource for outcome measures with direct relevance to clinical research. Importantly, because the emphasis is on effectiveness trials, the great bulk of the published COSs are labelled as also being available for use in clinical practice as well as research.

4.2 Syntactic and transport standards

Within interventional research the picture for syntactic standards is mixed. Data standards are available and are used – *at least within the commercial research sector*.

Up to about 15 years ago there was little consistency in the way result data was structured and coded within clinical trials, even for the commonly used types of data, such as adverse events and medical history, and thus very little direct interoperability between datasets. The FDA, however, faced with a bewildering array of datasets to assess, has provided steadily increasing pressure for the data submitted in pursuance of a marketing authorisation to be standardised. It now *requires* that such data are submitted, structured and coded using the CDISC Standard Data Tabulation Model, or SDTM. The Japanese authority, the PMDA, makes the same demand, and the use of SDTM is also preferred by the Chinese NMPA. [116] Of the major regulators, only the EMA has so far not set a similar requirement.

CDISC (originally known as the Clinical Data Interchange Standards Consortium, but now just as CDISC) originated in 1997 and became incorporated and funded, by its member organisations, from 2000. Though US-based, and with a close relationship to the FDA, it is global in scope. Over the past 20 years, CDISC has developed a wide range of standards and tools for different purposes [117], the most relevant of which are listed in Table 7. Of those listed, SDTM, CDASH and Define-XML are the ones that have the greatest potential to improve interoperability.

The commercial sector of clinical research has had no choice but to embrace data standards, and pharmaceutical companies, and / or their CROs (contract research organisations), have had to invest in personnel and systems to support their use. Not every interventional study’s data may need to be turned into SDTM, but study data collected by commercial entities are often collected using CDASH, in case SDTM is required.

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

In the non-commercial sector, however, there has not been the regulatory incentive or the resources to support a similar move towards interoperability. Many university and hospital trials units have experimented with elements of the CDASH standard, and a few use it extensively, but in general the uptake of CDISC standards in academic interventional research has been limited. Individual researchers, carrying out studies in isolation without the support of a trials unit, are even less likely to be aware of, or use, data standards from CDISC or elsewhere.

Name	Description / Use	Current Version
SDTM	SDTM provides a standard for organizing and formatting data in tables, to support data collection, management, analysis and reporting, data aggregation and warehousing, data mining, reuse and sharing, due diligence and other data review activities, and the regulatory approval process.	SDTM v1.8 17/9/2019 SDTMIG v3.3 20/11/2018
CDASH	CDASH establishes a standard way to collect data consistently across studies and sponsors, with data items then easily mapped to SDTM. This allows more transparency to regulators and others who conduct data review.	CDASH v1,.1 1/11/2019 CDASHIG v2.1 1/11/2019
Define-XML	An extension of ODM-XML that structures metadata that describes any tabular dataset structure. It can be used with SDTM but is not restricted to that standard.	Define-XML v2.1 8/5/2019
ODM-XML	ODM-XML is designed for exchanging and archiving clinical data, along with their associated metadata, administrative data, reference data, and audit information.	ODM-XML v1.3.2 1/12/2013
ADaM	The Analysis Data Model, ADaM, defines dataset and metadata standards that support efficient review of clinical trial statistical analyses.	ADaM v2.1 7/12/2009 ADaMIG v1.2 3/10/2019
TA Standards	Therapeutic Area User Guides (TAUGs) demonstrate how CDISC standards can be applied to specific disease areas.	Issued and revised on an ad hoc basis
CDISC CT	CDISC Controlled Terminology is the set of CDISC developed (or adopted) standard expressions / values used with data items within CDISC-defined datasets. Managed in collaboration with the US NCI's Enterprise Vocabulary Services (EVS).	Updated regularly and published quarterly.

Table 7: Key CDISC Standards (largely from the CDISC website. IG = implementation guide).

One recent development that may increase CDISC adoption has been the introduction of ‘platform trials’, e.g. as within the ECRAID [118] and VACCELERATE [119] projects. These are designed to provide a stable infrastructure of clinical observational research sites, each with expertise in a particular scientific area, and all collecting data against a master protocol, to which adaptive trials can be added as required.

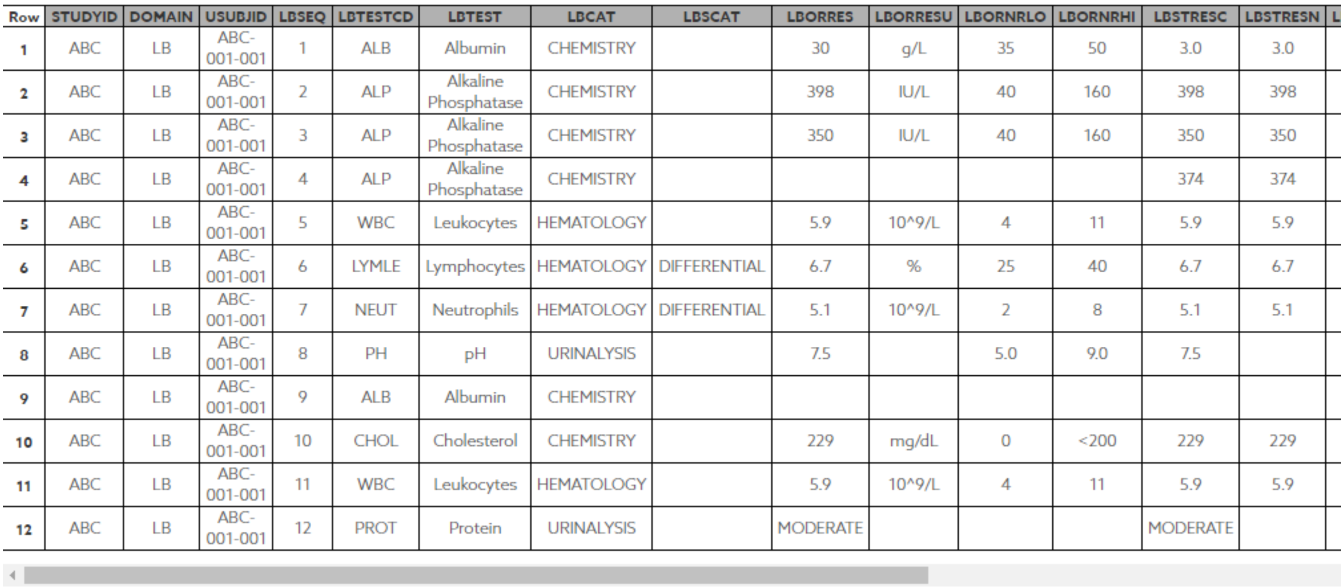
Both ECRAID and VACCELERATE were established within the EU’s Horizon 2020 research program, but are designed to encourage commercial sponsors to ‘embed’ trials within the platform, to maximize the utility of the data and to help ensure the sustainability of these large-scale initiatives. To make that proposition more attractive, for sponsors who may wish to use their trial data in an application to the FDA and / or PMDA, both these research networks have opted to use CDISC standards to structure data collection, specifically CDASH. The need to connect to sustainable funding has therefore provided non-commercial trials units with a powerful incentive to investigate, understand and use CDASH.

The SDTM standard is the main CDISC Common Data Model. It delivers data in a series of tables, each one dealing with a different 'domain' according to its content, and each of which becomes a separate (SAS) file in the submission dataset. Each domain is associated with a two letter code. The current domains are listed in appendix 1, and include

- 6 classified as 'events', (e.g. AE = Adverse events, DV = Protocol deviations),
- 7 as 'interventions' (e.g. PR = Procedures, CM = Concomitant / prior medication),
- 30 as 'findings' (e.g. LB = Laboratory test results, QS = Questionnaires) and
- 5 that are 'special purpose' (e.g. DM = Demographics, CO = Comments).

The system is comprehensive but so is the documentation – the SDTM specification [120] has 45 pages, but the implementation guide [121] has over 500. The system has grown steadily over the years and continues to do so – currently the CDISC wiki lists a further 7 draft domains (e.g. ER = Environmental and Social factors, GF = Genomic findings) and 17 'draft domains under construction' (e.g. BE = Biospecimen events, SI = Site summary). In most cases the content of these draft domains can also be used, but may be changed slightly in later versions [122].

SDTM records in most tables tend to have a relatively small cluster of fields that describe a single event, intervention or finding, e.g. laboratory test data has 'One record per lab test per time point per visit per subject ', while questionnaire data is structured as 'One record per questionnaire per question per time point per visit per subject'. There are therefore very often multiple records per visit for a single participant, as shown in figure 5.



Row	STUDYID	DOMAIN	USUBJID	LBSEQ	LBTESTCD	LBTEST	LBCAT	LBSCAT	LBORRES	LBORRESU	LBORNRLO	LBORNRHI	LBSTRESC	LBSTRESN	L
1	ABC	LB	ABC-001-001	1	ALB	Albumin	CHEMISTRY		30	g/L	35	50	3.0	3.0	
2	ABC	LB	ABC-001-001	2	ALP	Alkaline Phosphatase	CHEMISTRY		398	IU/L	40	160	398	398	
3	ABC	LB	ABC-001-001	3	ALP	Alkaline Phosphatase	CHEMISTRY		350	IU/L	40	160	350	350	
4	ABC	LB	ABC-001-001	4	ALP	Alkaline Phosphatase	CHEMISTRY						374	374	
5	ABC	LB	ABC-001-001	5	WBC	Leukocytes	HEMATOLOGY		5.9	10^9/L	4	11	5.9	5.9	
6	ABC	LB	ABC-001-001	6	LYMLE	Lymphocytes	HEMATOLOGY	DIFFERENTIAL	6.7	%	25	40	6.7	6.7	
7	ABC	LB	ABC-001-001	7	NEUT	Neutrophils	HEMATOLOGY	DIFFERENTIAL	5.1	10^9/L	2	8	5.1	5.1	
8	ABC	LB	ABC-001-001	8	PH	pH	URINALYSIS		7.5		5.0	9.0	7.5		
9	ABC	LB	ABC-001-001	9	ALB	Albumin	CHEMISTRY								
10	ABC	LB	ABC-001-001	10	CHOL	Cholesterol	CHEMISTRY		229	mg/dL	0	<200	229	229	
11	ABC	LB	ABC-001-001	11	WBC	Leukocytes	HEMATOLOGY		5.9	10^9/L	4	11	5.9	5.9	
12	ABC	LB	ABC-001-001	12	PROT	Protein	URINALYSIS		MODERATE				MODERATE		

Figure 5: Portion of a SDTM Lab Results (LB) table

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

The result is that SDTM provides relatively long, ribbon-like tables of data, often referred to as being 'normalised' (though they are not, in a formal relational database sense). Most of the data points in any particular table are prefixed with the same two letter domain code, whilst the rest of the data item's code consist of one of a set of standard suffixes that indicate its purpose, as defined, and often exemplified and discussed, in the CDISC documentation. For example, in Figure 5, –TESTCD and TEST refer to the lab test code and name, --CAT and SCAT to its category and subcategory, --ORRES and ORRESU to the result and units as originally reported, --ORNRLO and --ORNRHI as the low and high ends of the normal range in the original units, etc.

SDTM data structures therefore generally follow the EAV (Entity-Attribute-Value) pattern, where the attribute (the specific event, intervention or finding of interest) is identified as part of the record content, and not fixed as a table column heading. This maximises the system's flexibility. It is a reasonable fit for types of data that have often been collected this way (e.g. adverse events and concomitant medication) but not for others (e.g. laboratory test results or vital signs) where traditionally all data from a single participant / visit are tabulated and analysed together.

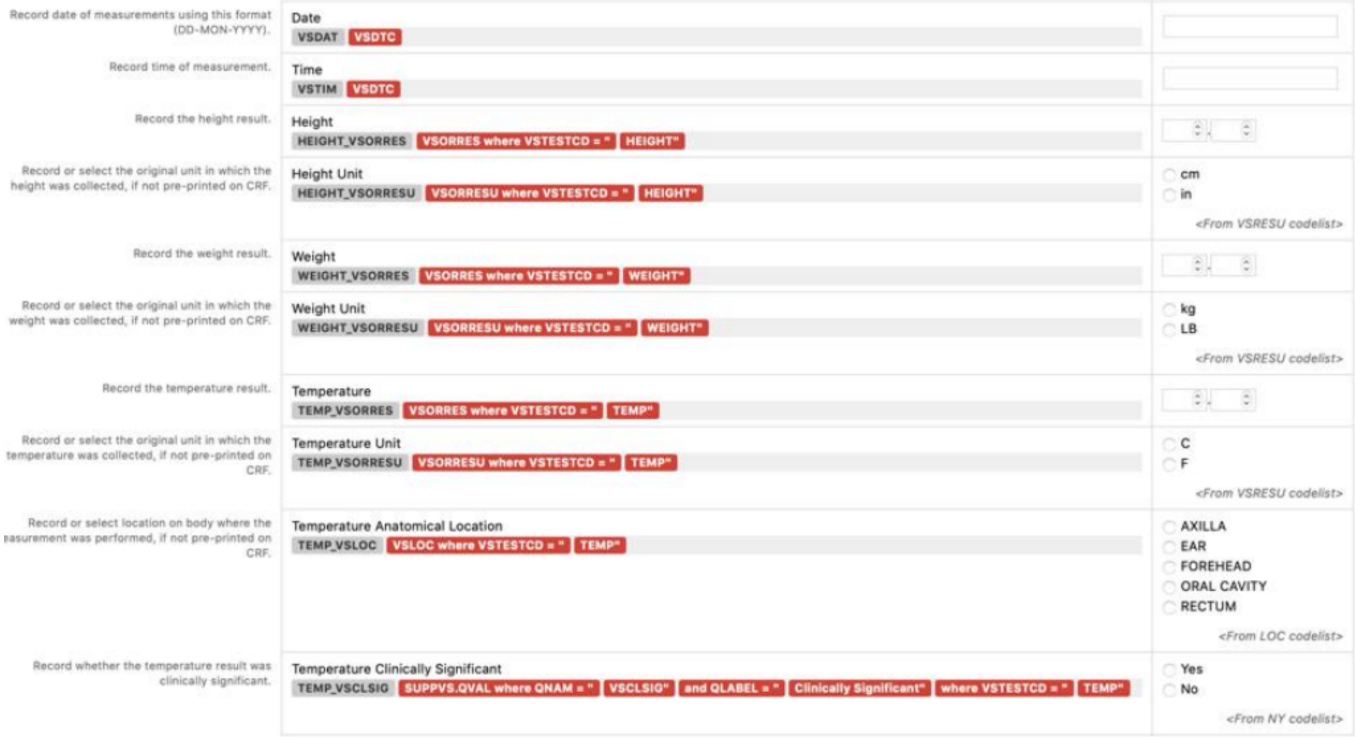
The consistent, relatively simple and fixed data structure of SDTM makes identifying and using interoperability between datasets much more straightforward. It can be difficult, however, to transform data, as traditionally collected and organised, into SDTM tables. To help support the use of SDTM, CDISC developed CDASH – as part of the 'CDISC Data Acquisition and Standards Harmonisation' project [123]. CDASH shares many of the features of SDTM: data is split into domains, using the same two letter prefixes, and again has code suffixes indicating the purpose of the data item.

Appendix 2 tabulates the suffixes available to all the 'findings' domains, though some domains (e.g. EG for ECG findings) have specific additional item types available to them. CDASH is an enabling technology – it is not as rigid as SDTM in the way the data is structured. Data can be provided in the 'normalised' format, but it can also be assembled in a traditional non- normalised pattern, with the implementation guide describing how this can be done [124]. Figure 6 is taken from that guide and illustrates a non-normalised form for collecting some vital signs data with, for example, the WEIGHT_VSORRES and WEIGHT_VSORRESU CDASH fields being transformed into the SDTM VSORRES and VSORRESU fields, alongside a VSTESTCD field that has the value 'WEIGHT' – a simple transformation into an EAV format.

In addition, CDASH recognises that data points are often pre-populated or implicit on eCRFs, when only a confirmation (e.g. of a particular reaction, test or diagnosis) is required, that data points may not be available (in which case a reason for their absence should be provided), and that some data points are populated not from the site but by later coding. The greater flexibility available for organising CDASH data items means direct compatibility between studies can be partially lost, though there is nothing to stop local policies making CDASH implementations within a group of studies as consistent as possible. The degree of interoperability between datasets provided by CDASH may be sufficient for easy aggregation or comparison, but if that is not the case the data can again be easily transformed to the more consistently structured SDTM, which should improve interoperability. Transformation to SDTM (outside a submission process) is an optional step that only needs to be taken, and funded, if and when required.

CDISC has also produced a set of therapeutic area user guides (or TAUGs), that show how SDTM, and often also CDASH, can be used to structure data within a particular disease area. 48 of these

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.



Record date of measurements using this format (DD-MON-YYYY).	Date VSDAT VSDTC	
Record time of measurement.	Time VSTIM VSDTC	
Record the height result.	Height HEIGHT_VSORRES VSORRES where VSTESTCD = " HEIGHT"	
Record or select the original unit in which the height was collected, if not pre-printed on CRF.	Height Unit HEIGHT_VSORRESU VSORRESU where VSTESTCD = " HEIGHT"	○ cm ○ in *<From VSRESU codelist>*
Record the weight result.	Weight WEIGHT_VSORRES VSORRES where VSTESTCD = " WEIGHT"	
Record or select the original unit in which the weight was collected, if not pre-printed on CRF.	Weight Unit WEIGHT_VSORRESU VSORRESU where VSTESTCD = " WEIGHT"	○ kg ○ LB *<From VSRESU codelist>*
Record the temperature result.	Temperature TEMP_VSORRES VSORRES where VSTESTCD = " TEMP"	
Record or select the original unit in which the temperature was collected, if not pre-printed on CRF.	Temperature Unit TEMP_VSORRESU VSORRESU where VSTESTCD = " TEMP"	○ C ○ F *<From VSRESU codelist>*
Record or select location on body where the easurement was performed, if not pre-printed on CRF.	Temperature Anatomical Location TEMP_VSLOC VSLOC where VSTESTCD = " TEMP"	○ AXILLA ○ EAR ○ FOREHEAD ○ ORAL CAVITY ○ RECTUM *<From LOC codelist>*
Record whether the temperature result was clinically significant.	Temperature Clinically Significant TEMP_VSCLSIG SUPPVS.QVAL where QNAM = " VSCLSIG" and QLABEL = " Clinically Significant" where VSTESTCD = " TEMP"	○ Yes ○ No *<From NY codelist>*

Figure 6: Extract of a sample vital signs eCRF showing CDASH (grey) and SDTM (red) fields

guides are currently listed, from Acute Kidney Injury and Alzheimer’s through to Vaccines and Virology, though they can only be downloaded by CDISC members.

4.3 Semantic standards

CDISC has its own controlled terminology system, CDISC CT [125], developed by the CDISC user community. The terms are published by US's NCI-EVS (National Cancer Institute - Enterprise Vocabulary Services), which also manages the development of CDISC CT value sets, published in the NCI’s Thesaurus. The EVS and thesaurus are also used by the FDA and NIH when developing controlled vocabulary, which ensures that the CDISC terminology is embedded in a larger ontology. Terms are provided as named lists – each list (or a subset of it) providing and defining arrays of questions, or the options or categories that can be used as allowed responses to a question.

The terminology files are available on the NCI’s EVS website - the CDASH file is a relatively small subset of the main SDTM file, which also includes very many questionnaires and other ‘standard instruments’ [126]. In general, the questionnaires and standard instrument lists are closed, i.e. they cannot be extended with new terms, whilst most of the other lists are extensible. Supporting information and documents are available on the CDISC website.

The strength of the CDISC CT is that it is tailored to the needs of clinical researchers, including lists like ‘Protocol milestones’, or ‘Completion / Reason for Non-Completion’ (of the study). Unless creating data for submission to the FDA its use is not mandated, but CDISC CT provides a useful,

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

standardised source of both questions and question options. A weakness is that a few of the non-technical lists (e.g., Race, Ethnicity) are too strongly US-centric.

CDISC also makes extensive use of MedDRA [127]. MedDRA is well known and widely used within clinical research because of its use in coding, reconciling, and summarising adverse events and SAEs. Originally developed by ICH (the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use) in the late 1990s, and updated regularly twice a year since, its use is mandated within the EU, the US and several other countries for adverse event reporting.

MedDRA has a hierarchical structure, with lower level terms converging on 'preferred terms', that converge to higher level terms, higher level group terms, and finally to 'system organ class' terms. The hierarchy is 'multi-axial', however, so that a particular item may have one or more parent items further up the tree, giving some flexibility in how an individual AE is classified. This and other issues mean that coding is usually done after data collection by staff specifically trained for the task. Although MedDRA browsers exist online, using the system for coding requires a licence, though for most non-commercial users the cost is small.

For coding drugs, SDTM often uses the WHODrug system [128]. Despite its name, WHODrug is not managed by WHO – though it had its origins in a WHO drug monitoring programme – it is coordinated instead by the Uppsala Monitoring Centre in Sweden, and is available under a paid subscription. It has been mandated by the FDA for drug coding (from 2019), as well as being 'preferred' by the PMDA – it is therefore used within the pharma industry but, partly because of costs, not widely within non-commercial clinical research.

Finally, (like OMOP), CDISC also supports LOINC, the Logical Observation Identifiers Names and Codes system [129], first developed in 1994 at the Regenstrief Institute [130], (a medical research institute linked to Indiana University), which still manages the system. For some time, LOINC was the only standard for naming and coding clinical tests, which were its original focus, though it has now expanded to cover other observations and measurements (just as other systems have expanded to cover tests). The system is free to use, and updated twice a year.

LOINC is used within healthcare generally, but in the US at least, LOINC codes are also mandated by the FDA in certain types of submissions for marketing authorisation [131], which is why they are supported by CDISC. Some, though not yet all, of the lab tests defined within CDISC CT have been mapped to LOINC codes, including all of the most commonly used tests as determined by an analysis of both LOINC and SDTM usage. (LOINC and SNOMED mappings also exist).

The usability of other major coding systems within CDISC SDTM – in particular SNOMED CT and ICD – is unfortunately very unclear. The SDTM implementation guide makes no mention of ICD and almost none of SNOMED – other than a brief reference as a parameter coding system allowed within the Trial Summary domain (but only that domain). In theory, codes and decodes (e.g. for medical history terms) can reference any CT dictionary, including ICD or SNOMED, as long as the dictionary and its version is identified in the associated metadata file [132]. In practice, there is a need for SDTM data to be compatible with the 'transport file' format defined by version 5 of the statistical analysis program SAS, because this is the submission format demanded by the FDA. This makes the use of some codes problematic, because they break the identifier rules for SAS v5 [133]. Many SNOMED codes are simply too long to use.

This is an anachronistic anomaly that urgently needs to be addressed, by the FDA as well as CDISC. As the author (Jozef Aerts) of the last reference asserts, in a blog post from March 2021, there is an urgent need for CDISC to make SAS V5 format files just one export format amongst many, rather than designing the system around its limitations, and for CDISC to "stop trying to keep … SNOMED-CT 'out of the door'".

4.4 Metadata standards

The metadata considered here is the detailed *descriptive metadata* that lists the data items in each table in a dataset, and for each indicates its name, code, type and meaning, plus any associated coding and / or categories. Such metadata is essential for understanding and using the data, for example within statistical analysis, and is usually created along with the data, by the data generators.

Descriptive metadata has often been just a simple table (like appendix 3) or a spreadsheet. This is a workable if clumsy approach when looking at an individual dataset but it prevents easy comparison or aggregation of metadata across different datasets, because the documents or spreadsheets are not structured consistently, and thus are not machine readable. What is required is a standard descriptive metadata format that is also machine readable, so that the metadata linked to several datasets can be presented, processed and queried more easily.

Within clinical research, CDISC has produced **Define-XML,** a metadata schema that has been developed (it is an extension of the older Operational Data Model, or ODM) specifically to describe datasets [132]. Define-XML's most common use is for describing the elements of the SDTM data sent as part of a submission to the FDA / PMDA, where its use is mandated. While describing non SDTM data is a little more work, it is still possible – in other words Define-XML *could* be used to describe any data set, including data in CDASH format.

Unfortunately, the creation of such files would normally have to be done manually, and so be resource intensive. Although several research database systems export ODM-based metadata files, few use the newer Define-XML standard. Detailed descriptive metadata presented in a consistent, machine-readable fashion, at least outside the pharmaceutical industry, is therefore likely to remain rare, unless more tools are developed to address this problem.

5. Interoperability of RWD and Interventional Research Data

This section considers the *practicalities* of interoperability, in particular of making RWD data, organised using the emerging standards in that area (OMOP, SNOMED etc.), interoperable with interventional research data, organised according to the standards existing in that area (CDISC, MedDRA etc.).

The wider questions, of whether such interoperability is necessary, or even desirable, and when, are left to the Conclusions.

The related questions, of whether interoperability should be increased *within* healthcare data, and *within* clinical trial data, are not considered in detail. The assumption is that most people would agree that they should be, to promote the FAIRness and the secondary re-use of data in

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

each domain, enabling a more powerful and more efficient pool of data for both research and healthcare management.

Increasing data standards within either domain would not necessarily be straightforward, or quick – there are many resourcing, system and training issues that would need to be overcome. For example, persuading more non-commercial researchers to use CDISC standards, or finding a way of safely linking organisation-centric health records so that they became person-centric, would be very substantial projects in their own right. We emphasise, however, that there is nothing *in the data* to prevent, in principle, the further development and uptake of data standards within each of these domains, and one might hope that such a progression will be encouraged by funders, regulators and other stakeholders. Of course, the greater the use of data standards in healthcare and interventional research, the greater the importance of the issue of interoperability between the two systems of standards.

5.1 Interoperability in Outcome measures

As outlined previously, there are some initiatives promoting common outcome measures, such as COMET and ICHOM, and there is considerable scope for overlap between the outcomes applied in comparative effectiveness research and those likely to be used in clinical practice. As usual, hard empirical evidence about the extent of usage of core outcome sets is hard to find, even within clinical trials, let alone within healthcare. The impression is that, whilst interventional research always defines outcomes, and the use of core outcome sets is slowly increasing, any alignment between clinical trial endpoints and the outcome measures available in RWD is, currently, likely to be down to luck rather than deliberate planning.

This represents a large potential source of data going untapped, especially with regard to patient-centric outcomes, which tend not to be dependent on specialist measuring techniques and which are therefore easier to gather (as well as being, one could argue, a more valid measure of outcome).

A recent paper by LoCasale et al. (*Bridging the Gap Between RCTs and RWE Through Endpoint Selection* [59]) tackles this issue from the point of view of the trialist. They provide a diagram, reproduced here as figure 7, that depicts a framework for deciding if a real world outcome could be used within a trial, and equally importantly, suggest possible actions if it does not.

Thus, if a trial endpoint is already available in the real world data, and used within routine practice, then that RWD is potentially usable as a data source (notwithstanding possible quality issues, as discussed previously). If the endpoint is not yet routinely available, but reflects a patient-centric measure, then it may be possible to create systems in clinical practice to capture it directly from the patient, e.g. using Patient Reported Outcome methods (e.g. a questionnaire, or an app). If neither is the case, then it may be possible to explore 'mosaic' designs, where trial-specific outcome measures are used alongside specialism-specific RWD outcomes, in the process providing

> "a broader spectrum of insights from across the different data sources while simultaneously generating evidence to appease the requirements of different decision-makers including regulators, HTA/payers and healthcare providers." [59]

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

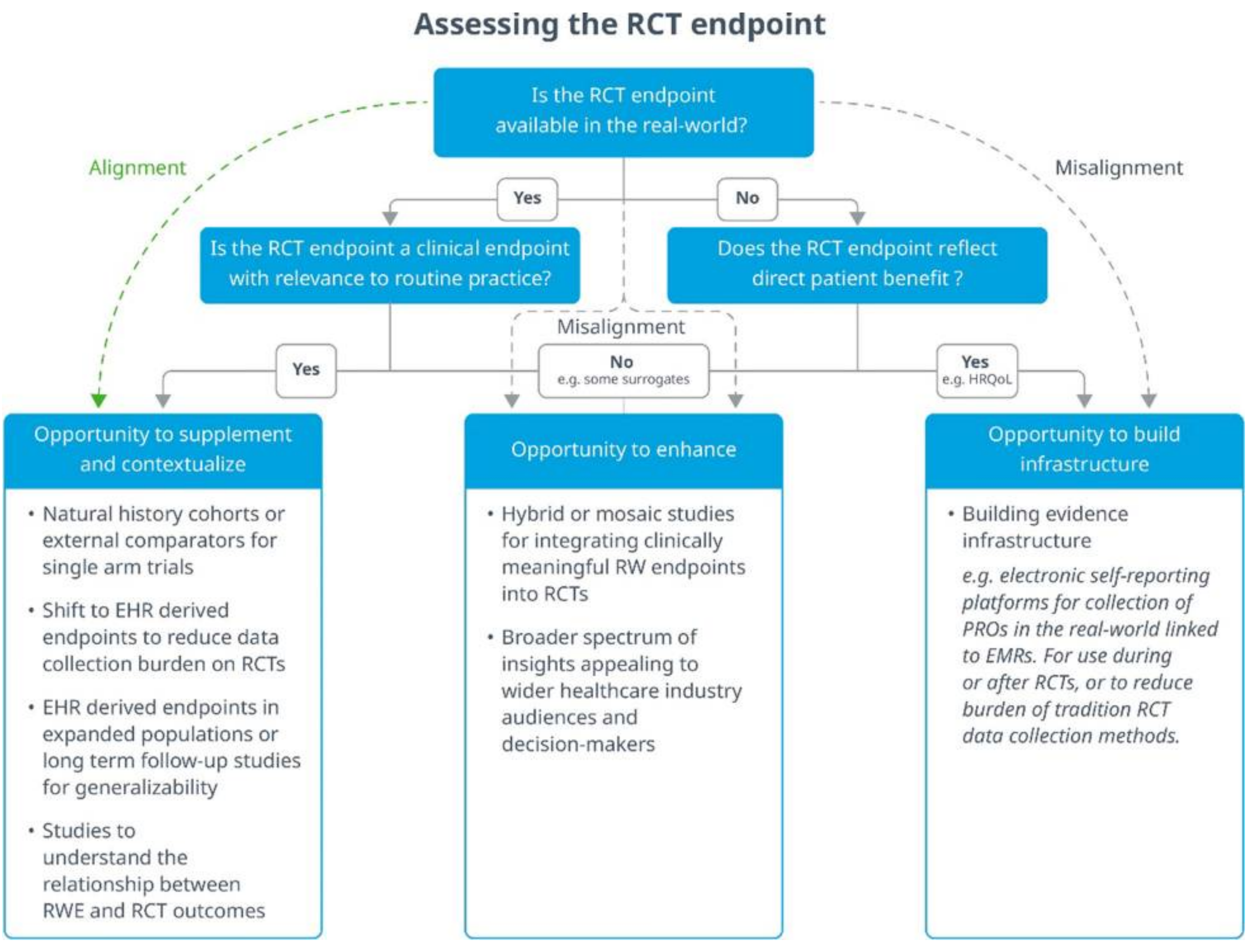

Assessing the RCT endpoint
Is the RCT endpoint available in the real-world?
Alignment
Misalignment
Yes
No
Is the RCT endpoint a clinical endpoint with relevance to routine practice?
Does the RCT endpoint reflect direct patient benefit ?
Misalignment
Yes
No
e.g. some surrogates
Yes
e.g. HRQoL
Opportunity to supplement and contextualize
• Natural history cohorts or external comparators for single arm trials
• Shift to EHR derived endpoints to reduce data collection burden on RCTs
• EHR derived endpoints in expanded populations or long term follow-up studies for generalizability
• Studies to understand the relationship between RWE and RCT outcomes
Opportunity to enhance
• Hybrid or mosaic studies for integrating clinically meaningful RW endpoints into RCTs
• Broader spectrum of insights appealing to wider healthcare industry audiences and decision-makers
Opportunity to build infrastructure
• Building evidence infrastructure
e.g. electronic self-reporting platforms for collection of PROs in the real-world linked to EMRs. For use during or after RCTs, or to reduce burden of tradition RCT data collection methods.

Figure 7: Framework of pathways to integrate RWD within clinical trial endpoints (from [59]).

Patient registries, especially those with a quality focus, are likely to have an important role both in standardising outcome measures for a particular condition or speciality, and in making that data available for research. Some registry based research has already taken place, e.g. the hybrid registry / RCT SWEDEHEARET-VALIDATE trial [30]. Recent attempts to use registries more pro-actively by the EMA and the rare disease research community may be indicative of a greater future role for registries in this respect.

In summary, it is clear that increased alignment of outcome measures has great potential value both to practice (the 'learning health system') and to pragmatic research strategies, and that a much greater degree of interoperability is almost certainly possible. The barriers to increased use are partly technical (how to extract the outcome data most easily from real world systems, though this problem is reduced if the data is already being extracted for management or quality monitoring purposes) but are also about resources and coordination of effort. What could be useful are additional cost-benefit analyses of pilot projects, in both financial and scientific terms, to clarify the issue for funders and researchers alike.

5.2 Interoperability in Syntactic and Transport Standards

In theory, it should always be possible to transform the organisation of data points, from one set of tables to another, using an ETL (extract – transform – load) process. While such transformations can be carried out, they can be complex, may introduce errors, require extensive validation and will need revision each time either the source or destination systems change. They also depend on being able to place all the source data points somewhere in the destination system and, in addition, there may also be technical constraints in the destination system that can impact the practicality of the transformation process.

The central question is therefore whether OMOP data can be converted into CDISC data using an ETL process, (leaving aside for the moment the issues of semantic compatibility). CDISC SDTM and the OMOP CDM are both 'traditional' relational database systems. They also both rely on an EAV approach – SDTM more explicitly than OMOP – to deal with the variety of data that they have to hold. Individual data points are located in different parts of each system but assuming their location is understood, then in most cases it should be possible to map the points to each other using standard data manipulation scripts, for example using SQL. **In other words, purely syntactic interoperability issues can normally be overcome**, with good knowledge of each system. The process becomes a data transformation exercise. So, for example, points in the Person table can be mapped to the DM demographics domain, points within the health system data tables to the HO domain, and points in the Drug exposure table to the EX domain.

In practice, as discussed in the previous section, there appear to be some limitations within CDISC that could impede any ETL process – in particular the need to conform to short codes and column names to maintain compatibility with the legacy version of SAS still favoured by the FDA. Further work would be needed to clarify the nature and extent of the problem. Despite this issue, a variety of combinations and mappings of the different standards are to be found as proposals or discussions in the literature [e.g. 134].

Significantly, a 2021 press release announced the collaboration of OHDSI and HL7, with the aim of linking OMOP and FHIR [135]. The material presented at the initial meeting to discuss this collaboration is available at [136]. This development seems likely to further augment the importance of both systems, and gives a clearer indication that the future of data standards in the healthcare sector – especially in Europe – is likely to be built around OMOP and HL7 FHIR.

At the same time, CDISC is also involved in a variety of initiatives with HL7, around linking FHIR and various forms of CDISC data [137]. This work includes 3 different projects involving CDISC, PHUSE and HL7 FHIR. It also reflects the results of a survey CDISC carried out in 2019, amongst its stakeholders, on working with RWD. The two priorities identified were that CDISC should focus on connecting directly with EHRs and with HL7 FHIR [138].

This raises the possibility of HL7 FHIR acting as an intermediary in any data transformation process, as summarised by figure 8. This may represent a more pragmatic solution to the problem of transferring data between the SDTM and OMOP.

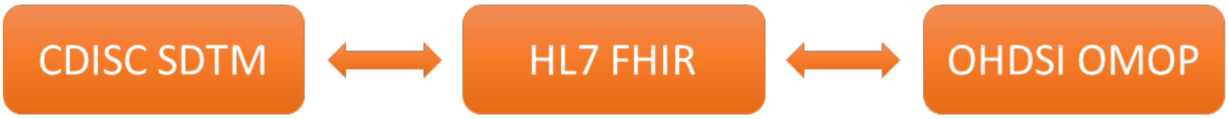


Figure 8: Possible intermediate role for HL7 FHIR

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

5.3 Interoperability in Semantic standards

As described in previous sections, and summarised in table 8, below, there is a clear dichotomy between the major CTs favoured by CDISC and those favoured by OMOP. This is partly due to their development histories and design decisions, partly due to regulatory requirements, and partly due to the need for compatibility with existing data.

CDISC	OMOP
CDISC CT	SNOMED
MedDRA	ICD
LOINC	LOINC
WHODrug	RxNorm

Table 8: Major controlled terminologies supported by CDISC and OMOP

The problem is that differences in CTs are embedded in the data and cannot be easily eliminated, except by a mapping strategy that almost always has inherent flaws. Mapping, identifying corresponding points in different CTs to allow translation between them, is often only partial because coding systems are devised for different purposes or based on different assumptions. Usually information and precision is lost in one or both directions, as related but distinct codes in one CT system are all mapped to a single, coarser, classification in the other.

This is clearly a major barrier for data interoperability between the healthcare and clinical research sectors, and there is no easy way around this problem – one cannot easily 'map' SNOMED (352,567 concepts and 1.36 million links, covering all of healthcare) to CDISC CT (about 32,000 terms focused on interventional research). It might be possible, if a great deal of work, to map the CDISC CT and MedDRA systems to SNOMED, but, as discussed above, any such mapping is likely to be partial – switching controlled terminologies involves much more than substituting one set of codes for another, because it changes the ways in which a domain is conceptualised and categorised.

It might also be possible to use a set of CDISC NSVs (non standard variables) to receive SNOMED codes and decodes, but this may not find favour with regulatory authorities and in any case defeats much of the purpose of importing the data into a CDISC format – an important part of the data will not truly be in that format, and available only as an 'add-on' file, making it much more difficult to use. For analysis purposes, in most situations, it will still need mapping.

SNOMED seems to becoming more widely used in healthcare systems, and therefore also observational research, (although again there is a lack of empirical data to give exact figures) and it may become a *de facto* standard. Its use in interventional research, however, appears to be relatively low. Although integrated into EHR systems, it does not seem to be integrated into the clinical data management systems used by trialists, perhaps because of the system's complexity, licensing issues, or training needs.

The healthcare sector, which of course is much bigger than the health research sector, is not going to start to use CDISC CT, nor is there any suggestion that they should. The obvious answer to this problem is therefore for CDISC to explicitly allow (even encourage) the use of SNOMED and ICD, including explaining how best that can be done. Similarly, the insistence on using WHODrug needs to be replaced by also allowing RxNorm. These changes would require agreement by the

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

FDA and PMDA, but these are the very bodies who, with their interest in using RWD, potentially have the most to gain from such a change. It would also need the main CDISC stakeholders and funders to agree, but these are largely pharmaceutical companies, who again would have a lot to gain from the easier integration of RWD into research data.

This change would be a major move for CDISC but it seems unavoidable if they (and their major stakeholders) are to seriously engage with RWD and support its use in interventional research and related activities. In the longer term, as discussed in the Conclusions section below, an even more radical approach is required.

5.4 Interoperability in Metadata

There is very limited support for metadata standards in healthcare, and while interventional research has Define.xml this is not as easy as it should be to use unless the described dataset is in CDISC SDTM. There is an urgent need for a single, machine readable metadata schema for both types of datasets. A more generic and simpler version of define-XML would probably be the best choice for such a schema, and it could probably be developed relatively easily.

The caveat is that the need to understand the item names used within any metadata file, for example for consistent querying, returns us to the problem of controlled terminologies. Metadata files describing datasets from SDTM and OMOP might be constructed using the same schema, which would certainly be an improvement on the current situation, but full interoperability of the metadata only comes when the field and table names listed in the metadata use the same vocabulary.

6. Conclusions

6.1 Summary of current position

The ‘standards landscape’, described in the previous chapters, can be summarised as follows:

- There is an almost total lack of independent empirical data about the level of use of data standards, of any sort and in any sector. Claims are made but are difficult to verify. The statements below should be read bearing this in mind.
- Data standards are used extensively for interventional research in the commercial sector, based on CDISC standards and associated vocabularies, because of the pressure from regulatory authorities. These standards cover both syntactic and semantic aspects of the data.
- There is a very variable but generally low usage of data standards within non-commercial interventional research. The main syntactic standard, if one is used at all, is CDISC’s CDASH. The principal semantic system is probably MedDRA but this is only used for adverse event reporting.
- Standardisation of data structure in RWD is largely achieved by extracting the data and transferring it to a ‘bolt-on’ common data model or CDM. In Europe the most important CDM is OMOP, and OMOP installations appear to be rapidly growing in numbers. Despite this, probably only a few percent of healthcare providers have access to any form of data standardisation.

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

- There appears to be an increasing use of controlled terminology systems within healthcare (and thus also within a lot of interventional research) – in particular SNOMED, but also LOINC and ICD. The same controlled vocabularies are found within OMOP.
- Transforming healthcare data in an OMOP structure into a CDISC SDTM structure should normally be possible, but for some data could be impeded by structural constraints within CDISC related to file formats. These constraints could be removed, however, relatively easily.
- Transforming data points expressed in OMOP linked CTs into those currently supported by CDISC will be very difficult. At best it is likely to be approximate and labour-intensive. The mismatch between CTs is the single most important barrier to interoperability.
- HL7 FHIR has quickly assumed an important role in standardising messaging between healthcare systems, and may facilitate use and transformation of data standards.
- The alignment of outcome measures between healthcare and registry data and that from late phase or comparative effectiveness trials is currently low, despite various initiatives, but could be greatly increased with potential benefit to both sectors.
- Standards for metadata are present within interventional research but rarely used outside submission datasets. A generic and easy to use metadata schema, with a linked toolset would be very useful in all sectors.

6.2 In the short and medium term

In this context there are several issues that could be usefully considered in the next few years (assuming the COVID pandemic does not continue to disrupt many activities):

[a] Do we even need interoperability?

As stated previously, there is little controversy about the desirability of improving interoperability *within* either or both of healthcare data and interventional research data, even if neither task would be trivial, but there remains the question of whether we really need interoperability *between* healthcare data and interventional research data. It is suggested that there are two aspects to this question:

- Is there a practical need, in the sense of being able to, as efficiently as possible, turn healthcare data into ‘research friendly’ formats so that it can be more easily aggregated, compared, submitted or processed along with interventional research data?
- Is there a scientific need, perhaps more concerned with the underlying justifications for FAIR data in science, to be able to turn real world data into a scientific resource, that can be more easily aggregated and / or compared with other data, derived from interventional research?

For the first question, the answer lies mostly with the data consumers. If the FDA and PMDA say they will consider RWD, as either prime or supplementary evidence of efficacy and / or safety, but that it does not have to be structured in the same way as ‘normal’ interventional research data, then clearly there is no imperative to convert it to a CDISC format. If they say that RWD also has to be presented as SDTM then interoperability assumes much greater importance.

Other data consumers, for example HTA agencies, may not have a clear policy, but they may express a preference – if only to make life as simple as possible. Data producers, in particular drug companies, may also want to simplify their own data processing by only dealing with (documenting, describing, analysing) data in a single structure.

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

Some of the answers to these questions may depend on the volumes concerned. If RWD is relatively rare it can be treated and 'excused' as an exception. If its use becomes more common there will be greater pressure to make it conform to the norm represented by conventional interventional research data.

The answer to the second question depends on a more fundamental, philosophical view of the nature of data. If the function of data is essentially to test a specific hypothesis, then there is no real need to make it available for comparison with other data. But such a view would be out of step with the prevailing idea, that data is a scientific resource that should be available, and understandable, as widely as possible, and that it should be FAIR.

For RWD to be FAIR does not mean that it must be collected, categorised and structured in the same way as other related data – most obviously that from interventional research – but it does mean that it should be possible to convert it to that format if and when required – interoperability should be available, even if not always implemented.

Ultimately these are questions for funders to clarify with stakeholders. But they need to be clarified, otherwise we risk either wasting effort, creating interoperabilities that are never used, or wasting data, because barriers to interoperability make it too difficult to fully exploit.

[b] Can we obtain better data about standard use?

A recurring theme in this report has been the lack of available empirical data, in particular about the real usage levels of various standards and systems – rather ironic for a report about data.

One of the things that would be very useful would therefore be mechanisms for gathering periodic, independent data about (for example) the number of healthcare facilities with OMOP, i2b2, and HL7 FHIR servers installed, or the number of regulatory submissions using RWD (and how that data has been standardized), or even the numbers of research studies using ICD, MedDRA, SNOMED, or any other CT. At the moment, such data comes from isolated research projects or in the shape of claims made by the proponents of particular schemes that do not always hold up to close examination.

Ensuring that researchers and health care systems report their use (or non-use) of standards will be difficult, but ideas should at least be explored, for example by including (and then publishing) such data in returns to governments, by routinely providing such data within regulatory decision summaries, or even by adding a structured section to published papers about the use of controlled terminologies.

Investigating, developing and testing mechanisms for obtaining data about data standards use could provide very useful data both from and to the European Open Science Cloud, and further guide input on data standards.

[c] Can we simplify licensing and funding of standards systems?

One of the features of standard systems is that they are available under a variety of different business models. Some are free, others use a membership model, others a licencing model – which in some cases (e.g. SNOMED) can be quite complex. This not only affects collaborative projects, where some users might have permission to use a licence, and some might not, it can also affect discussions about standards and systems.

It is difficult to discuss the merits of CDISC CT and SNOMED, for example, unless one can access both systems. But that requires an employer to have a licence for both systems, and few organisations will fall into that group – depending on their orientation (e.g., research university versus healthcare provider) they are much more likely to have one or the other. Staff therefore become 'locked in' to one system or another because that is all they can access. Discussions about systems, options, and future actions become compromised.

Standards development organisations obviously need financial support, but this problem raises the question of whether this can be simplified in some way, so that end users, and those examining different options for their own organisations, can access standards for free. For example, the SNOMED 'national' licencing model may be worth exploring and applying more widely, at least for non-commercial organisations. This is again a question for funders, but it is of central importance for the future take up of standards, as well as their future development.

[d] Can CDISC be supported in 'opening up' to other CTs?

In the section on interoperability it was asserted that for semantic interoperability to be possible in the short term CDISC SDTM needs to be able to accommodate SNOMED and ICD codes. Currently this seems to be difficult, or at least CDISC provides no guidance on how it can be done, outside of a very limited application within trial design data. More generally, there is an apparent need to drop the constraints that originate from adherence to the default export format of SAS v5 transport files (a format that originated in the 1980s). This is obviously a substantial change, and whether it can be implemented or not may depend in turn on the requirements of the FDA and PMDA – though a version of CDISC that was not so tightly tied to the submission function could be simpler to use for non-commercial researchers in any case.

Preliminary discussions with CDISC have indicated that they are not averse to discussing these issues, but it does raise the question of if and how CDISC could be persuaded and supported in making such a change. A necessary initial step would be to clarify the precise impact of the v5 format on SDTM and other standards – in particular on future interoperability. A joint project with CDISC experts, if one has not been carried out already on this topic, would require a relatively small investment but could be potentially very useful. Assuming such an investigation confirmed and clarified the nature of the problem, and that there was a consensus on the need for interoperability that involved CDISC (as discussed in [a] above), then further actions, including approaches to various agencies, could be discussed.

6.3 In the longer term

In the longer term there are two very substantial problems that need to be addressed, if the interoperability issue is ever to be resolved.

[e] Creating FAIR data by design

Systems like OMOP, using ETL to transform data into a CDM, are only necessary because of the heterogeneity of the source systems. Ideally, these standards would be pushed 'upstream', to be used within the source data itself, so that it became inherently FAIR. Funding, policy and accreditation pressures could be applied to promote this change. This is easier said than done, however, for a variety of reasons:

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

- There would need to be a consensus on what set of standards and models (OMOP?, PCORNet?, OpenEHR?) should be used.
- Previous initiatives in this area have often been unsuccessful, and are likely to be viewed with suspicion by vendors and users.
- Source system vendors have a commercial interest in differentiating themselves and 'locking in' their customers. Vendors have also, traditionally, been wary of sharing internal details of their systems.
- EHR systems are often integrated with a variety of other services. Unless that integration is only through well designed interfaces (unlikely given the age of many EHR systems) making substantial changes in internal structures will be very difficult and costly.
- The 'add-on' model, whether with OMOP, i2b2 and / or HL7 FHIR, seems well established.

What might be more feasible is the sort of semi-standardisation that CDASH provides for SDTM – a system where individual data elements are designed to make the ETL process simpler and safer, and where the controlled terminologies in the source data are already in line with those in the destination system. In a largely market-driven sector, we need governments, insurance companies and other funders to put pressure on healthcare providers, for them to demand that vendors provide, for example full SNOMED, ICD, LOINC, DICOM, RxNorm (etc.) capabilities within their systems. Again there would need to be discussion about exactly which CTs should be promoted, including which versions. That leads to the next, more general, but even more difficult point.

[f] Managing semantic interoperability

We have a plethora of CTs and ontologies – the BioPortal website of biomedical ontologies lists 947 (as of December 2021), of which 211 are classified as dealing with 'Health' [139], whilst the FairSharing site lists 186 standards using a search of 'Terminology artifact, Life Science, Humans' [140]. This is an embarrassment of riches, and while many are specialist schemas with restricted scopes, many also overlap.

The effort required for mapping even a fraction of these schemas against each other, as and when required for interoperability, but in perpetuity, is intimidating. Furthermore, the point has already been made that mapping is rarely completely accurate – a scheme that can be accurately mapped is redundant, but accurate mapping is rare because CTs and ontologies normally embody different assumptions and conceptualisations of the data.

The alternative to mapping would be a rationalisation of at least some of the CTs – using some combination of convergence and aggregation to reduce the number of CTs into a smaller set. In the particular context of healthcare and interventional research data, for example, to merge the RxNorm, WHODrug and ATC systems (as well as all the many national pharmacopoeias) into a single scheme for categorising drugs, to merge MedDRA and ICD (though first to merge the various versions of ICD back together) to a single scheme for diagnoses and symptomology, to merge CDISC CT, LOINC and SNOMED to a single ontology for both healthcare and research activity.

Of course such a proposal is very ambitious, even technically, without taking into account the organisations involved and their various traditions, connections and plans. It is also true that this approach has been less than successful in the past, as illustrated (literally) by an infamous xkcd cartoon [141].

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Nevertheless, the alternative to some degree of rationalisation is simply to carry on mapping imperfectly between schemes, essentially forever, with all that implies for the effort required to maintain data interoperability and FAIRness, and ultimately for the value of both healthcare and interventional research data.

Any CT rationalisation programme:

- Is likely to take decades, and will necessarily be planned and funded on a piecemeal basis, but that planning should occur within a larger, shared programme with clear long term goals.
- Will need to be pragmatic and make use of existing standards and schemes, and not try to re-invent the wheel.
- Must include all stakeholders and schema users, including industry and healthcare providers. In particular, it should strive to avoid becoming a purely academic exercise, with over elaborate ontological models that are then difficult to use in practice.
- Needs to respect and meet the different use cases behind existing CT schemes. For example, it is important that the research focus of the CDISC CT terms is not lost in any amalgamation with the larger SNOMED ontology.
- Should allow base terms to be grouped and classified in a variety of different ways, to help support different use cases.
- Should support different cultures and languages by optional levels of granularity, not different versions or systems.
- Should use mapping, but not as an end in itself. Mapping should be seen instead as an important device to identify the areas of similarity and difference between any two CTs.

Rationalisation of CTs requires that all the stakeholders involved can easily access the current schemes. In addition, to be widely taken up, any rationalised schemes should be free at the point of use. There will therefore also be a need to reduce the differences in the funding models used by the various standards bodies, and in particular move towards systems of national or international funding, rather than individual member organisations having to purchase licences.

There is an overarching body, the Joint Initiative Council, founded in 2007, that includes amongst its members CDISC, HL7, LOINC and SNOMED and whose first stated aim is to "Promote interoperability and seek to avoid overlaps and inconsistencies between standards used in health informatics" [142]. That council should therefore be involved in this programme, along with all the major standards development organisations, regulatory bodies and other data consumers, researchers in academia and industry, and healthcare providers and other data creators.

There is no dispute that such a programme would be difficult, lengthy and costly – but the long-term costs of not attempting it, financially, scientifically and ultimately, at a time when many challenges to health are emerging, to the health of populations, could be even greater. It is not likely ever to be a single named programme – but it is a 'direction of travel' that could be agreed upon by all stakeholders. Doing so would allow later funding and development decisions to be taken in that context. If FAIR data is ever to be a reality in clinical care and research, with data flowing easily between healthcare and research domains, then a programme like that described here, where all those involved in creating and using data work towards rationalising the way it is conceptualised, used and described, will be necessary.

References

[1] Thiese, M. Observational and Interventional Study Design Types; an Overview. *Biochemia Medica* 24, no. 2 (15 June 2014): 199–210. https://doi.org/10.11613/BM.2014.022.

[2] Makady A, de Boer A, Hillege H et al. What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. *Value in Health* 20, no. 7 (1 July 2017): 858–65. https://doi.org/10.1016/j.jval.2017.03.008.

[3] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016), available at https://www.nature.com/articles/sdata201618, accessed 27/12/2021.

[4] EOSC Executive Board Working Groups FAIR and Architecture. EOSC Interoperability Framework. February 2021, available at https://op.europa.eu/en/publication-detail/-/publication/d787ea54-6a87-11eb-aeb5-01aa75ed71a1/language-en/format-PDF/source-190308283, accessed 28/10/2021

[5] The Secure Anonymised Information Linkage Databank, (Wales) available at https://saildatabank.com/, accessed 27/12/2021.

[6] The Health Data Hub, (France) available at https://www.health-data-hub.fr/, accessed 27/12/2021.

[7] Medical Informatics Initiative (Germany), available at https://www.medizininformatik-initiative.de/, accessed 27/12/2021.

[8] The eClinical Forum and PhRMA EDC/eSource Taskforce. The Future Vision of Electronic Health Records as eSource for Clinical Research. 11/10/2006. Available at https://eclinicalforum.org/Downloads/2006-sept-the-future-vision-of-electronic-health-records-as-esource-for-clinical-research, accessed 27/12/2021.

[9] Green J. How EHR and meaningful use has transformed healthcare. EHR in practice. 11/04/2019, available at https://www.ehrinpractice.com/ehr-meaningful-use-transformed-healthcare.html, accessed 27/12/2021.

[10] Evans R. Electronic Health Records: Then, Now, and in the Future. Yearbook Med Inform. 2016; (Suppl 1): S48–S61. Published online 2016 May 20. doi: 10.15265/IYS-2016-s006.

[11] Booth C and Tannock I. Randomised Controlled Trials and Population-Based Observational Research: Partners in the Evolution of Medical Evidence. *British Journal of Cancer* 110, no. 3 (4 February 2014): 551–55. https://doi.org/10.1038/bjc.2013.725.

[12] Averitt A, Weng C, Ryan P and Perotte A. Translating Evidence into Practice: Eligibility Criteria Fail to Eliminate Clinically Significant Differences between Real-World and Study Populations. *Npj Digital Medicine* 3, no. 1 (11 May 2020): 1–10. https://doi.org/10.1038/s41746-020-0277-8.

[13] Pastorino R, De Vito C, Migliara G et al. Benefits and Challenges of Big Data in Healthcare: An Overview of the European Initiatives. *European Journal of Public Health* 29, no. Supplement_3 (1 October 2019): 23–27. https://doi.org/10.1093/eurpub/ckz168.

[14] Naidoo P, Bouharati C, Rambiritch V et al. Real-World Evidence and Product Development: Opportunities, Challenges and Risk Mitigation. *Wiener Klinische Wochenschrift*, 9 April 2021. https://doi.org/10.1007/s00508-021-01851-w.

[15] Eichler H, Bloechl-Daum B, Broich K et al. Data Rich, Information Poor: Can We Use Electronic Health Records to Create a Learning Healthcare System for Pharmaceuticals? *Clinical Pharmacology & Therapeutics* 105, no. 4 (2019): 912–22. https://doi.org/10.1002/cpt.1226.

[16] Cowie M, Blomster J, Curtis L et al. Electronic Health Records to Facilitate Clinical Research. *Clinical Research in Cardiology: Official Journal of the German Cardiac Society* 106, no. 1 (January 2017): 1–9. https://doi.org/10.1007/s00392-016-1025-6.

[17] Jeon C, Pandol S, Wu B et al. The Association of Statin Use after Cancer Diagnosis with Survival in Pancreatic Cancer Patients: A SEER-Medicare Analysis. *PLoS ONE* 10, no. 4 (1 April 2015): e0121783. https://doi.org/10.1371/journal.pone.0121783.

[18] Vashisht R, Jung K, Schuler A et al. Association of Hemoglobin A1c Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin. *JAMA Network Open* 1, no. 4 (24 August 2018): e181755. https://doi.org/10.1001/jamanetworkopen.2018.1755.

[19] Suchard M, Schuemie M, Krumholz H et al. Comprehensive Comparative Effectiveness and Safety of First-Line Antihypertensive Drug Classes: A Systematic, Multinational, Large-Scale Analysis. *The Lancet* 394, no. 10211 (16 November 2019): 1816–26. https://doi.org/10.1016/S0140-6736(19)32317-7.

[20] Castro V, Clements C, Murphy S et al. QT Interval and Antidepressant Use: A Cross Sectional Study of Electronic Health Records. *BMJ* 346 (29 January 2013): f288. https://doi.org/10.1136/bmj.f288.

[21] Vickers-Smith R, Sun J, Charnigo R et al. Gabapentin Drug Misuse Signals: A Pharmacovigilance Assessment Using the FDA Adverse Event Reporting System'. *Drug and Alcohol Dependence* 206 (1 January 2020): 107709. https://doi.org/10.1016/j.drugalcdep.2019.107709.

[22] Onukwugha E. Visualising data for hypothesis generation using large volume claims data. Methodology.Jan / Feb 2017. Available at ISPOR VOS_Feb 2017_web.indd (https://www.ispor.org/docs/default-source/publications/value-outcomes-spotlight/january-february-2017/vos-visualizing-data.pdf?sfvrsn=58ea1c0c_2); accessed 27/08/2021.

[23] Visweswaran S, Becich M, D'Itri V et al. Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. *JAMIA Open* 1, no. 2 (21 August 2018): 147–52. https://doi.org/10.1093/jamiaopen/ooy033.

[24] Quint, J, Moore E, Lewis A et al. Recruitment of Patients with Chronic Obstructive Pulmonary Disease (COPD) from the Clinical Practice Research Datalink (CPRD) for Research. *Npj Primary Care Respiratory Medicine* 28, no. 1 (December 2018): 21. https://doi.org/10.1038/s41533-018-0089-3.

[25] Gökbuget N, Kelsh M, Chia V et al. Blinatumomab vs historical standard therapy of adult relapsed/ refractory acute lymphoblastic leukemia. : Blood Cancer Journal (2016) 6, e473; doi:10.1038/bcj.2016.84. Available at https://discovery.ucl.ac.uk/id/eprint/1519995/1/bcj201684a.pdf, accessed 27/08/2021.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

[26] Makady A, Ten Ham R., de Boer A et al. Policies for Use of Real-World Data in Health Technology Assessment (HTA): A Comparative Study of Six HTA Agencies. *Value in Health* (2017) *20*(4), 520-532. https://doi.org/10.1016/j.jval.2016.12.003.

[27] Bell H, Wailoo A, Hernandez M et al. The Use of Real World Data for the Estimation of Treatment Effects in NICE Decision Making, 2016, 60. Available at http://nicedsu.org.uk/wp-content/uploads/2018/05/RWD-DSU-REPORT-Updated-DECEMBER-2016.pdf, accessed 27/12/2021

[28] Marquis-Gravel G, Roe M, Robertson H et al. Rationale and Design of the Aspirin Dosing—A Patient-Centric Trial Assessing Benefits and Long-Term Effectiveness (ADAPTABLE) Trial. *JAMA Cardiology* 5, no. 5 (1 May 2020): 598–607. https://doi.org/10.1001/jamacardio.2020.0116.

[29] Davies G, Jordan S, Brooks C et al. Long Term Extension of a Randomised Controlled Trial of Probiotics Using Electronic Health Records. *Scientific Reports* 8, no. 1 (16 2018): 7668. https://doi.org/10.1038/s41598-018-25954-z.

[30] Erlinge D, Koul S, Eriksson P et al. Bivalirudin versus Heparin in Non-ST and ST-Segment Elevation Myocardial Infarction—a Registry-Based Randomized Clinical Trial in the SWEDEHEART Registry (the VALIDATE-SWEDEHEART Trial). *American Heart Journal* 175 (1 May 2016): 36–46. https://doi.org/10.1016/j.ahj.2016.02.007.

[31] Albertson T, Murin S, Sutter M, and Chenoweth J. The Salford Lung Study: A Pioneering Comparative Effectiveness Approach to COPD and Asthma in Clinical Trials. *Pragmatic and Observational Research* 8 (20 September 2017): 175–81. https://doi.org/10.2147/POR.S144157.

[32] Lee S, Monz B, Clemens A et al. Representativeness of the Dabigatran, Apixaban and Rivaroxaban Clinical Trial Populations to Real-World Atrial Fibrillation Patients in the United Kingdom: A Cross-Sectional Analysis Using the General Practice Research Database. *BMJ Open* 2, no. 6 (1 January 2012): e001768. https://doi.org/10.1136/bmjopen-2012-001768.

[33] Targeted Oncology. Palbociclib Approved by FDA for Treatment of Male Patients With HR+/HER2- Breast Cancer (Press release). Available at https://www.targetedonc.com/view/palbociclib-approved-by-fda-for-treatment-of-male-patients-with-hrher2-breast-cancer, accessed 27/08/2021.

[34] FDA. Use of Electronic Health Record Data in Clinical Investigations Guidance for Industry. July 2018. Available at https://www.fda.gov/media/97567/download, accessed 27/12/2021.

[35] FDA. Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics Guidance for Industry. May 2019. Available at https://www.fda.gov/media/124795/download, accessed 27/12/2021.

[36] HMA-EMA Joint Big Data Taskforce – summary report. Available at https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf, accessed 27/08/2021.

[37] Li M, Chen S, Lai Y et al. Integrating Real-World Evidence in the Regulatory Decision-Making Process: A Systematic Analysis of Experiences in the US, EU, and China Using a Logic Model. *Frontiers in Medicine* 8 (31 May 2021): 669509. https://doi.org/10.3389/fmed.2021.669509.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

[38] EMA. EMA Regulatory Science to 2025 - Strategic Reflection, 2020. Available at https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/ema-regulatory-science-2025-strategic-reflection_en.pdf

[39] Lucadou, M, Ganslandt T, Prokosch H, and Toddenroth D. Feasibility Analysis of Conducting Observational Studies with the Electronic Health Record. *BMC Medical Informatics and Decision Making* 19, no. 1 (28 October 2019): 202. https://doi.org/10.1186/s12911-019-0939-0.

[40] Requena G, Wolf A, Williams R et al. Feasibility of Using Clinical Practice Research Datalink Data to Identify Patients with Chronic Obstructive Pulmonary Disease to Enrol into Real-world Trials. *Pharmacoepidemiology and Drug Safety* 30, no. 4 (April 2021): 472–81. https://doi.org/10.1002/pds.5188.

[41] Benchimol E, Smeeth L, Guttmann A et al. The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement. *PLoS Medicine* 12, no. 10 (6 October 2015): e1001885. https://doi.org/10.1371/journal.pmed.1001885.

[42] Callahan A, Shah N, and Chen J. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. *Annals of Internal Medicine* 172, no. 11 Suppl (2 June 2020): S79–84. https://doi.org/10.7326/M19-0873.

[43] Mc Cord, K, Ewald H, Ladanie A et al. Current Use and Costs of Electronic Health Records for Clinical Trial Research: A Descriptive Study. *CMAJ Open* 7, no. 1 (29 January 2019): E23–32. https://doi.org/10.9778/cmajo.20180096.

[44] Swift B, Jain L, White C et al. Innovation at the Intersection of Clinical Trials and Real-World Data Science to Advance Patient Care. *Clinical and Translational Science* 11, no. 5 (September 2018): 450–60. https://doi.org/10.1111/cts.12559.

[45] Zou K Li J, Salem L et al. Harnessing Real-World Evidence to Reduce the Burden of Noncommunicable Disease: Health Information Technology and Innovation to Generate Insights. *Health Services & Outcomes Research Methodology*, 6 November 2020, 1–13. https://doi.org/10.1007/s10742-020-00223-7.

[46] Mc Cord K and Hemkens L. Using Electronic Health Records for Clinical Trials: Where Do We Stand and Where Can We Go? *CMAJ : Canadian Medical Association Journal* 191, no. 5 (4 February 2019): E128–33. https://doi.org/10.1503/cmaj.180841.

[47] Sheikhalishahi S, Miotto R, Dudley J et al. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Medical Informatics* 7, no. 2 (27 April 2019): e12239. https://doi.org/10.2196/12239.

[48] Savova G, Danciu I, Alamudun F et al. Use of Natural Language Processing to Extract Clinical Cancer Phenotypes from Electronic Medical Records. *Cancer Research* 79, no. 21 (1 November 2019): 5463–70. https://doi.org/10.1158/0008-5472.CAN-19-0579.

[49] Wu R, Datta S et al. Deep Learning in Clinical Natural Language Processing: A Methodical Review. *Journal of the American Medical Informatics Association: JAMIA* 27, no. 3 (1 March 2020): 457–70. https://doi.org/10.1093/jamia/ocz200.

[50] Personal communication with UK NHS records staff. July 2021.

[51] NHS Digital – Summary Care Records. Available at Summary Care Records (SCR) - NHS Digital (https://digital.nhs.uk/services/summary-care-records-scr); accessed 27/08/2021.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

[52] Bian J, Lyu T, Loiacono A et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc. 2020 Dec; 27(12): 1999–2010, available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7727392/; accessed 28/08/2021.

[53] Kohane I, Aranow B, Avillach P et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res 2021 | vol. 23 | iss. 3 | e22219 | p. 1. Available at https://pubmed.ncbi.nlm.nih.gov/33600347/ ; accessed 28/08/2021.

[54] Mehra M, Desai S, Ruschitzka F et al. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. The Lancet 2020 May doi: 10.1016/S0140-6736(20)31180-6. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7255293/ ; accessed 28/08/2021.

[55] Mehra M, Desai S, Kuy S et al. RETRACTED: Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med 2020 Jun 18;382(25):e102. doi: 10.1056/nejmoa2007621. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7206931/ ; accessed 28/08/2021.

[56] Data taken from the database behind the ECRIN Metadata Repository (https://crmdr.org). Total number of studies registered in one or more trial registries at 08/12/2020 were 617,351, at 05/12/2021 the figure was 667,346.

[57] NIHR: Health technology Assessment, available at Health Technology Assessment | NIHR (https://www.nihr.ac.uk/explore-nihr/funding-programmes/health-technology-assessment.htm); accessed 29/08/2021.

[58] ICHOM, the International Consortium for Health Outcomes Measurement, available at https://www.ichom.org/, accessed 29/08/2021.

[59] LoCasale R, Pashos C, Gutierrez B et al. Bridging the Gap Between RCTs and RWE Through Endpoint Selection. *Therapeutic Innovation & Regulatory Science* 55, no. 1 (2021): 90–96. https://doi.org/10.1007/s43441-020-00193-5.

[60] McGrath, D. French Centres Adopt ICHOM Standards. *EuroTimes* (blog), 1 September 2019. Available at https://www.eurotimes.org/french-centres-adopt-ichom-standards/, accessed 11/12/2021

[61] Evans S, Millar J, Moore C et al. Cohort Profile: The TrueNTH Global Registry - an International Registry to Monitor and Improve Localised Prostate Cancer Health Outcomes. *BMJ Open* 7, no. 11 (1 November 2017): e017006. https://doi.org/10.1136/bmjopen-2017-017006.

[62] Bak J, Serné E, Kramer M et al. National Diabetes Registries: Do They Make a Difference? *Acta Diabetologica* 58, no. 3 (1 March 2021): 267–78. https://doi.org/10.1007/s00592-020-01576-8.

[63] EMA – Patient Registries, available at https://www.ema.europa.eu/en/human-regulatory/post-authorisation/patient-registries, accessed 27/12/2021.

[64] McGettigan P, Olmo A, Plueschke K et al. Patient Registries: An Underused Resource for Medicines Evaluation. Drug Saf 42, 1343–1351 (2019). https://doi.org/10.1007/s40264-019-00848-9.

[65] RD Connect: FAIRification of rare disease registries, available at https://rd-connect.eu/what-we-do/data-linkage/fairification/, accessed 27/12/2021.

[66] Tan A, Armstrong E, Close J, and Harris I. Data Quality Audit of a Clinical Quality Registry: A Generic Framework and Case Study of the Australian and New Zealand Hip Fracture Registry. *BMJ Open Quality* 8, no. 3 (July 2019): e000490. https://doi.org/10.1136/bmjoq-2018-000490.

[67] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. J Clin Epidemiol. 2008 Apr; 61(4):344-9. PMID: 18313558

[68] WHO. WHO Recommended Surveillance Standards. Second Edition, 1999. Available at https://www.who.int/publications/i/item/who-recommended-surveillance-standards, assessed 28/10/2021.

[69] Fairchild G, Tasseff B, Khalsa H et al. Epidemiological Data Challenges: Planning for a More Robust Future Through Data Standards. *Frontiers in Public Health* 6 (2018): 336. https://doi.org/10.3389/fpubh.2018.00336

[70] Badker R, Miller K, Pardee C, et al. Challenges in reported COVID-19 data: best practices and recommendations for future epidemics. BMJ Glob Health. 2021;6(5):e005542. doi:10.1136/bmjgh-2021-005542.

[71] International Severe Acute Respiratory and Emerging Infection Consortium, at https://isaric.org/, accessed 27/12/2021.

[72] Dunning J, Merson L, Rohde G et al. Open Source Clinical Science for Emerging Infections. *The Lancet Infectious Diseases* 14, no. 1 (1 January 2014): 8–9. https://doi.org/10.1016/S1473-3099(13)70327-X.

[73] ISARIC. ISARIC/WHO Clinical Characterisation Protocol for Severe Emerging Infections, 2020. Available at https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fisaric.org%2Fwp-content%2Fuploads%2F2020%2F10%2FISARIC_CCP.docx&wdOrigin=BROWSELINK, accessed 28/10/2021.

[74] Infectious Disease Data Observatory, at https://www.iddo.org/, accessed 27/12/2021.

[75] EMA. ICH E2B (R3) Electronic Transmission of Individual Case Safety Reports (ICSRs) - Data Elements and Message Specification Implementation Guide. European Medicines Agency, 17 September 2018. https://www.ema.europa.eu/en/ich-e2b-r3-electronic-transmission-individual-case-safety-reports-icsrs-data-elements-message.

[76] Centre for public impact - The Electronic Health Records System in the UK, available at https://www.centreforpublicimpact.org/case-study/electronic-health-records-system-uk/, accessed 27/12/2021.

[77] Adler-Milstein J, Holmgren A, Kralovec et al. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc. 2017 Nov 1;24(6):1142-1148. doi: 10.1093/jamia/ocx080. Available at https://pubmed.ncbi.nlm.nih.gov/29016973/; accessed 29/08/2021.

[78] Apathy N, Holmgren A, Adler-Milstein J. A decade post-HITECH: Critical access hospitals have electronic health records but struggle to keep up with other advanced functions. J Am Med Inform Assoc. 2021 Jul 1;28(9):1947–54. https://doi.org/10.1093/jamia/ocab102 Epub ahead of print.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

[79] USF Health - The US Core Data for Interoperability: Purpose, Classes, Growth and Importance. 30/11/2020. Available at https://www.usfhealthonline.com/resources/health-informatics/the-us-core-data-for-interoperability-purpose-classes-and-growth-and-importance/, accessed 27/12/2021.

[80] HealthIT.gov - United States Core Data for Interoperability (USCDI), available at https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi, accessed 27/12/2021.

[81] The U.S. Core Data for Interoperability: A National Policy Journey. Center for Digital Health Innovation at UCSF. 13/05/2021. Available at https://www.centerfordigitalhealthinnovation.org/posts/the-us-core-data-for-interoperability-a-national-policy-journey, accessed 29/08/2021.

[82] What is OpenEHR, available at https://www.openehr.org/about/what_is_openehr, accessed 28/10/2021.

[83] Purkayastha, S, Allam R, Pallavi M and Gichoya J. Comparison of Open-Source Electronic Health Record Systems Based on Functional and User Performance Criteria. *Healthcare Informatics Research* 25, no. 2 (April 2019): 89–98. https://doi.org/10.4258/hir.2019.25.2.89. Accessed 28/10/2021.

[84] Meredith J. What is OpenEHR and why is it important?. Digital Health.Wales. 2nd February 2021. Available at https://digitalhealth.wales/news/what-openehr-and-why-it-important, accessed 28/10/2021.

[85] i2b2 – Informatics for integrating biology and the bedside, at https://www.i2b2.org/, accessed 29/08/2021.

[86] i2b2 Web Client Architecture Guide, at https://community.i2b2.org/wiki/display/webclient/Web+Client+Architecture+Guide, accessed 29/08/2021.

[87] Entity – Attribute – Value Model. Wikipedia. At Entity–attribute–value model - Wikipedia https://en.wikipedia.org/wiki/Entity%E2%80%93attribute%E2%80%93value_model, accessed 28/10/2021.

[88] Visweswaran S, Becich M, D'Itri V et al. Accrual to Clinical Trials (ACT): A Clinical and Translational Science Award Consortium Network. JAMIA Open. 2018 Oct;1(2):147-152. https://doi.org/10.1093/jamiaopen/ooy033 Epub 2018 Aug 21.

[89] Morrato E, Lennox L, Sendro E et al. Scale-up of the Accrual to Clinical Trials (ACT) network across the Clinical and Translational Science Award Consortium: a mixed-methods evaluation of the first 18 months. J Clin Transl Sci. 2020 Jun 30;4(6):515-528. https://doi.org/10.1017/cts.2020.505.

[90] OMOP Common Data Model, at https://www.ohdsi.org/data-standardization/the-common-data-model/, accessed 27/12/2021.

[91] OHDSI – The Observational Health Data Sciences and Informatics programme, at https://www.ohdsi.org/, accessed 27/12/2021.

[92] OHDSI Software tools, available at https://www.ohdsi.org/software-tools/, accessed 27/12/2021.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

[93] EHDEN Data Partner Calls, available at https://www.ehden.eu/data-partner-calls/, accessed 27/12/2021.

[94] EHDEN almost doubles SME network in 19 countries following certification of third SME cohort. Press release 02/10/2020, available at https://www.ehden.eu/ehden-almost-doubles-sme-network, accessed 28/10/2021.

[95] Mapping UK Biobank to the OMOP CDM using the flexible ETL framework Delphyne. The Hyve. 19th August 2021. Available at https://www.thehyve.nl/cases/mapping-uk-biobank-to-omop-using-delphyne, accessed 28/10/2021.

[96] The Book of OHDSI – Chapter 4: The Common Data Model, at https://ohdsi.github.io/TheBookOfOhdsi/CommonDataModel.html, accessed 27/12/2021.

[97] HL7 FHIR – Welcome to FHIR, available at http://hl7.org/fhir/index.html, accessed 27/12/2021.

[98] HL7 FHIR – Resource Index, available at https://www.hl7.org/fhir/resourcelist.html, accessed 27/12/2021.

[98] HL7 UK INTEROPen CareConnect FHIR profiles, available at https://fhir.hl7.org.uk/, accessed 27/12/2021.

[100] The Book of OHDSI – Chapter 5: Standardized Vocabularies, available at https://ohdsi.github.io/TheBookOfOhdsi/StandardizedVocabularies.html; accessed 29/08/2021.

[101] SNOMED International, at https://www.snomed.org/, accessed 27/12/2021.

[102] The Global Patient Set, at https://gps.snomed.org/. accessed 27/12/2021.

[103] SNOMED – Our Partners, at https://www.snomed.org/snomed-international/our-partners, accessed 27/12/2021.

[104] ICD – The International Classification of Diseases, at (ICD 11) https://www.who.int/standards/classifications/classification-of-diseases, accessed 27/12/2021.

[105] Sylvestre E, Bouzillé G, McDuffie M et al. A Semi-Automated Approach for Multilingual Terminology Matching: Mapping the French Version of the ICD-10 to the ICD-10 CM, 2020, 5; European Federation for Medical Informatics, available to download at https://ebooks.iospress.nl/publication/54116, accessed 27/12/2021.

[106] RxNorm, at https://www.nlm.nih.gov/research/umls/rxnorm/index.html, accessed 27/12/2021.

[107] Belenkaya R, Gurley M, Golozar A et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. *JCO Clinical Cancer Informatics*, no. 5 (1 October 2021): 12–20. https://doi.org/10.1200/CCI.20.00079.

[108] Michael C, Sholle E, Wulff R et al. Mapping Local Biospecimen Records to the OMOP Common Data Model. AMIA Summits on Translational Science Proceedings 2020 (30 May 2020): 422–29.

[109] Kim J, Kim S, Ryu B et al. Transforming Electronic Health Record Polysomnographic Data into the Observational Medical Outcome Partnership's Common Data Model: A Pilot Feasibility Study. *Scientific Reports* 11, no. 1 (29 March 2021): 7013. https://doi.org/10.1038/s41598-021-86564-w.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

[110] Lamer A, Depas N, Doutreligne M et al. Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. *Applied Clinical Informatics* 11, no. 1 (January 2020): 13–22. https://doi.org/10.1055/s-0039-3402754.

[111] Maier C., Lang L, Storf H et al. Towards Implementation of OMOP in a German University Hospital Consortium. *Applied Clinical Informatics* 9, no. 1 (January 2018): 54–61. https://doi.org/10.1055/s-0037-1617452.

[112] Klann J., Joss M, Embree K and Murphy S. Data Model Harmonization for the All Of Us Research Program: Transforming I2b2 Data into the OMOP Common Data Model. *PLoS ONE* 14, no. 2 (19 February 2019). https://doi.org/10.1371/journal.pone.0212463.

[113] Kim, H, Choi J, Jang I et al. Feasibility of Representing Data from Published Nursing Research Using the OMOP Common Data Model. AMIA Annual Symposium Proceedings 2016 (10 February 2017): 715–23.

[114] The Comet Initiative - Core outcome measures in effectiveness trials. Available at https://comet-initiative.org/, accessed 27/12/2021.

[115] Gargon E, Gorst SL, Matvienko-Sikar K, Williamson PR (2021) Choosing important health outcomes for comparative effectiveness research: 6th annual update to a systematic review of core outcome sets for research. PLOS ONE 16(1): e0244878 https://doi.org/10.1371/journal.pone.0244878.

[116] Global Regulatory Requirements, available at https://www.cdisc.org/resources/global-regulatory-requirements, accessed 27/12/2021.

[117] CDISC – Standards, available at https://www.cdisc.org/standards, accessed 27/12/2021.

[118] European Clinical Research Alliance on Infectious Diseases (ECRAID), at https://www.ecraid.eu/, accessed 27/12/2021.

[119] VACCELERATE, at https://vaccelerate.eu/, accessed 27/12/2021.

[120] CDISC – Study Data Tabulation Model, Version 1.8 (Final), available at https://www.cdisc.org/system/files/members/standard/foundational/sdtm/SDTM%20v1.8.pdf, accessed 27/12/2021 *[accessible to CDISC members only]*

[121] CDISC – Study Data Tabulation Model Implementation Guide: Human Clinical Trials, Version 3.3 (Final), available at https://www.cdisc.org/standards/foundational/sdtmig/sdtmig-v3-3/html, accessed 27/12/2021 *[accessible to CDISC members only]*

[122] SDTM Draft Domains Home, on the CDISC wiki, available at https://wiki.cdisc.org/display/SDD/SDTM+Draft+Domains+Home, accessed 29/10/2021

[123] CDISC – Clinical Data Acquisition Standards Harmonization Model, Version 1.1 (Final), available at https://www.cdisc.org/standards/foundational/cdash/cdash-model-v1-1-0, accessed 27/12/2021 *[accessible to CDISC members only]*

[124] CDASH Implementation Guide Version 2.1 (Final), available at https://www.cdisc.org/system/files/members/standard/foundational/CDASHIG%20v2.1-Final_Rev.pdf, accessed 27/12/2021 *[accessible to CDISC members only]*

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

[125] CDISC Controlled Terminology, available at https://evs.nci.nih.gov/ftp1/CDISC/SDTM/SDTM%20Terminology.pdf, accessed 27/12/2021.

[126] CDISC SDTM Controlled Terminology, 2021-06-25. Available at https://evs.nci.nih.gov/ftp1/CDISC/SDTM/CDASH%20Terminology.html, accessed 30/08/2021.

[127] MedDRA – Medical Dictionary for Regulatory Activities, at https://www.meddra.org/, accessed 30/08/2021.

[128] WHODrug Global, at https://who-umc.org/whodrug/whodrug-global/, accessed 30/08/2021.

[129] LOINC - The international standard for identifying health measurements, observations, and documents, at https://loinc.org/, accessed 30/08/2021.

[130] Regenstrief Institute, at https://www.regenstrief.org/, accessed 27/12/2021.

[131] LOINC Working Group: (FDA, NIH, CDISC, and Regenstrief Institute). Recommendations for the Submission of LOINC® Codes in Regulatory Applications to the U.S. Food and Drug Administration. November 2017, available at https://www.fda.gov/media/109376/download, accessed 27/12/2021.

[132] CDISC Define-XML Specification Version 2.1 (Final), available at https://www.cdisc.org/standards/foundational/define-xml/define-xml-v2-1, accessed 29/10/2021.

[133] Aerts, J. Working on and with CDISC Standards: LOINC-SDTM Mapping for Drug and Toxicology Lab Test. *Working on and with CDISC Standards* (blog), 1 March 2021. Available at https://cdiscguru.blogspot.com/2021/03/loinc-sdtm-mapping-for-drug-and.html, accessed 29/08/2021.

[134] van Bochove, K. EHDEN - D4.5 Roadmap for interoperability solutions. 16/12/2020. Available at https://zenodo.org/record/4474373, accessed 28/10/2021.

[135] OHDSI - HL7 International and OHDSI Announce Collaboration to Provide Single Common Data Model for Sharing Information in Clinical Care and Observational Research (Press release), available at https://www.ohdsi.org/ohdsi-hl7-collaboration/, accessed 30/08/2021.

[136] Inaugural HL7-OHDSI meeting on FHIR and OMOP - 4-August 2021. Available at https://confluence.hl7.org/display/OOF/Inaugural+HL7-OHDSI+meeting+on+FHIR+and+OMOP+-+4-August+2021, accessed 28/10/2021.

[137] Real World data (and CDISC), Available at https://www.cdisc.org/standards/real-world-data, accessed 11/12/2021.

[138] Facile R and Wurst B. CDISC RWD Connect Report, 2020. Available at https://www.cdisc.org/sites/default/files/2021-08/CDISC_RWD_Connect_Report_07-27-2021.pdf, accessed 11/12/2021.

[139] BioPortal, available at https://bioportal.bioontology.org/, accessed 27/12/2021.

[140] FAIRSharing.org – standards, databases, policies, available at https://fairsharing.org/, accessed 27/12/2021.

[141] XKCD – Standards, available at https://xkcd.com/927/, accessed 27/12/2021.

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

[142] Joint Initiative Council, at Welcome to Joint Initiative Council http://jointinitiativecouncil.org/; accessed 30/08/2021.

Abbreviations

CCP	Clinical Characterisation Protocol
CDASH	Data Acquisition and Standards Harmonisation
CDM	Common Data Model
CDISC	Clinical Data Interchange Standards Consortium (originally, now CDISC)
COMET	Core Outcome Measures in Effectiveness Trials
COS	Core Outcome Set
CRO	Contract Research Organisation
CT	Controlled Terminology
DICOM	Digital Imaging and Communication in Medicine
EAV	Entity, Attribute, Value
eCRFs	Electronic Case Report Forms
EHDEN	European Health Data and Evidence
EHR	Electronic Health Record
EMA	European Medicines Agency
eRDC	Electronic Remote Data Capture
ETL	Extract, Transform and Load
EVS	Enterprise Vocabulary Services
FAIR	Findable, Accessible, Interoperable, Re-usable
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
HL7	Health level 7
HTA	Health Technology Assessment
HITECH	Health Information Technology for Economic and Clinical Health Act
i2b2	Informatics for integrating biology at the bedside
ICD	International Statistical Classification of Diseases and Related Health Problems
ICHOM	International Consortium for Health Outcomes Measurement
IDDO	Infectious Disease Data Observatory
ISARIC	International Severe Acute Respiratory and emerging Infections Consortium

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

LOINC	Logical Observation Identifiers Names and Codes system
MedDRA	Medical Dictionary of Adverse Reactions
NCI	National Cancer Institute (US)
NLP	Natural Language Processing
NMPA	National Medical Products Administration (China)
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PAES	Post Authorisation Efficacy Study
PASS	Post-authorisation safety studies
PCT	Pragmatic clinical trial
PCORNet	National Patient Centred Clinical Research Network (US)
PMDA	Pharmaceuticals and Medical Devices Agency (Japan)
RCT	Randomised Controlled Trial
RWD	Real World Data
RWE	Real World Evidence
SAS	Statistical Analysis System (originally, now SAS)
SDTM	Standard Data Tabulation Model
SNOMED	Systematic Nomenclature for Medicine
WHO	World health organisation

Delivery and Schedule

D4.4 “Data standards for observational and interventional studies, and interoperability between healthcare and research data.” was due for M30 (August 2021) as rescheduled with the 2nd Grant Agreement Amendment.

The delivery has slightly deviated from this deadline due to additional COVID-19 related workload for the co-authors and difficulties that the partners encountered in recruiting additional human resources during the pandemic. As a result, the actual delivery date is M34 (December 2021).

Adjustments

Adjustments made:

None

This project has received funding from the *European Union’s Horizon 2020 research and innovation programme* under grant agreement No 824087.

Appendices

Appendix 1. Domains within SDTM and CDASH (from SDTMIG v3.3)

Dataset	Description	Class	Structure
CO	Comments	Special Purpose	One record per comment per subject
DM	Demographics	Special Purpose	One record per subject
SE	Subject Elements	Special Purpose	One record per actual Element per subject
SM	Subject Disease Milestones	Special Purpose	One record per Disease Milestone per subject
SV	Subject Visits	Special Purpose	One record per subject per actual visit
AG	Procedure Agents	Interventions	One record per recorded intervention occurrence per subject
CM	Concomitant/Prior Medications	Interventions	One record per recorded intervention occurrence or constant-dosing interval per subject
EC	Exposure as Collected	Interventions	One record per protocol-specified study treatment, collected-dosing interval, per subject, per mood
EX	Exposure	Interventions	One record per protocol-specified study treatment, constant-dosing interval, per subject
ML	Meal Data	Interventions	One record per food product occurrence or constant intake interval per subject
PR	Procedures	Interventions	One record per recorded procedure per occurrence per subject
SU	Substance Use	Interventions	One record per substance type per reported occurrence per subject
AE	Adverse Events	Events	One record per adverse event per subject
CE	Clinical Events	Events	One record per event per subject
DS	Disposition	Events	One record per disposition status or protocol milestone per subject
DV	Protocol Deviations	Events	One record per protocol deviation per subject
HO	Healthcare Encounters	Events	One record per healthcare encounter per subject
MH	Medical History	Events	One record per medical history event per subject
CV	Cardiovascular System Findings	Findings	One record per finding or result per time point per visit per subject
DA	Drug Accountability	Findings	One record per drug accountability finding per subject
DD	Death Details	Findings	One record per finding per subject
EG	ECG Test Results	Findings	One record per ECG observation per replicate per time point or one record per ECG observation per beat per visit per subject
FA	Findings About Events or	Findings	One record per finding, per object, per time point, per visit per subject

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Dataset	Description	Class	Structure
	Interventions		
FT	Functional Tests	Findings	One record per Functional Test finding per time point per visit per subject
IE	Inclusion/Exclusion Criteria Not Met	Findings	One record per inclusion/exclusion criterion not met per subject
IS	Immunogenicity Specimen Assessments	Findings	One record per test per visit per subject
LB	Laboratory Test Results	Findings	One record per lab test per time point per visit per subject
MB	Microbiology Specimen	Findings	One record per microbiology specimen finding per time point per visit per subject
MI	Microscopic Findings	Findings	One record per finding per specimen per subject
MK	Musculoskeletal System Findings	Findings	One record per assessment per visit per subject
MO	Morphology	Findings	One record per Morphology finding per location per time point per visit per subject
MS	Microbiology Susceptibility	Findings	One record per microbiology susceptibility test (or other organism-related finding) per organism found in MB
NV	Nervous System Findings	Findings	One record per finding per location per time point per visit per subject
OE	Ophthalmic Examinations	Findings	One record per ophthalmic finding per method per location, per time point per visit per subject
PC	Pharmacokinetics Concentrations	Findings	One record per sample characteristic or time-point concentration per reference time point or per analyte per subject
PE	Physical Examination	Findings	One record per body system or abnormality per visit per subject
PP	Pharmacokinetics Parameters	Findings	One record per PK parameter per time-concentration profile per modelling method per subject
QS	Questionnaires	Findings	One record per questionnaire per question per time point per visit per subject
RE	Respiratory System Findings	Findings	One record per finding or result per time point per visit per subject
RP	Reproductive System Findings	Findings	One record per finding or result per time point per visit per subject
RS	Disease Response and Clin Classification	Findings	One record per response assessment or clinical classification assessment per time point per visit per subject per assessor per medical evaluator
SC	Subject Characteristics	Findings	One record per characteristic per subject.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Dataset	Description	Class	Structure
SR	Skin Response	Findings	One record per finding, per object, per time point, per visit per subject
SS	Subject Status	Findings	One record per finding per visit per subject
TR	Tumour/Lesion Results	Findings	One record per tumour measurement/assessment per visit per subject per assessor
TU	Tumour/Lesion Identification	Findings	One record per identified tumour per subject per assessor
UR	Urinary System Findings	Findings	One record per finding per location per visit per subject
VS	Vital Signs	Findings	One record per vital sign measurement per time point per visit per subject
TA	Trial Arms	Trial Design	One record per planned Element per Arm
TD	Trial Disease Assessments	Trial Design	One record per planned constant assessment period
TE	Trial Elements	Trial Design	One record per planned Element
TI	Trial Inclusion/Exclusion Criteria	Trial Design	One record per I/E criterion
TM	Trial Disease Milestones	Trial Design	One record per Disease Milestone type
TS	Trial Summary Information	Trial Design	One record per trial summary parameter value
TV	Trial Visits	Trial Design	One record per planned Visit per Arm
RELREC	Related Records	Relationships	One record per related record, group of records or dataset
RELSUB	Related Subjects	Relationships	One record per relationship per related subject per subject
SUPP--	Supplemental Qualifiers for [domain name]	Relationships	One record per IDVAR, IDVARVAL, and QNAM value per subject
OI	Non-host Organism Identifiers	Study Reference	One record per taxon per non-host organism

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Appendix 2. CDASH Variables, for Findings domains (from CDASH v1.1)

CDASH Var.	CDASH Variable Label	DRAFT CDASH Definition
--OBJ	Object of the Observation	Describes the event or intervention whose property is being measured in -- TESTCD/--TEST.
--YN	Any [Finding]	An indication whether or not any data was collected for the finding topic.
--PERF	[Observation] Performed	An indication of whether or not a planned measurement, series of measurements, test, observation or specimen was performed or collected.
-- TESTCD	Short Name of Measurement, Test or Examination	Short character value code for the test being performed.
--TEST	Name of Measurement, Test or Examination	Descriptive name for the test being performed. Examples: Platelet, Systolic Blood Pressure, Summary (Min) RR Duration, Eye Examination.
-- TSTDTL	Measurement, Test or Examination Detail	A further description of --TESTCD and -- TEST.
--CAT	Category	A grouping of topic- variable values based on user-defined characteristics.
--SCAT	Subcategory	A sub-division of the -- CAT values based on user-defined characteristics.
--ORRES	Result or Finding in Original Units	Result of the measurement or finding as originally received or collected.
-- ORRESU	Original Units	The unit of the result as originally received or collected.
--CRESU	Collected Non- Standard Unit	The unit of the result if it were collected as a non-standard unit.
--DESC	Description of Finding	Text description of any findings.
--RES	Collected Result or Finding	The result of the measurement or finding as originally received or collected.
-- RESOTH	Result Other	A free text result which provides further information about the original received or collected result.
-- RESCAT	Result Category	A categorization of the result of a finding.
-- ORNRLO	Normal Range Lower Limit- Original Units	The lower end of normal range or reference range for continuous results stored in --ORRES.
-- ORNRHI	Normal Range Upper Limit- Original Units	The upper end of normal range or reference range for continuous results stored in --ORRES.
-- CSTNRC	Collected Character/Ordinal Normal Range	The normal references ranges that are expressed as characters ("Negative to Trace") or ordinal (-1 to 1).

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

CDASH Var.	CDASH Variable Label	DRAFT CDASH Definition
--NRIND	Normal/Reference Range Indicator	An indication or description about how the value compares to the normal range or reference range.
--STAT	Completion Status	The variable used to indicate that data are not available by having the site recording the value as "Not Done".
-- REASND	Reason Not Done	An explanation of why the data are not available.
--NAM	Laboratory/Vendor Name	Name or identifier of the vendor (e.g., laboratory) that provided the test results.
--LOINC	LOINC Code	The Logical Observation Identifiers Names and Codes (LOINC) code for the topic variable such as a lab test.
--SPEC	Specimen Material Type	The type of specimen used for a measurement.
-- ANTREG	Anatomical Region	The specific anatomical or biological region of a tissue, organ specimen or the region from which the specimen is obtained, as defined in the protocol, such as a section or part of what is described in the -- SPEC variable.
-- SPCCND	Specimen Condition	The condition of the specimen.
-- CSPUFL	Collected Specimen Usability Flag	An indication about the usability of the specimen for obtaining the test result.
--POS	Position of Subject During Observation	The position of the subject during a measurement or examination.
--LOC	Location Used for the Measurement	The anatomical location of the subject relevant to the collection of the measurement.
--LAT	Laterality	Qualifier for anatomical location further detailing the side of the body.
--DIR	Directionality	Qualifier further detailing the position of the anatomical location relative to the center of the body, organ, or specimen.
-- LOCDTL	Location Detail	A detail description of the location of the identified finding.
-- PORTOT	Portion or Totality	Qualifier for anatomical location further detailing the distribution, which means arrangement of, apportioning of.
-- METHOD	Method of Test or Examination	The method of the test or examination.
--LEAD	Lead Identified to Collect Measurements	The lead or leads identified to capture the measurement for a test from an instrument.
-- CSTATE	Consciousness State	The consciousness state of the subject at the time of measurement.
--FAST	Fasting Status	An indication that the subject has abstained from food/water for the specified amount of time.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

CDASH Var.	CDASH Variable Label	DRAFT CDASH Definition
--EVAL	Evaluator	The role of the person who provided the evaluation.
--EVALID	Evaluator Identifier	An identifier used to distinguish multiple evaluators with the same role recorded in - -EVAL.
-- ACPTFL	Accepted Record Flag	An indication that the evaluation is considered, by an independent assessor, to be the accepted or final evaluation.
--TOX	Toxicity	The description of toxicity quantified by -- TOXGR such as NCI CTCAE Short Name.
--TOXGR	Toxicity Grade	The toxicity grade using a standard toxicity scale (such as the NCI CTCAE).
--SEV	Severity	The severity or intensity of a particular finding.
-- DTHREL	Relationship to Death	An indication of the relationship of a particular finding to the death of a subject.
--CLLOQ	Collected Lower Limit of Quantitation	The collected lower limit of quantitation for an assay, represented in text format or as a range, such as less than a specified numeric value.
--CULOQ	Collected Upper Limit of Quantitation	The collected upper limit of quantitation for an assay, represented in text format or as a range, such as greater than a specified numeric value.
--COND	Test Condition Met	An indication whether the testing conditions defined in the protocol were met (e.g., Low fat diet).
--CLSIG	Clinical Significance	An indication whether the test results were clinically significant.
-- REPNUM	Repetition Number	The instance number of a test that is repeated within a given timeframe for the same test. The level of granularity can vary, e.g., within a time point or within a visit.
--DATFL	Same as Previous Sample Collection Date	A flag indicating that the date (or start date) is the same as the previous specimen collection date (or start date).
-- ENDATF	Same as Current Sample Collection Start Date	A flag indicating that the specimen/sample collection ended on the same date as the current/previous specimen collection started.
COVAL	Comment	A free text comment.
-- MODIFY	Modified Term	If the value for -- ORRES is modified for coding purposes, then the modified text is placed here.
-- BODSYS	Body System or Organ Class	Body System or Organ Class that is involved for a finding from the standard hierarchy for dictionary-coded results.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

Appendix 3. Observation Table Field definitions – OMOP CDM 6.0

CDM Field	User Guide	ETL Conventions	Type	Key
observation_id **(**required)	The unique key given to an Observation record for a Person. Refer to the ETL for how duplicate Observations during the same Visit were handled.	Each instance of an observation present in the source data should be assigned this unique key.	bigint	PK
person_id (required)	The PERSON_ID of the Person for whom the Observation is recorded. This may be a system generated code.		bigint	FK
observation_ concept_id (required)	The OBSERVATION_CONCEPT_ID field is recommended for primary use in analyses, and must be used for network studies.	The CONCEPT_ID that the OBSERVATION_SOURCE_CONCEPT_I D maps to. There is no specified domain that the Concepts in this table must adhere to. The only rule is that records with Concepts in the Condition, Procedure, Drug, Measurement, or Device domains MUST go to the corresponding table.	int	FK
observation_ date	The date of the Observation. Depending on what the Observation represents this could be the date of a lab test, the date of a survey, or the date a patient's family history was taken.	For some observations the ETL may need to make a choice as to which date to choose.	date	No
observation_da tetime (required)		If no time is given set to midnight (00:00:00).	date-time	No
observation_ type_ concept_id (required)	This field can be used to determine the provenance of the Observation record, as in whether the measurement was from an EHR system, insurance claim, registry, or other sources.	Choose the OBSERVATION_TYPE_CONCEPT_ID that best represents the provenance of the record, for example whether it came from an EHR record or billing claim. Accepted Concepts.	int	FK
value_as_ number	This is the numerical value of the Result of the Observation, if applicable and available. It is not expected that all Observations will have numeric results, rather, this field is here to house values should they exist.		float	No

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

CDM Field	User Guide	ETL Conventions	Type	Key
value_as_string	This is the categorical value of the Result of the Observation, if applicable and available.		var char (60)	No
value_as_ concept_id	It is possible that some records destined for the Observation table have two clinical ideas represented in one source code. This is common with ICD10 codes that describe a family history of some Condition, for example. In OMOP the Vocabulary breaks these two clinical ideas into two codes; one becomes the OBSERVATION_CONCEPT_ID and the other becomes the VALUE_AS_CONCEPT_ID. It is important when using the Observation table to keep this possibility in mind and to examine the VALUE_AS_CONCEPT_ID field for relevant information.	Note that the value of VALUE_AS_CONCEPT_ID may be provided through mapping from a source Concept which contains the content of the Observation. In those situations, the CONCEPT_RELATIONSHIP table in addition to the 'Maps to' record contains a second record with the relationship_id set to 'Maps to value'. For example, ICD10 Z82.4 'Family history of ischaemic heart disease and other diseases of the circulatory system' has a 'Maps to' relationship to 4167217 'Family history of clinical finding' as well as a 'Maps to value' record to 134057 'Disorder of cardiovascular system'.	Int	FK
qualifier_ concept_id	This field contains all attributes specifying the clinical fact further, such as as degrees, severities, drug-drug interaction alerts etc.	Use your best judgement as to what Concepts to use here and if they are necessary to accurately represent the clinical record. There is no restriction on the domain of these Concepts, they just need to be Standard.	int	FK
unit_concept_id	There is currently no recommended unit for individual observation concepts. UNIT_SOURCE_VALUES should be mapped to a Standard Concept in the Unit domain that best represents the unit as given in the source data.	There is no standardization requirement for units associated with OBSERVATION_CONCEPT_IDs, however, it is the responsibility of the ETL to choose the most plausible unit.	int	FK

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

CDM Field	User Guide	ETL Conventions	Type	Key
provider_id	The provider associated with the observation record, e.g. the provider who ordered the test or the provider who recorded the result.	The ETL may need to make a choice as to which PROVIDER_ID to put here. Based on what is available this may or may not be different than the provider associated with the overall VISIT_OCCURRENCE record. For example the admitting vs attending physician on an EHR record.	bigint	FK
visit_ occurrence_id	The visit during which the Observation occurred.	Depending on the structure of the source data, this may have to be determined based on dates. If an OBSERVATION_DATE occurs within the start and end date of a Visit it is a valid ETL choice to choose the VISIT_OCCURRENCE_ID from the visit that subsumes it, even if not explicitly stated in the data. While not required, an attempt should be made to locate the VISIT_OCCURRENCE_ID of the observation record. If an observation is related to a visit explicitly in the source data, it is possible that the result date of the Observation falls outside of the bounds of the Visit dates.	bigint	FK
visit_detail_id	The VISIT_DETAIL record during which the Observation occurred. For example, if the Person was in the ICU at the time the VISIT_OCCURRENCE record would reflect the overall hospital stay and the VISIT_DETAIL record would reflect the ICU stay during the hospital visit.	Same rules apply as for the VISIT_OCCURRENCE_ID.	bigint	FK
observation_ source_value	This field houses the verbatim value from the source data representing the Observation that occurred. For example, this could be an ICD10 or Read code.	This code is mapped to a Standard Concept in the Standardized Vocabularies and the original code is stored here for reference.	var char (50)	No
observation_ source_ concept_id (required)	This is the concept representing the OBSERVATION_SOURCE_VALUE and may not necessarily be standard. This field is discouraged from use in analysis because it is not required to contain Standard	If the OBSERVATION_SOURCE_VALUE is coded in the source data using an OMOP supported vocabulary put the concept id representing the source	int	FK

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.

CDM Field	User Guide	ETL Conventions	Type	Key
	Concepts that are used across the OHDSI community, and should only be used when Standard Concepts do not adequately represent the source detail for the Observation necessary for a given analytic use case. Consider using OBSERVATION_CONCEPT_ID instead to enable standardized analytics that can be consistent across the network.	value here. If not available, set to 0.		
unit_ source_value	This field houses the verbatim value from the source data representing the unit of the Observation that occurred.	This code is mapped to a Standard Condition Concept in the Standardized Vocabularies and the original code is stored here for reference.	var char (50)	No
qualifier_ source_value	This field houses the verbatim value from the source data representing the qualifier of the Observation that occurred.	This code is mapped to a Standard Condition Concept in the Standardized Vocabularies and the original code is stored here for reference.	var char (50)	No
observation_ event_id	If the Observation record is related to another record in the database, this field is the primary key of the linked record.	Put the primary key of the linked record, if applicable, here. See the ETL Conventions for the OBSERVATION table for more details.	bigint	No
obs_event_field _concept_id	If the Observation record is related to another record in the database, this field is the CONCEPT_ID that identifies which table the primary key of the linked record came from.	Put the CONCEPT_ID that identifies which table and field the OBSERVATION_EVENT_ID came from.	int	FK
value_as_dateti me	It is possible that some Observation records might store a result as a date value.		date-time	No

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 824087.