Beyond One Million Genomes

D3.6

Phenotypic and clinical metadata framework - 3v0

Project Title (grant agreement No)	Beyond One Million Genomes (B1MG) Grant Agreement 951724		
Project Acronym	B1MG		
WP No & Title	WP3 - Standards & Quality Guidelines		
WP Leaders	Ivo Gut (CRG), Jeroen Belien (VUmc)		
Deliverable Lead Beneficiary	19 - VUmc		
Deliverable	D3.6 - Phenotypic and clinical metadata framework - 3v0		
Contractual delivery date	31/05/2023	**Actual delivery date**	31/10/2023
Delayed	Yes		
Authors	Jeroen Belien (VUmc, B1MG WP3 & 1+MG WG3, NL), Ivo Gut (CRG, B1MG WP3 & 1+MG WG4, SP), Harmke Groot (Nictiz, B1MG WP3, NL), Maarten Ligtvoet (Nictiz, B1MG WP3, NL), Pim Volkert (Nictiz, B1MG WP3, NL), Wei Gu (1+MG WG3, LU), Jan Korbel (1+MG WG3, GER), Michela Tebaldi (IRST, 1+MG WG3, IT), Milena Urbini (1+MG WG3, IT), Giovanni Martinelli (1+MG WG3, IT), Michela Riba (Ospedale San Raffaele IT, 1+MG WG3, IT), Catia Pinto (1+MG WG3, PT), Vanessa Mendes (1+MG WG3, PT) Inês Lourenço (1+MG WG3, PT). Ulrika Hermansson (1+MG WG3, SE), Alfonso Valencia (1+MG WG3, SP), Laima Ambrozaityte (1+MG WG3, LT),		

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

	Sonata Jarmalaite (1+MG WG3, LT), Kristiina Aittomaki (1+MG WG3, FI), Hannele Laivuori (1+MG WG3, FI), Rutt Lindström (1+MG WG3, EE), Peter-Bram 't Hoen (1+MG WG5, NL), Evita Maria Lindholm (1+MG WG3, NO)
Contributors	Regina Becker (UNILU, B1MG WP2 & 1+MG WG2, LU) Pablo Serrano (Hospital 12 Octubre, Spanish 1+MG mirror group, SP) Carlos Parra (Hospital Virgen del Rocio, Servicio Andaluz de Salud, Spanish 1+MG mirror group, SP)
Acknowledgements (not grant participants)	
Deliverable type	Report
Dissemination level	Public

Document History

Date	Mvm	Who	Description
01/10/2022	0v1	Jeroen Beliën (VUmc)	Initial draft created based on publication of D3.5
17/10/2023	0v2	Jeroen Beliën (VUmc) Nikki Coutts (ELIXIR Hub)	WP comments addressed.
25/10/2023	0v2	Nikki Coutts (ELIXIR Hub)	Version circulated to B1MG-OG, B1MG-GB & Stakeholders for feedback
31/10/2023	0v3	Jeroen Beliën (VUmc)	B1MG-OG, B1MG-GB & Stakeholder comments addressed.
31/10/2023	1v0	Nikki Coutts (ELIXIR Hub)	Version uploaded to the EC Portal

Table of Contents

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

1. Executive Summary

The aim of the 1+MG member states initiative with coordination support of the Beyond 1 Million Genomes (B1MG) consortium is to develop a pan-European genome-based health data infrastructure to further develop and operationalise personalised medicine and to understand pharmacogenomics. In order to support the 1+MG member states initiative ambitions for sustainability policies regarding assets in the field of personalised medicine interoperability, a set of simple but well-aligned instruments needs to be prepared. One of the crucial components for 1+MG is a phenotypic and clinical metadata framework which describes, in a commonly understandable language, the principles, models and recommendations for sharing and linking of phenotypic and clinical metadata and genetic metadata between the member states.

The current framework document proposes to adhere to standards for data capture and exchange. The main target is to provide guidance on which standards, terminologies and tools to use. Best practices, describing which ontologies are currently implemented in each member state, are described elsewhere (Deliverable 3.8: Documented best practices in sharing and linking phenotypic and genetic data - live working document[1]).

This phenotypic and clinical metadata framework document will be part of the entire B1MG project's outputs and guidance where other crucial instruments are taken care of by other work packages/working groups like Governance, ELSI and Technical Infrastructure. The relevant B1MG outputs will be part of the 1+MG Framework[2], once released and supported by the 1+MG WGs leaders, and will be used during implementation in the Genomic Data Infrastructure project[3] that has started November 1st 2022 as well as by the Genome of Europe Project to be granted early in 2024. While the 1+MG framework first of all aims at the 1+MG initiative, at the same time it could be used as a base for the European Health Data Space[4] (EHDS).

[1] https://doi.org/10.5281/zenodo.7342854
[2] https://framework.onemilliongenomes.eu/
[3] https://gdi.onemilliongenomes.eu/
[4] https://ec.europa.eu/health/ehealth-digital-health-and-care/european-health-data-space_en

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

2. Contribution towards project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives/key results:

	Key Result No and description	Contributed
Objective 1 Engage local, regional, national and European stakeholders to define the requirements for cross-border access to genomics and personalised medicine data	**1.** B1MG assembles key local, national, European and global actors in the field of Personalised Medicine within a B1MG Stakeholder Coordination Group (WP1) by M6.	Yes
	2. B1MG drives broad engagement around European access to personalised medicine data via the B1MG Stakeholder Coordination Portal (WP1) following the B1MG Communication Strategy (WP6) by M12.	No
	3. B1MG establishes awareness and dialogue with a broad set of societal actors via a continuously monitored and refined communications strategy (WP1, WP6) by M12, M18, M24 & M30.	No
	4. The open B1MG Summit (M18) engages and ensures that the views of all relevant stakeholders are captured in B1MG requirements and guidelines (WP1, WP6).	Yes
Objective 2 Translate requirements for data quality, standards, technical infrastructure, and ELSI into technical specifications and implementation guidelines that captures European best practice	**Legal & Ethical Key Results**	
	1. Establish relevant best practice in ethics of cross-border access to genome and phenotypic data (WP2) by M36	No
	2. Analysis of legal framework and development of common minimum standard (WP2) by M36.	No
	3. Cross-border Data Access and Use Governance Toolkit Framework (WP2) by M36.	No
	Technical Key Results	
	4. Quality metrics for sequencing (WP3) by M12.	No
	5. Best practices for Next Generation Sequencing (WP3) by M24.	No
	6. Phenotypic and clinical metadata framework (WP3) by M12, M24 & M36.	Yes
	7. Best practices in sharing and linking phenotypic and genetic data (WP3) by M12 & M24.	Yes
	8. Data analysis challenge (WP3) by M36.	No
	Infrastructure Key Results	
	9. Secure cross-border data access roadmap (WP4) by M12 & M36.	No
	10. Secure cross-border data access demonstrator (WP4) by M24.	No

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

Objective 3 Drive adoption and support long-term operation by organisations at local, regional, national and European level by providing guidance on phased development (via the B1MG maturity level model), and a methodology for economic evaluation	**1.** The B1MG maturity level model (WP5) by M24.	Yes
	2. Roadmap and guidance tools for countries for effective implementation of Personalised Medicine (WP5) by M36.	Yes
	3. Economic evaluation models for Personalised Medicine and case studies (WP5) by M30.	No
	4. Guidance principles for national mirror groups and cross-border Personalised Medicine governance (WP6) by M30.	Yes
	5. Long-term sustainability design and funding routes for cross-border Personalised Medicine delivery (WP6) by M34.	No

3. Methods

The collection, analysis, use and sharing of genomic data promises major breakthroughs in health research, more specifically for personalised medicine and for population health. Personalised medicine research relies on more than just data generated by genome sequencing; it also entails the study of a patient's overall health, thus the need to link (or match) genomic data with relevant and accurate (time-dependent) phenotypic data, such as environmental data, information in medical records (e.g. lifestyle factors, comorbidities) and administrative data. As such, to ensure optimal use of genomic datasets for research and development of personalised medicine, linkage of genomic and health related data is a cornerstone for realising the potential that genomic data offers to improve health.

Across Europe there are different data sources of health related data, different taxonomy and ontology codes to label the same condition, making comparisons of different datasets challenging.

Semantic unification is therefore a core priority when developing a European framework for exchange of health data both for primary use, where data has been collected for the specific purpose or analysis under consideration, i.e. data from experiments, interviews, surveys and direct observations. Additionally, such unification is important for secondary use, where data has been collected in the past by someone else, and later made available for others to reuse for a similar or a different purpose than originally intended. Since data can be collected to answer both specific questions, as well as for other purposes, the same data can function both as primary and secondary data sets.

Semantic unification ensures that the defined health/medical concepts are unambiguous and that both primary data users (e.g. a medical specialists) as well as secondary data users have the same interpretation of the exchanged (health/medical) concepts. Harmonising and standardising coded health(care) data improves safety and efficacy of healthcare and facilitates secondary use, i.e. for healthcare services as well as research and management (quality, value-based health

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

care, administration), without loss of "meaning" (HIMMS[5]), in other words 'what is sent is what is understood'.

Moreover, this document will serve as a semantic interoperability framework component as laid out in the European Interoperability Framework and depicted in the picture below showing the interactions between the generic European Interoperability Framework (EIF), the National Interoperability Framework (NIFs) and the Domain specific Interoperability Framework (DIFs). The European interoperability framework is a commonly agreed approach to the delivery of European public services in an interoperable manner. It defines basic interoperability guidelines in the form of common principles, models and recommendations. The EIF provides a common core of interoperability elements to European NIFs and DIFs. Compliance with the EIF guarantees that NIFs and DIFs are developed in a coordinated and aligned way while providing the necessary flexibility to address specific requirements coming from national or domain specific requirements. It remains to be investigated whether, or not, this concerns being part of an existing DIF, the refined eHealth European Interoperability Framework (ReEIF[6]), or if the 1+MG Interoperability Framework will be a domain specific interoperability framework overlapping with other defined Interoperability Frameworks.

[5] https://www.himss.org/sites/hde/files/d7/FileDownloads/HIMSS%20Interoperability%20Definition%20FINAL.pdf
[6] https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5b56dffdc&appId=PPGMS

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

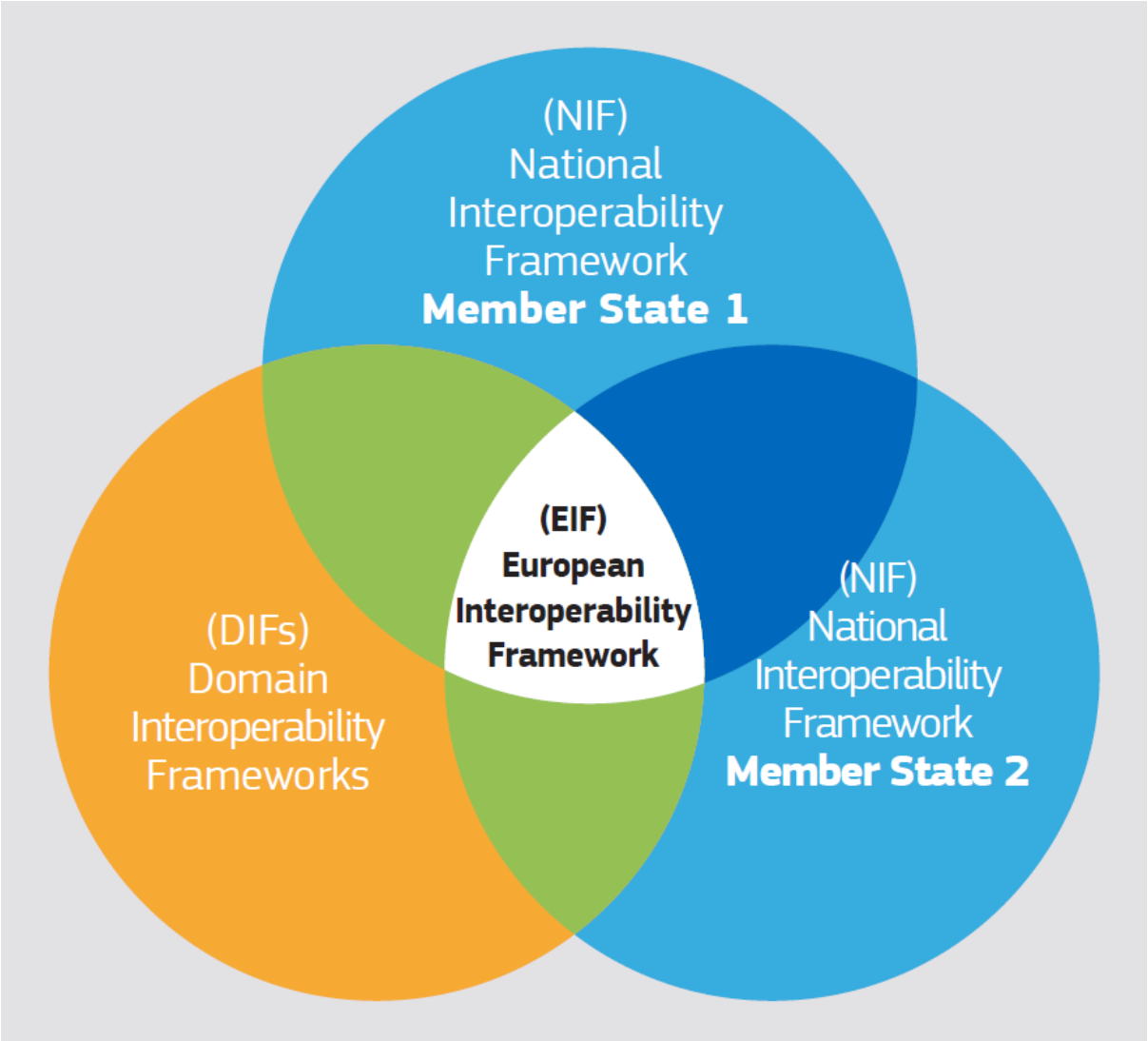

Figure 1: The relationship between EIF, NIFs and DIFs as presented in https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf[7]

Within the EIF an interoperability model is presented as shown in Figure 2 below:

[7] https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

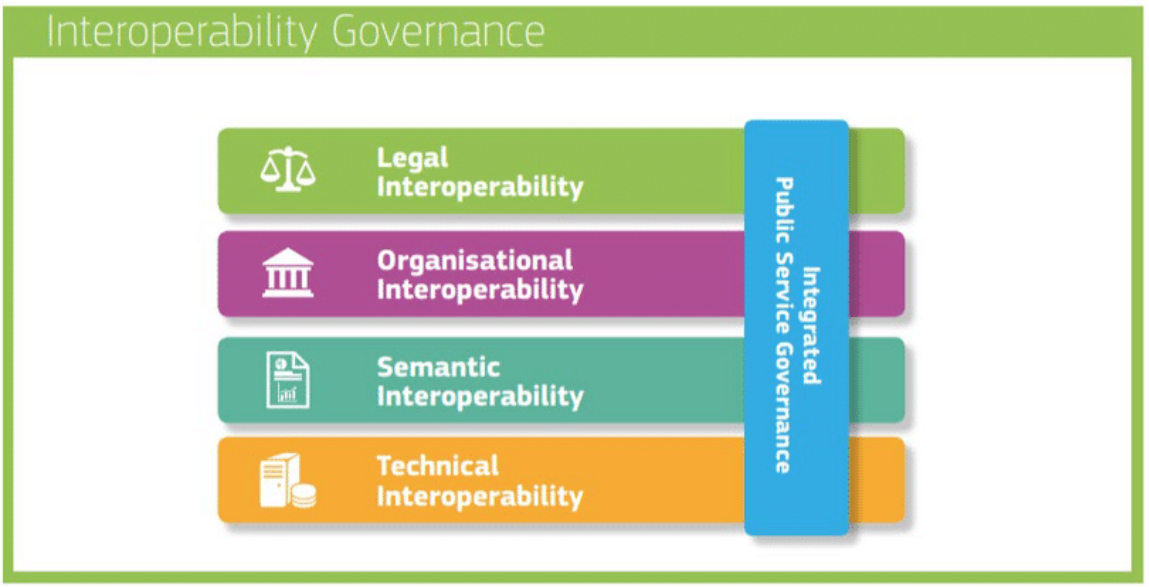

Figure 2: The new European Interoperability Framework (EIF) model which includes the semantic interoperability model (In the EIF, semantic interoperability covers both semantic and syntactic aspects) covered by this framework document.

Compliance with the EIF guarantees that the 1+MG specific DIF is developed in a coordinated and aligned way while providing the necessary flexibility to address specific requirements for sharing and linking phenotypic and genetic metadata between the member states.

In parallel with working on this framework document, best practices were identified for each member state, describing which ontologies are currently implemented in each member state for several use cases in which health related data will be recorded (amongst which Pathology (including digital pathology), cancer and rare disease diagnoses (including medical imaging reports), clinical laboratory data, medication and other related personal data). Of note, best practices identified for individual states could range from being explored in specific healthcare levels or judicial areas, to being implemented nation-wide. It is of high relevance to align which ontologies are used as much as possible to facilitate data exchange and interoperability when developing a genomic data infrastructure. The results of this inventorisation are described elsewhere (**Deliverable 3.8: Appendix III: National implementation of standards** Documented best practices in sharing and linking phenotypic and genetic data - live working document[8] or the stable published version at Zenodo[9]).

Besides ultimately aiming for semantic interoperability, identifying and accessing relevant existing datasets and/or initially (as an agile approach) agreeing to a minimal representative datasets is challenging by itself.

In July 2020 the 1+MG Coordination team and Working Group leaders launched the 1+MG - Survey on accessible genomes[10] to understand what existing genomic datasets and corresponding phenotypic/clinical information are effectively available for participation in the 1+MG. The survey results helped to understand what genomic datasets and corresponding

[8] https://docs.google.com/document/d/1GOvcR3l3t8T4cJLDVx7kVxbYa9F2oo7deQveMG6Og6c/edit
[9] https://doi.org/10.5281/zenodo.7342854
[10] https://ec.europa.eu/eusurvey/runner/1plusMG_Survey2020

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

phenotypic/clinical information exist. This survey contained detailed questions about availability of data dictionaries and which ontologies have or are being used. Based on the responses of this survey we have contacted some of the submitters to discuss their submission to identify the challenges and bottlenecks for sharing, as well as jointly making recommendations for design of a European framework for sharing genomic and associated clinical data (this framework document). The results obtained show that most genomic datasets have linked clinical data, but that since these datasets for the larger part all have been generated in the past a substantial effort is needed to align all these datasets on format, structure, interoperability and quality. To make those datasets findable as well as accessible (i.e. if one is interested in a dataset one can contact the dataset owner) a proof of concept has been initiated to create the European Genome Dashboard (B1MG task 3.6). This dashboard is now live at: https://dashboard.onemilliongenomes.eu/[11]

To make existing as well future datasets interoperable we, in collaboration with the 1+MG use case working groups, propose to define minimal datasets following a multi-level standardisation model as depicted in Figure 3, The sunflower metaphor.

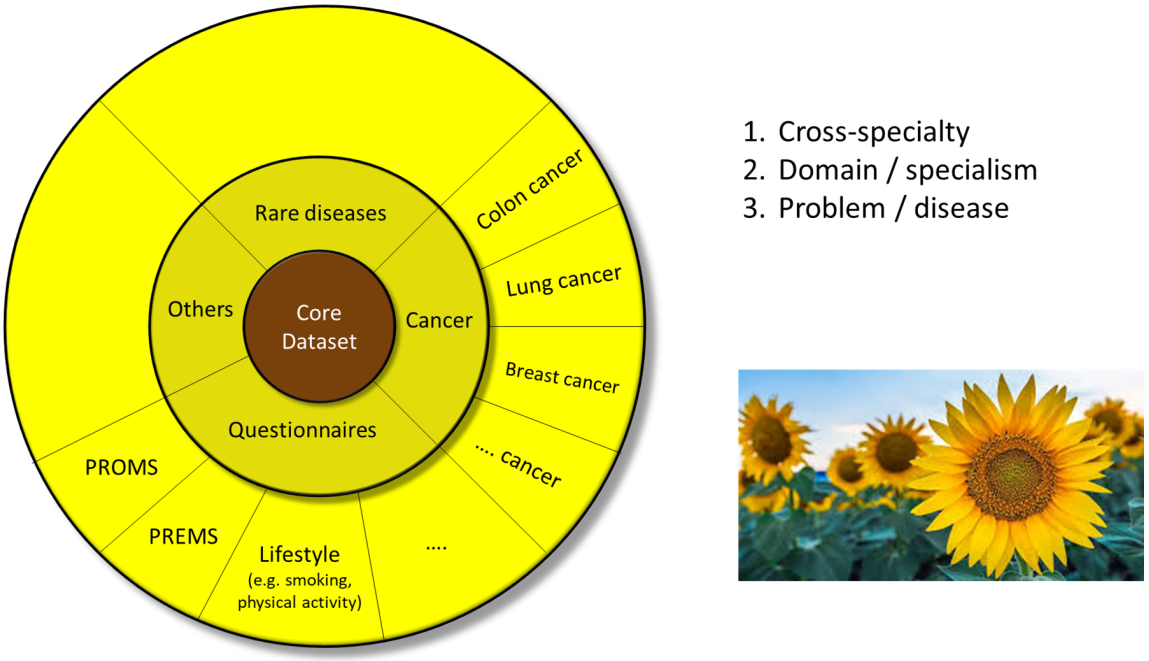

Figure 3: Multi-level standardisation. The core minimal dataset contains cross-specialty mandatory fields, the next layer inherits the mandatory fields of the core and adds mandatory fields specific for a domain or specialism. The outer layer (petals of sunflower) adds the problem/disease specific mandatory fields.

Guidance is needed in order to reach interoperability. Which standard is prioritised in which context? The International Classification of Diseases (ICD). The International Classification of Disease (ICD) is currently the most widely used classification system for diseases in the world. It serves a broad range of uses globally, and provides critical knowledge on the extent, causes and consequences of human disease and death worldwide via data that is reported and coded with the ICD. All Member States are committed to use the most recent version of ICD, and currently over 100 countries use ICD-10. However, several of these are in different stages of ICD-11

[11] https://dashboard.onemilliongenomes.eu/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

implementation, which is currently the newest release. ICD-11 offers several advantages over ICD-10, including content updates, tooling improvements and additional language versions. In contrast to ICD, SNOMED Clinical Terms (CT) is a rich comprehensive clinical terminology which has been used by medical professionals in over 50 countries for more than decades. SNOMED CT facilitates data capture, like disorders and procedures, in multiple languages and patient-friendly terms. Another advantage is that this polyhierarchical ontology can be queried using Expression Constraint Language (ECL) and concepts can be pre-or post-coordinated.

SNOMED CT and ICD-10 have complementary strengths, and situations may benefit from use of both code systems. With this respect, the workgroup will follow the guidelines written by the eHN Subgroup on Semantics. It is anticipated that the WHA meeting in 2025 will provide further details on how these standards may be used simultaneously and provide a roadmap. Since terminologies have different structures and intended uses, mappings provide a linkage between different terminologies or between terminologies and classifications. This task involves identifying related terms or concepts in each system, and although performed consistently to reduce incompatibilities it includes the risks of mistakes, misinterpretations and a relatively high cost for resources, maintenance and translations. These aspects should be considered in cost-benefit analysis before implementing different terminologies that overlap in use.

It is important to define which terminologies are best suited for describing the different areas within healthcare/clinical terminology, which can complement existing requirements and, where possible, define standards suitable for general use. The definition level scheme is essential for providing information at the required level of granularity and for homogenising definitions provided by different sources at different levels of granularity. In addition, the schemes must be flexible enough to include extensions to the levels and new classes as required by domain experts. These operations must be performed respecting the hierarchical structure and format of the data. The appropriate tools for those operations should be considered part of the infrastructure necessary for the use of the dictionaries and ontologies

As part of the framework document, and based on the best practices inventory, as well as the experiences from working towards a minimal dataset with each of the use case working groups, we provide:

- an overview on relevant code systems and terminologies (dictionaries) and their characteristics, including information on which systems are mapped to others, and where their domain of interest overlaps, as well as known data architecture models. The live version of this overview has been added as Appendix I
- advice on which standards are of preference or recommended: see paragraph 4.1.4

3.1 An approach for working with common standards

Introduction

Data sharing works best if the data is based on, or can be modelled along a set of well-defined high-quality standards. The steps outlined below describe the steps that can be taken to establish such a set of standards.

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

Measure the current state of maturity

The B1MG MLM (Maturity level model)[12] can be used to assess the current effectiveness of the set of standards. The MLM defines levels of effectiveness that start with ad hoc practices and end with stable general practices. The MLM can also be used to assess at which levels the current healthcare and research systems support standards. The MLM is a multidimensional model among which topics are covered such as: governance, strategy, economic, legislation, policy, education, standards, infrastructure and public awareness. The model can be used to develop a roadmap for optimising maturity in the standards domain and adoption of standards by research projects and healthcare systems.

Collect use cases

A set of use case definitions should be collected. Besides describing aspects like purpose of data access and other specifics of the data request, a functional specification of the data is needed. What kind of data is needed (genomic type of genomic information, symptoms, treatment, phenotypic information, demographic data, etc.).

ReEIF
The eHealth Network (eHN) has created the Refined eHealth European Interoperability Framework (ReEIF[13]), which describes an interoperability model and template for documenting use cases. The ReEIF model documents actions that are needed per interoperability layer:

Layer		Topic	Holder
•	Legal and regulatory	Compatible legislation and regulations	B1MG WP2
•	Policy	Collaboration agreements	B1MG WP2
•	Care Process	Alignment of care processes and workflows	
•	Information	Datamodel, terminologies, formatting	B1MG WP3
•	Applications	Integration in healthcare systems	
•	IT Infrastructure	Communication- and network protocols	B1MG WP4

Although this framework is mainly focussed on *care* processes, the concepts are useful for broader interoperability projects. The top layer describes how these projects can function within the legal and regulatory environment. At the policy level, agreements can be made between parties that exchange health information. The information layer contains the functional description of the data model, data elements and linking to terminologies. The application layer details the technical specification of how information is recorded or transported.

The following steps can be distinguished when creating a data model:

- Use case quality check
  Starting from the collected use cases, perform a quality check. Are the collected use cases complete and the semantics clear? If this is not the case, the author(s) should be contacted to further document the use case.

- Use case harmonisation

[12] https://doi.org/10.5281/zenodo.8205631
[13] https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5b56dffdc&appId=PPGMS

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

Use cases can contain generic or more specific content. During the harmonisation phase, parts can be identified that are used in multiple use cases, or multiple times in a single use case. Usually, more generic elements, such as for instance demographic data elements, appear in almost all use cases. More specific data elements might be limited to one use case. This closely links to the multi-level standardisation as proposed in Figure 3: generic (mandatory) elements at the heart/core and multiple specific elements around it.

Adopt existing standards

Wherever possible, existing standards should be re-used when creating a data model. Data elements contained in use cases may have been built on already existing standards. If that is the case, analysis can be done to identify exactly which standards. Depending on the type of standard, the *scale* on which the standard can be adopted can differ, such as international, national, regional. Depending on the limitations contained in the standard and the use case at hand, it might be applicable to either extend or specialise parts of the existing standard. When a relevant code is not available in the desired ontology for a specific use case, then it may be possible to suggest the missing code(s) for inclusion following existing governance procedures.

Creating, adapting and/or extending standards

If the use case cannot be mapped in its entirety or in part to existing standards, it can be applicable to create a new standard (but only if really none of the elements can be mapped), or preferably adapt and/or extend existing standards that are modelled after the collected use cases. When modelling, the more generic data elements can be used multiple times for different use cases. Each use case can have different actors involved and can specify a different cardinality and conformance for each element. An actor can be a person (such as a doctor or a patient), an organisation, biobank, or even a computer system providing a healthcare service. Cardinality specifies the allowable occurrences within the use case of a certain data element. Conformance expresses whether an actor must support an element or not. Data elements can have a number of attributes such as: version, id, status, name, description, source, rationale, operationalization, comment, value type, value properties, value example and relations to other data elements.

Linking data elements to terminology

Data elements (e.g. a variable in a dataset or a column name in a spreadsheet) can be linked to terminology in a number of ways. One common association is the link between a data element and a specific code from a code system. By associating data elements to terminology codes in a structured format, this metadata makes the data element meaningful and machine processable. Another type of association can be created between a data element and a value set (e.g. the allowed values in a cell of a spreadsheet) to specify the allowed codes associated with the data element.

Figure 4 shows an example of how existing ontologies are interrelated. Some (local) standards (the orange shapes) and existing international (domain specific) endorsed standards and/or aggregation terminologies (dark blue shapes) are (or will be) mapped, while others might have distinct purposes (e.g. HPO and LOINC) and cannot be or only partly be mapped. In reality, there are many more domain-specific or more general standards available; therefore, to reach maximum interoperability, it is of high importance to specify which are standards of preference for which particular set of usage (e.g. 1+MG use case for rare disease or cancer), variables or clinical features. (see also work in progress as mentioned above in Appendix I)

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

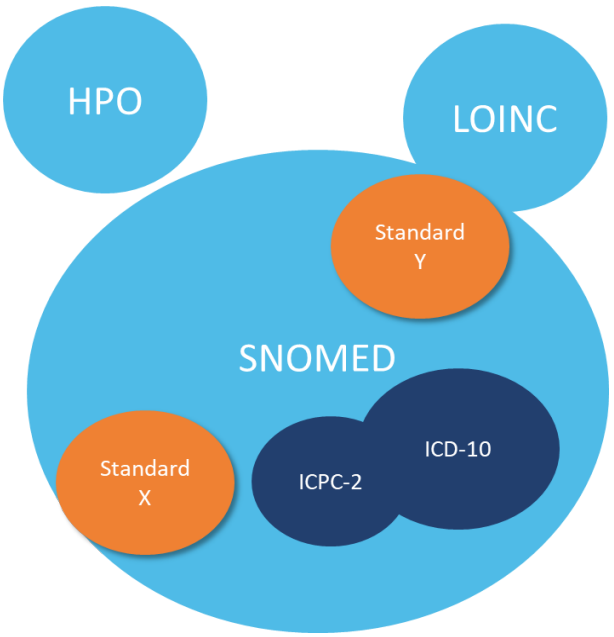

Figure 4: Core terminology overview on (inter)national level: standardisation and mapping

The recently released ISO standard 13972:2022 provides guidance on definition and maintenance of clinical information models. This standard facilitates semantic and technical interoperability on the European level. It also describes clinical information models, their content, structure and context and specification of their data elements. [reference: https://www.iso.org/standard/79498.html.]

Project administration and governance

Parties that contribute to the project can function in different roles. NEN 7522:2020 nl[14] - *Health Informatics – Development and maintenance of standards systems of standards* is a norm written in the context of Dutch healthcare (originally limited to terminology) which contains a description of roles that can be used in a wider context.
A brief summary of these roles is included here:

- Holder of the data standard: owner, responsible for development and maintenance
- Financier of the data standard: decides upon a finance structure
- Authorizer of the data standard: responsible for decision making, for instance for changes
- Functional manager of the data standard: performs the functional management and maintenance as well as version management
- Technical manager of the data standard: performs technical management
- Distributor of the data standard: responsible for distribution
- Expert for a specific data standard or functional (clinical/phenotypic) data domain: contributes expertise relating to the information technology or purpose of use
- Computer Science experts working on the tools for ontology mapping and alignment to reduce the conflicts and/or increasing the concordance between ontologies - with applications to the biomedical domain
- User of the data standard: implements standards into applications. For example, an application vendor

[14] https://www.nen.nl/en/nen-7522-2021-nl-283706

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

- End user of the data standard: user of a component that implements standards. A source or consumer of data.

When defining a governance scheme for a standard, each role should have a natural or legal person assigned to it. A governance scheme should also include a rights policy that may include copyright and licence statements.

Quality check with expert or stakeholder

The development process should include clear phases during which experts and/or stakeholders are consulted on the contents of the standard. Depending on the maturity of the standard (e.g. Changing culture with the Dataset Maturity Model (elixir-europe.org), Identifying ELIXIR Core Data Resources | F1000Research[15]), a public consultation may be included in the development cycle.

Applications layer

Once the functional information layer is established, the standard can be represented in a technical format after which a quality check and revision can be scheduled. The technical format should be linked to the functional information layer so that metadata can be leveraged and becomes machine readable. Users of the standard should be presented with an explicit, implementable format so that healthcare information can be recorded and exchanged unambiguously. Implementing national terminology servers could benefit and increase the potential use of established standards. Examples of such national terminology services (i.e. by providing a server or terminology API) can be found in Australia (NCTS[16]), the UK (The NHS Digital Terminology Server: a digital standard server[17]), the Netherlands (De Nationale Terminologieserver - Nictiz[18]), New Zealand (NZ Health Terminology Service (NZHTS)[19]) and Switzerland[20].

In summary we therefore propose the following approach to help organisations and/or countries to arrive at high quality and governed FAIR clinical and phenotype metadata and when combined with genome data arrive at FAIR genomes. The steps are more or less the best practices but can be adopted to local situations as deemed appropriate.

- Step 1: Contact all relevant centres/organisations/stakeholders and invite them to actively participate/contribute
- Step 2: Learn from the participants what they consider important to find/reuse (their and others') data and discuss what they need and have in the context of the known standards (international, national as well as local)
- Step 3: Distil the commonalities into a 'metadata' schema: define what metadata is needed to find, share and reuse clinical and phenotypic (as well as genome) data in research and healthcare
    - Step 3.1…Step 3.n Have as many iterations with experts to arrive at a common metadata schema: Forming an evolving semantic schema of essential elements
- Step 4: Create prototype systems. Test-drive the schema. Improve. Become FAIR in theory.

[15] https://faircookbook.elixir-europe.org/content/recipes/maturity.html
[16] https://www.digitalhealth.gov.au/newsroom/product-releases/new-national-clinical-terminology-service-website
[17] https://digital.nhs.uk/services/terminology-servers#:~:text=The%20NHS%20Digital%20Terminology%20Server%20is%20a%20FHIR%C2%AE%20compliant,Data%20Model%20and%20Dictionary%20codes
[18] https://nictiz.nl/wat-we-doen/activiteiten/terminologie/de-nationale-terminologieserver/
[19] https://www.tewhatuora.govt.nz/our-health-system/digital-health/terminology-service/
[20] https://sphn.ch/2021/09/10/dcc-terminology-service/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

- Step 5: Implement and use the metadata schema. Become FAIR in practice

The above approach is based on the Maelstrom data harmonisation guidelines for retrospective data harmonisation , see https://www.maelstrom-research.org/page/maelstrom-guidelines[21] and has been adopted and modified towards FAIR clinical and phenotype metadata.

4. Description of work accomplished

4.1 Introduction

A robust infrastructure is needed for pooling data from different domains and which enables data extraction and analysis, multilingual representation, and facilitates output on the individual level. Alignment of information standards is therefore important to ensure secure digital cross-border exchange of both clinical/phenotype and sequencing data, and data pooling. Interoperable standards for sharing these data are essential for data re-use, compiling large European cohorts (e.g. on rare diseases) but also to facilitate the enrichment of the genomic and clinical dataset with other types of phenotypic data (e.g. lifestyle, quality of life, and health status, including diagnosis of cancer or cardiovascular diseases).

4.1.1 Data capture: development towards a minimum clinical dataset

All member states completed a survey[22] to identify accessible genomes with, if available, linked, clinical data. Results are presented via the European genomes dashboard: https://dashboard.onemilliongenomes.eu/. In addition, potentially relevant minimal datasets (domains: research, industry, existing international networks) were identified and workshops were held with domain experts from both the B1MG project as well as the 1+MG WGs to define the relevant minimal dataset for specified use cases. The creation of these agreed upon minimal (meta)datasets by all the use-case WGs of 1+MG followed the approach for working with common standards as described at the end of paragraph 3.1. All intermediate results have and will be shared with the relevant WPs/WGs. The minimal (meta)datasets of each of the 1+MG use-case WGs will be compared to detect commonalities as well as to arrive at harmonisation between them where possible. The targeted end-user of these datasets is a researcher, health(care) professional, or patient.

The status quo of the 1+MG minimal phenotypic datasets of cancer, rare disease, infectious disease and common and complex diseases/GoE as discussed and confirmed during the 1+MG/B1MG face to face meeting in Brussels September 11th - 13th 2023 is listed below:

The 1+MG WG8 minimal dataset for rare disease consists of the 16 agreed upon Common Data Elements (CDEs) which have been published.

- 1+MG minimal dataset for rare diseases:
  - https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en
  - https://eu-rd-platform.jrc.ec.europa.eu/system/files/public/CDS/EU_RD_Platform_CDS_Final.pdf

[21] https://www.maelstrom-research.org/page/maelstrom-guidelines
[22] https://ec.europa.eu/eusurvey/runner/1plusMG_Survey2020

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

The 1+MG WG9 minimal dataset for cancer encompasses 140 items (37 mandatory, 40 recommended and 63 optional) and is organised in eight conceptual domains for the collection of cancer-related clinical information and genomics metadata.

- 1+MG minimal dataset for cancer
  - 1+MG WG9/WG3 joined pre-publication of paper outlining the entire process from start till current result: https://doi.org/10.1101/2023.10.07.561259
  - 1+MG WG9/WG3 minimal dataset at Zenodo (master DOI: https://zenodo.org/doi/10.5281/zenodo.8239362)
  - 1+MG WG9/WG3 minimal dataset at ART-DECOR: https://art-decor.org/ad/#/b1mgsetcancer-/project/overview

The 1+MG WG11 minimal dataset for infectious diseases is still work in progress (initial version being reviewed) and only has taken the first steps as described :

- Conceptual domains currently identified are: identification, Country of hospitalisation, Case status, Demographics, Comorbidities, Hospitalisation information, Infection diagnosis and Vaccination
- A total of 24 variables (concepts) and where relevant accompanying value sets

The 1+MG combined WG10 and WG12, with a focus Genome of Europe, have defined a very limited minimal dataset consisting of 3 mandatory items: biological sex at birth, Body Mass Index (BMI) and age (at time of retrieving sample for sequencing).

4.1.2 Operationalising genotype and phenotype data

Which phenotype data[23] should be coded and how can interoperability be ensured? The minimum datasets are initially/primarily targeted to contain clinical information that can be linked to, or integrated with human genome/ sequencing data such that the initial questions as defined by the 1+MG use case working groups can be answered. For additional or specific questions/use cases, the minimal datasets can also be extended with data on lifestyle (e.g. smoking, physical activity), quality of life, incidence of certain disorders like cancer and cardiovascular disease, vital status, and/or others, given the sunflower metaphor as shown in Figure 3.
Next to the 1+MG use cases per domain, suitable terminology standards will be identified and these will be aligned with existing best practices in member states. This ensures interoperability and suitability for use in 1+MG, medicine and research.

4.1.3 Interoperability: using common terminology standards

The eHealth Network (eHN) developed the Refined eHealth European Interoperability Framework (ReEIF). B1MG WP3/1+MG WG3 targets the information layer of this framework. We have evaluated the maturity status of the most commonly deployed terminology standards (ontologies) in the participating member states (as part of the MLM model, as set-up by B1MG WP5). In parallel, relevant EU and national projects were consulted as well: Healthycloud, European Health Data Space, FAIRplus, FAIR genomes. Standards were classified by use case and/or domain where relevant (e.g. cancer, rare diseases). These results have been published as part of B1MG Deliverable 3.8[24] The aggregated results are shown below (Figures 5.1 - 5.5). In total 18 member states participated so far. Based on the responses we were able to aggregate results for the domains of pathology and laboratory, the 1+MG WG use cases domains of cancer and rare disease, and reporting diagnosis.

[23] Phenotype data and metadata as defined by https://www.genome.gov/about-nhgri/Policies-Guidance/Data-Sharing-Policies-and-Expectations/metadata-phenotypic-data-sharing and where we define clinical data as a subset of Phenotype data that relates to healthcare records, clinical studies and other clinical relevant data.
[24] https://doi.org/10.5281/zenodo.7342854

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

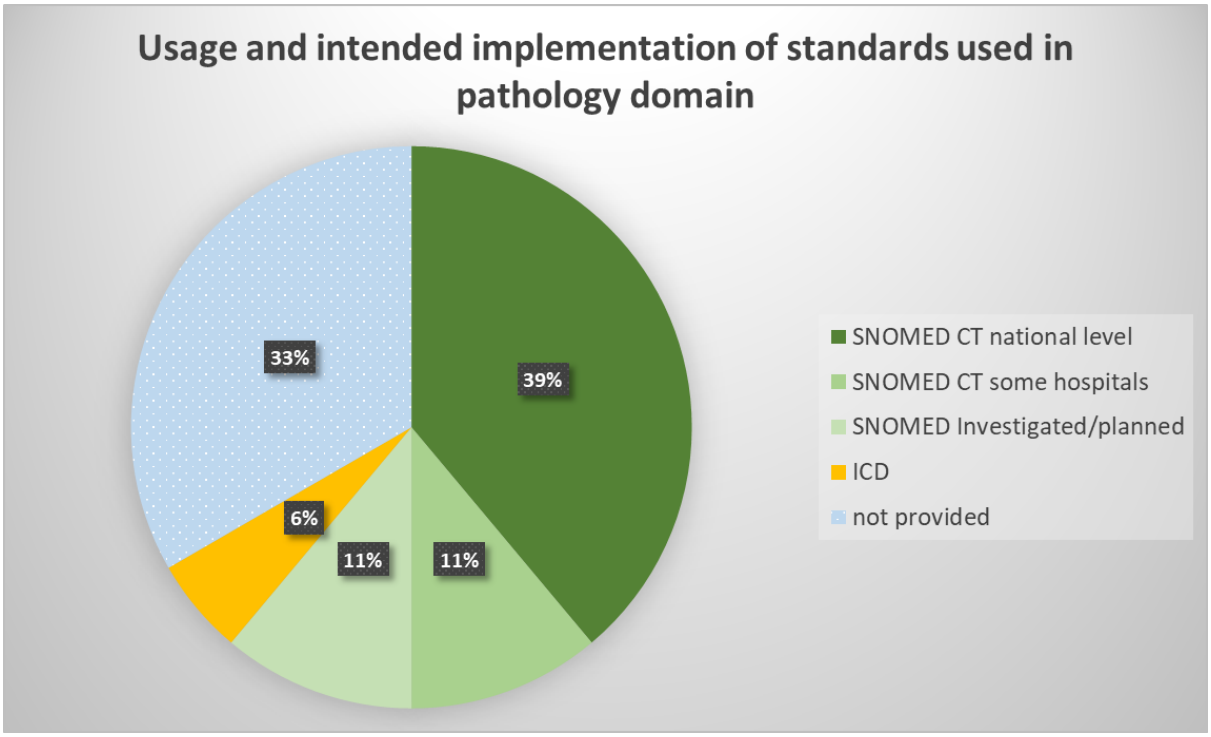

Figure 5.1: Usage and intended implementation of standard used in the pathology domain. Note that 33% of the member states did not provide any specific detail/response for the pathology domain.

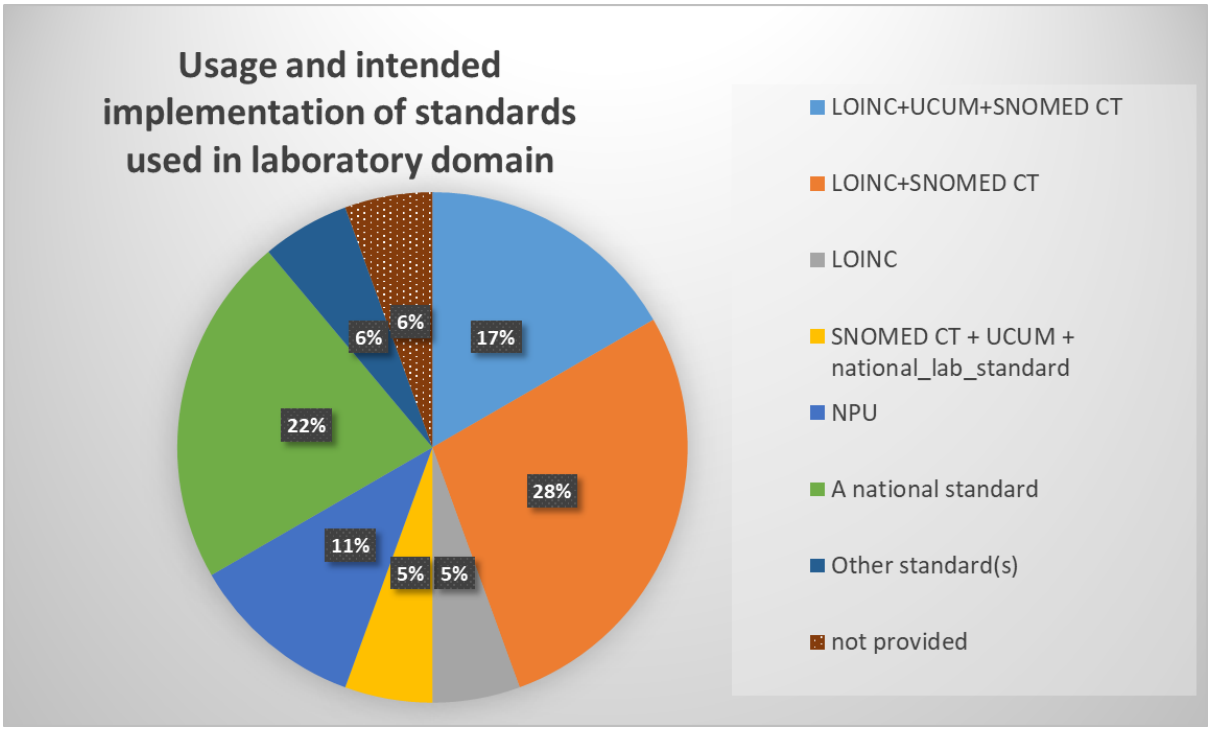

Figure 5.2: Usage and intended implementation of standard used in the laboratory domain. Note that 6% of the member states did not provide any specific detail/response for the laboratory domain.

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

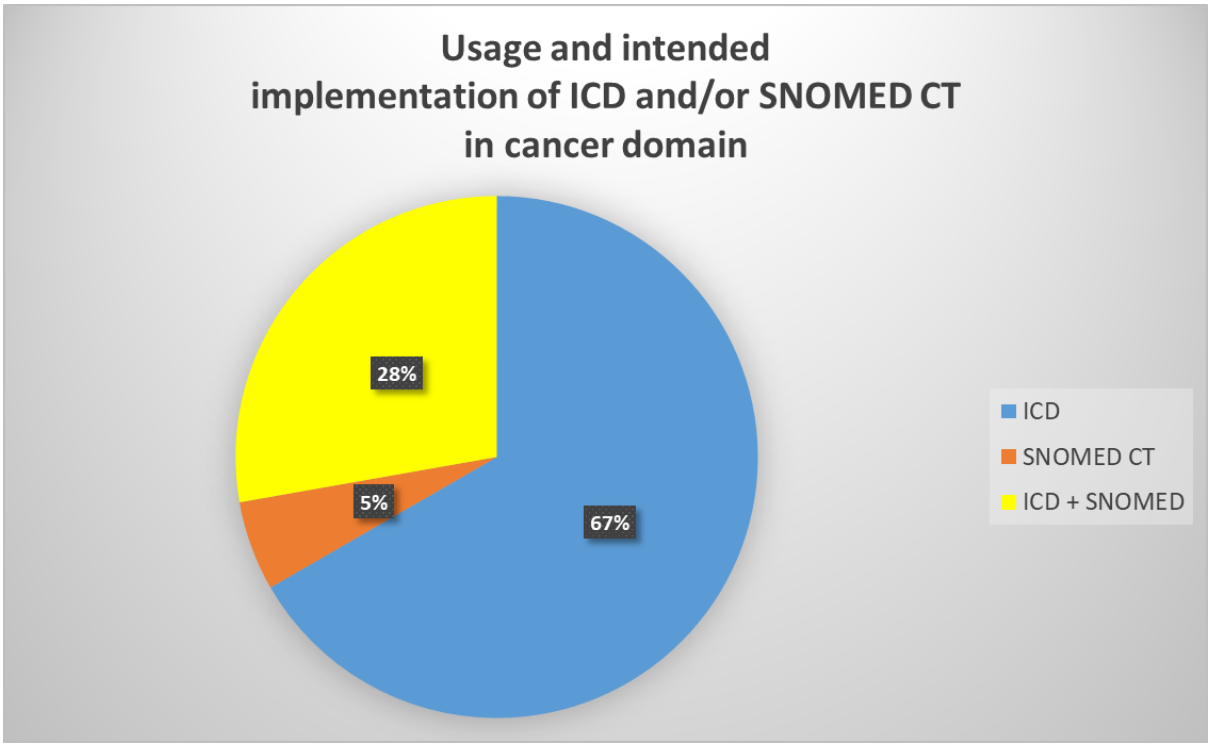

Figure 5.3.a: Usage and intended implementation of standard used in the cancer domain.

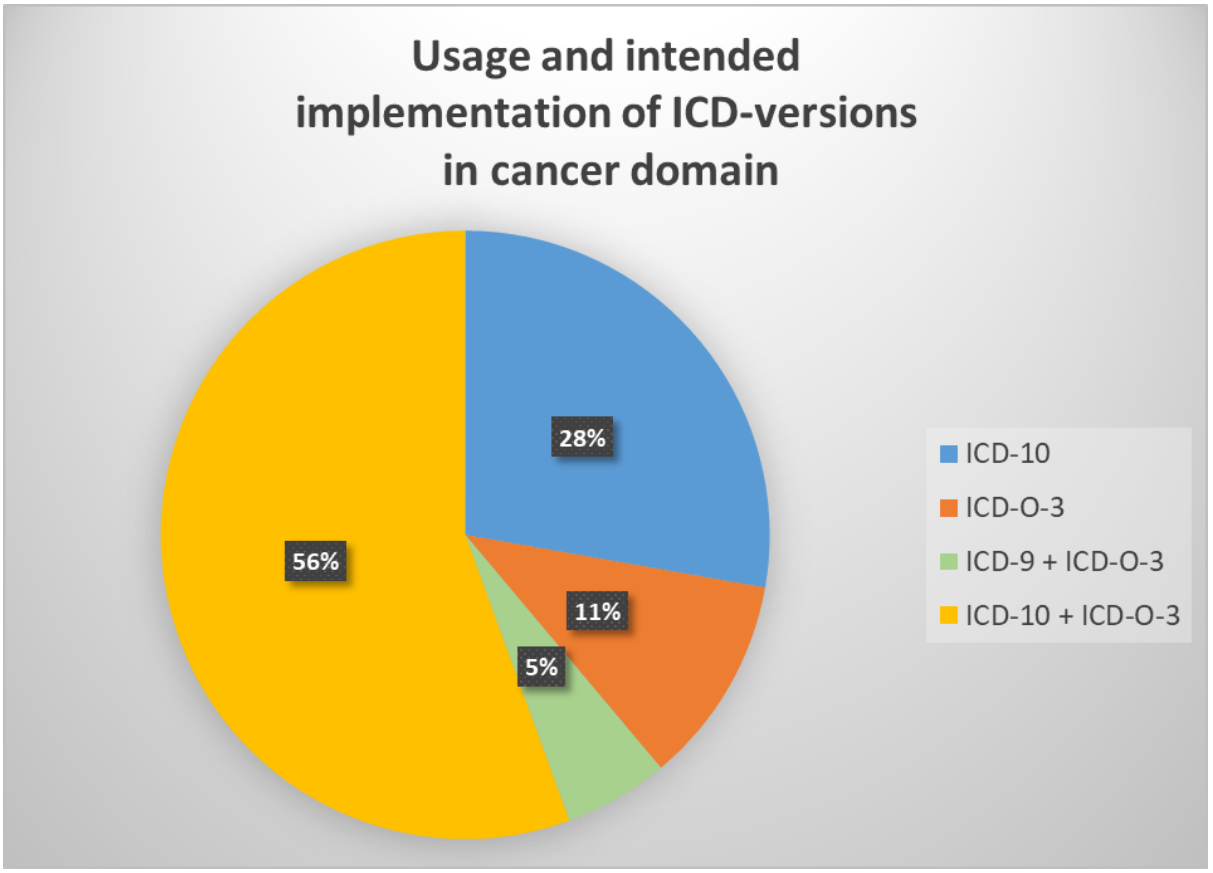

Figure 5.3.b: A detailed graph on various ICD versions used within the slice ICD of Figure 5.3.a.

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

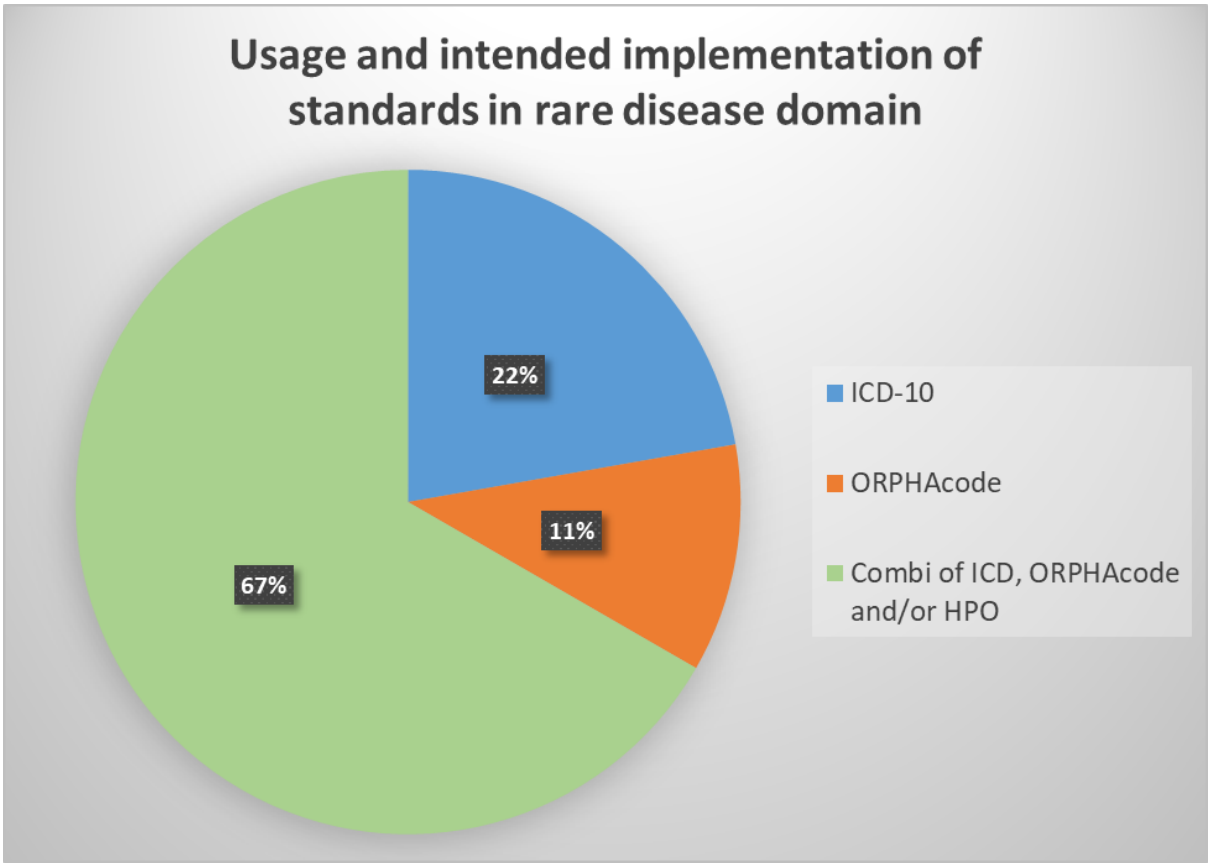

Figure 5.4: Usage and intended implementation of standard(s) used in the rare disease domain.

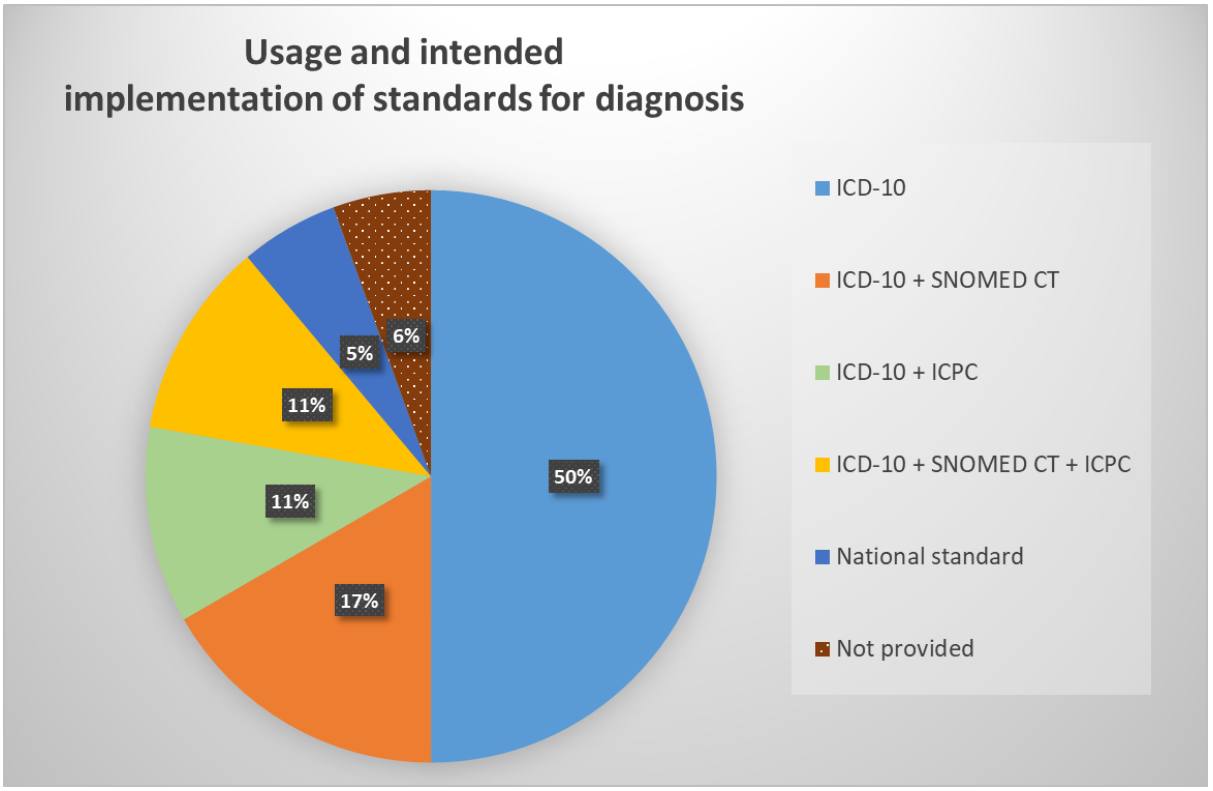

Figure 5.5: Usage and intended implementation of standard(s) used for reporting diagnosis.

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

A white paper published in 2021 by MedTech Europe and COCIR states that SNOMED, LOINC, DICOM, HL7 standards should be defined as 'base standards' to reach semantic interoperability. It is noted that this white paper only focusses and discusses the initiatives that are most relevant to the purposes of this paper, namely DICOM, HL7, IICC, IHE, LOINC, SNOMED CT, UCUM, Continua, and IEEE. Since there are many terminologies available, and there is increasing overlap, it is difficult to reach interoperability[25]. Minimum domain-specific datasets thus provide valuable insight into the overlap (or discrepancies) in which data fields are required, how they are coded, and which terminologies are used. It is of high importance to identify preferred standards for those generic data items that overlap (i.e. are common) to each data set. Subsequently, it is also relevant to specify which information models are visioned for primary data recording, data exchange as well as secondary use (e.g. research) **(Appendix I)**, Advice on which ontologies or data models are of preference is in line with the Common Semantic Strategy for Health in EU, which aims to provide guidance for EU level decisions on health semantics by 2025.

Current prioritised standards for cross-border exchange[26] (acute healthcare; eHN patient summary guidelines[27]) are LOINC, SNOMED, UCUM, WHO ATC, ICD-10, ORPHAnet, EMA SMS substances, EDQM, HL7 v3/FHIR, ISCO-08, EMDN and ISO Country and language codes. It is therefore of importance to use these standards for common data items in the minimal datasets, where possible.

Based on the inventory of the most commonly deployed terminology standards (ontologies) classified by use case and/or domain where relevant (e.g. cancer, rare diseases) in the participating member states, the literature, the results from the joined B1MG/1+MG WGs meeting in Brussel (Sept 2023), and the current prioritised standards for cross-border exchange of EHR data, we for now, in summary, advise the usage of the following standards/terminologies:

- For pathology: SNOMED CT (preferred)
- For laboratory: the combination of LOINC, SNOMED CT and UCUM (preferred)
- For cancer: SNOMED CT (preferred) with mapping to ICD-10 and ICD-O-3
- For rare diseases: a combination of ICD-10 with ORPHAcodes (part of Orphanet)
- For phenotypic abnormalities: HPO
- For common and complex diseases: when related to diagnosis then mainly ICD-10 and for the near future ICD-11 but also SNOMED CT
- For direct and indirect cause of death: ICD-10 and for the near future ICD-11
- For cardiovascular diseases or comorbidities: SNOMED CT
- For capturing medicinal data: ISO IDMP[28] is recommended
- For exposure ascertainment , the following validated questionnaires are recommended (just a snapshot (some examples), this list is extensive; see also ICHOM):
    - Quality of Life: SF-12, SF-36 or EORTC-QLQ-C30,
    - PROMS/PREMS,
    - Smoking: GATS, lifetime smoking status, pack-years,
    - Physical activity: IPAQ,
    - Obesity: BMI, waist circumference

For the following standards, a licence is needed:

[25] https://www.cocir.org/fileadmin/Publications_2021/2021-10_COCIR_-_MTE_Interoperability_standards_in_digital_health.pdf
[26] https://webgate.ec.europa.eu/fpfis/wikis/pages/viewpage.action?pageId=912786435#CodeSystemsintheMVC(MasterValueSetsCatalogue)-BackgroundinformationontheCodeSystemsusedintheMVC. (need a registered account for access)
[27] https://health.ec.europa.eu/publications/ehn-guideline-patient-summary-release-notes_en
[28] https://www.ema.europa.eu/en/human-regulatory/overview/data-medicines-iso-idmp-standards-overview

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

- SNOMED CT, of which the European Commission currently contributes 60% of the annual base licence fee. For research purposes, no licence is needed.

For individual member states, it should be evaluated whether mappings should be developed to comply with the recommended standards, as was mentioned above, as well as evaluating which standards support data capture in and/or translation into the local language (e.g. SNOMED CT and ICD-11 supports translation and language preferences[29])

The interoperability framework should conform with as well as drive development of the Global Alliance for Genomics and Health (GA4GH) standards, in particular when it comes to APIs, data use conditions (ADA-M, Data Use Ontology), and phenopackets as standardised object to share phenotype data (https://github.com/phenopackets). Beside standards, application specific ontologies also exist, that reuse definitions and terms from existing ones and add specific missing elements, like:

- For the human genome: FAIRgenomes[30]
- A semantic version of the phenopackets object (https://github.com/LUMC-BioSemantics/phenopackets-rdf-schema) that incorporates the ontology standards mentioned above and increases interoperability with other FAIR resources.

Besides being able to link data elements to a terminology to make each data element meaningful and machine processable the data elements often are transformed into a common format (data architecture model), as well as a common representation (terminologies, vocabularies, coding schemes), like the OHDSI OMOP CDM as listed in Appendix I on the data architecture tab.

4.1.4 Data quality assurance

Successful clinical decisions, as well as clinical research, require high-quality data. High-quality data means data that represents its underlying real-world phenomena correctly. To achieve high data quality and sustain it, organisations must implement data quality assurance procedures. The quality of data becomes more prominent since the effectiveness of Artificial Intelligence (AI)/Machine Learning (ML) directly depends on it.
Data quality assurance is the process of determining and screening anomalies by means of data profiling, removing obsolete information, and data cleaning. Throughout the lifecycle of data, it is at risk of being distorted by the influence of people and other external factors. Thus, in order to protect or sustain the (high) data quality, a data quality assurance strategy is needed that includes governance measures as well as technical interventions/solutions.
The data quality of a dataset boils down to the comparison of the actual state of that dataset compared to the desired state. Together with the stakeholders of 1+MG, the expectations, specifications, and requirements in terms of characteristics or dimensions of the data should be defined, like completeness, consistency, accuracy, timeliness, versioning, accessibility, etc.

4.1.5 Data access

Data should be made accessible at the appropriate levels of authorization given its scope and means of usage, and in accordance with the European Data Protection Regulation. Certain levels of aggregated (non-related) metadata could be made publicly available, not requiring any authentication or authorization, while restricted data obviously requires a certain form of authentication and authorization. B1MG WP2/1+MG WG2 has written a position paper on the

[29] https://confluence.ihtsdotools.org/pages/viewpage.action?pageId=26837136#:~:text=Today%2C%20SNOMED%20CT%20is%20available,being%20done%20by%20member%20countries
[30] https://fairgenomes.org/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

scope of the 1+MG initiative[31] as well as drafted the central elements of the recommended 1+MG data access governance framework for research purposes[32]. Data access for healthcare and policy-making purposes will be addressed separately and subsequently by B1MG WP2/1+MG WG2. The metadata on data access, also called rights metadata, is a type of administrative metadata that essentially describes data access and use conditions. What holds and applies for clinical and phenotype metadata probably also holds and applies for rights metadata and the framework proposed in this document might be used/applied to explore opportunities and barriers to developing, approving, and implementing a standard rights metadata vocabulary too. Preferred ontologies related to data access are DUO[33] (Data Use Ontology) as well as the ODRL (W3C Recommendation[34]) (Open Digital Rights Language), by taking into account the requirements from WG2 The final mapping is an ongoing effort. See also B1MG D3.8 Documented best practices in sharing and linking phenotypic and genetic data —2v0 | Zenodo[35].

4.1.6 Maintenance and versioning

Properly maintaining and caring for data is essential to ensuring that data remains safely accessible and usable for its intended purposes. With respect to the clinical and phenotype metadata we propose that:

- Changes in terminology standards and annual updates should be aligned with the minimum dataset's characteristics.
- Regular updates ensure correct clinical modelling and facilitate multilingual representation.
- Backward compatibility should be ensured for retrospective analyses.
- It is advised to annually identify currently used terminology standards versioning and update the minimum datasets where necessary.
- A data element should be linked to a value set or a specific code from a terminology system, preferably also specifying type and versioning of the ontology as well as date of administration or date of assessment, or if applicable date of deprecation (item is no longer valid or no longer in use).

4.2 A potential tool for in part operationalizing the framework: ART-DECOR

It is proposed to implement and apply the framework in a collaboration tool to harmonise standards and to define content as well as dataset provenance characteristics (using metadata). ART-DECOR, in use in various standardizations in multiple countries and identified as a best practice[36] is a useful tool to specify and maintain datasets and has a quality control cycle. Each minimal dataset should be characterised using metadata, in order to provide general information on the dataset or variable characteristics, i.e. how were the genomes sequenced/generated, what was the data source, as well as whether the dataset contains subject-level data or not (e.g., data use conditions based on the consent of the individual). The specified minimum dataset prototypes should be thoroughly tested with simulated data. A dataset with data access roles and security procedures to verify an individual's identity is needed. A procedure is needed for maintenance after withdrawal of patients' consent. Case Report Form (CRF) metadata can be made accessible in metadata or in a separate dataset.

[31] https://zenodo.org/record/6363119#.YjIQGxD7T0o
[32] https://zenodo.org/record/6363157#.YjIVMhD7T0o
[33] https://bioportal.bioontology.org/ontologies/DUO
[34] https://www.w3.org/TR/odrl-model/
[35] https://doi.org/10.5281/zenodo.7342854
[36] https://doi.org/10.5281/zenodo.734285

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

Cohort and patient record identifiers and time variables are needed to identify the cohort and location from which the information was derived, in order to be able to 1) compile personalised medicine reports on the individual level or country level, and 2) to make use of the longitudinal data for research purposes and identify relative risks of specific clinical outcomes. Additional details may be requested by the analysing party, therefore, contact details are needed. Optional data elements may also be shared and appended to the minimum dataset, in advance; metadata for these variables should be requested in the dataset maintained in ART-DECOR.
ART-DECOR can be used to model functional specifications that are agnostic from technical communication standards. Because of this approach these specifications can be implemented in a wider array of use cases, in one example in both HL7v3, FHIR STU3 and FHIR R4. ART-DECOR can also be used as a modelling tool towards implementers and supports the following technical standards: HL7v2, HL7v3, and FHIR (currently DSTU2, STU3 and R4).
ART-DECOR supports the following FHIR artefacts:

- dataset or transaction as FHIR Logical Model
- transaction as FHIR Questionnaire
- value set as FHIR ValueSets
- OID Registry info as FHIR NamingSystem

Support for the FHIR Code system and FHIR profile is planned, but not currently supported.

4.3 Re-use of developments in cross-border health data exchange

For several years now, the EU Commission, in collaboration with the member states, has been actively working to support cross-border healthcare, especially regarding e-health. This is intended to promote free movement in the internal market and create opportunities for innovation within the EU. As stated before, the eHealth Network has drawn up rules and specifications for how the data exchange takes place in the different areas and has developed requirements to be implemented in the participating countries to achieve sufficient interoperability (i.e. the Refined eHealth European Interoperability Framework (ReEIF[37])).
Since March 2021, two cross-border services[38], the Patient Summary (based on ISO 27269:2021[39] Health informatics — International Patient Summary) and ePrescription, are operational. The two services will make it easier for EU-citizens/patients to seek care, including collecting medicines in another country. In the near future, based on this infrastructure, services such as medical imaging, discharge letters, laboratory results, and other health data exchanges, will be added.

The implemented and still to be developed services apply semantic unification via standardisation of Healthcare information using Health and Care Information Models (HCIMs). The exchange of information has been defined by HL7 both for CDA and FHIR (Fast Healthcare Interoperability Resources[40]). FHIR is gaining momentum for its ease of implementation via API (Advanced Programming Interface) technology. Two HL7 implementation guides for the international patient summary for CDA[41] and FHIR[42], are available as practical examples of derived implementations conformant with the standard

Although this exchange of information is per patient, it is of interest to make this information available for secondary use via specifically designed connectors like OMOPonFHIR[43] and FHIR to

[37] https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5b56dffdc&appId=PPGMS
[38] https://ec.europa.eu/health/ehealth-digital-health-and-care/electronic-cross-border-health-services_en
[39] https://www.iso.org/standard/79491.html
[40] https://www.hl7.org/fhir/
[41] https://art-decor.org/art-decor/decor-templates--hl7ips-
[42] https://github.com/HL7/fhir-ips
[43] https://omoponfhir.org/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

mCODE[44] while other initiatives like OSIRIS and ICGC-Argo, at this point do not have a specific FHIR connector (API). The first connector supports various mapping combinations between versions of FHIR and OMOP CDM (OMOP: Observational Medical Outcomes Partnership; CDM: common data model). The OMOP CDM allows for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes), and then perform systematic analyses using a library of standard analytic routines that have been written based on the common format (https://www.ohdsi.org/data-standardization/the-common-data-model/). The IMI project European Health Data and Evidence (EHDEN[45]) is currently mapping 140+ data sources in Europe to the OMOP-CDM (EHDEN datapartners[46]). Moreover, EHDEN has been certifying many companies in Europe to support the mapping process (business-directory of EHDEN[47]). There is a large European OHDSI community[48].
The second one is of interest for the WG9-cancer use case, as part of the minimal dataset as being defined is based on mCODE. mCODE (short for Minimal Common Oncology Data Elements) is an initiative intended to assemble a core set of structured data elements for oncology electronic health records.
By taking advantage of semantic unification and exchange of this data/information within 1+MG, we could focus on those clinical and phenotype items that are not yet part of it.

5. Results

This document describes the proposed and still evolving (work continues via 1+MG WG3 as well as GDI) phenotypic and clinical metadata framework.

6. Discussion

This document forms the base for the 1+MG phenotypic and clinical metadata framework, guiding member states in maturing their semantic unification of phenotypic and clinical metadata. The framework will continue to evolve with input from all relevant stakeholders, optimising recommendations on standards and mappings to apply as well as trying to create an operational environment to apply (parts of) the framework. A joined 1+MG WG9 (cancer use case) & WG3 paper has been prepared that describes the process followed to arrive at the minimal (meta)dataset for cancer which has been and can be used as an example for other 1+MG WGs (like WG11 currently does) that are still working on defining their relevant minimal phenotypic (meta)datasets. This document is part of the 1+MG framework and will be maintained within the 1+MG initiative as well as used in the implementation of the GDI.

[44] https://github.com/HL7/fhir-mCODE-ig
[45] http://www.ehden.eu/
[46] https://www.ehden.eu/datapartners/
[47] https://www.ehden.eu/business-directory/
[48] https://www.ohdsi-europe.org/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

7. Conclusions

This document provides an updated (third) version of the phenotypic and clinical metadata framework to support 1+MG in obtaining semantic interoperability, facilitating sharing and linking of phenotypic and clinical metadata and genetic metadata between the member states.

8. Next steps

In the upcoming version the framework will evolve further and also will be linked to recent developments with respect to the EHDS. The advice on which standards are recommended will be extended as well as evaluated whether mappings should be developed to comply with the recommended standards taking into account guidelines of the eHN Subgroup on Semantics. Together with the working groups (1+MG WG8,9,10,11 and 12), the definition of the minimal datasets per use case will be finalised as well as being compared to each other and harmonised. They will be part of the 1+MG Framework[49] as part of the section on Data models, standards & ontologies[50] and handed over to the GDI project for adoption, implementation (including governance). We will also proceed with the PoC with ART-DECOR to test operationalizing (parts of) the framework by, together with 1+MG WG9 and external cancer experts, collaborating on optimising the minimal dataset for the cancer use case.

9. Impact

The phenotypic and clinical metadata framework will guide and advise member states in maturing their semantic unification of phenotypic and clinical metadata.

[49] https://framework.onemilliongenomes.eu/
[50] https://framework.onemilliongenomes.eu/data-models-ontologies

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

Appendix I - Overview of standards: ontologies and data architectures

This appendix describes common standards in the field of coding medical data and how they are interrelated or mapped, as well as provides an overview of various data architectures.

Overview of standards[51]

[51] https://zenodo.org/records/10058527

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724