Beyond One Million Genomes

D3.4

Phenotypic and clinical metadata framework – 1v0

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	D3.4 - Phenotypic and clinical metadata framework - 1v0		
Contractual delivery date	31/05/2021	Actual delivery date	23/05/2022
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), Flávio Soares (1+MG WG3, PT). Ulrika Hermansson (1+MG WG3, SE), Alfonso Valencia (1+MG WG3, SP),		

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

	Laima Ambrozaityte (1+MG WG3, LT), 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)
Contributors	Regina Becker (UNILU, B1MG WP2 & 1+MG WG2, LU) Pablo Serrano (Hospital 12 Octubre, Spanish 1+MG mirror group, SP)
Acknowledgements (not grant participants)	
Deliverable type	Report
Dissemination level	Public

Document History

Date	Mvm	Who	Description
13/10/2020	0v1	Jeroen Beliën (VUMC)	Initial draft created based on presentation on ART-DECOR as a potential tool/framework for the 1+MG project and especially 1+MG-WG3. Team members invited to co-write the first version.
15/04/2022	0v2	Jeroen Beliën (VUMC)	Version circulated to WP & comments addressed.
21/04/2022	0v3	Nikki Coutts (ELIXIR Hub)	Version circulated to B1MG-OG, B1MG-GB & Stakeholders for feedback.
20/05/2022	0v4	Jeroen Beliën (VUMC) Harmke Groot (Nictiz)	B1MG-OG, B1MG-GB & Stakeholder comments addressed.
23/5/2022	1v0	Nikki Coutts (ELIXIR Hub)	Finalised version uploaded to the EC Portal

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

Table of Contents

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 instruments 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 framework document will be part of the entire B1MG framework where other crucial instruments are taken care of by other work packages/working groups like Governance, ELSI and Technical Infrastructure. The B1MG framework document first of all aims at the 1+MG initiative while at the same time could be used as a base for the European Health Data Space[2] (EHDS) and/or genomic data space.

[1] https://docs.google.com/document/d/1GOvcR3l3t8T4cJLDVx7kVxbYa9F2oo7deQveMG6Og6c/edit
[2] 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

The current document is the first version of the 1+MG phenotypic and clinical metadata framework. This document will be updated continuously and updated versions will be published as deliverable D3.5 and D3.6. This deliverable contributes 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 phenotypic data, such as environmental data, information in medical records 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 healthcare data both for primary as well as secondary use. It ensures that the defined medical concepts are unambiguous and that the medical specialists involved have the same clinical interpretation of the exchanged medical concepts, as well as for those reusing the data. Harmonising and standardising coded healthcare 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 care, administration), without loss of "meaning" *[reference: Eenheid van Taal in de Nederlandse zorg Van eenduidige informatie-uitwisseling tot hulpmiddel voor betere zorg. C.H. van Gool et al. RIVM Rapport 2018-0081]* http://dx.doi.org/10.21945/RIVM-2018-0081 *;HIMMS* https://www.himss.org/sites/hde/files/d7/FileDownloads/HIMSS%20Interoperability%20Definition%20FINAL.pdf *],* in other words 'what is sent is what is understood'.

Moreover, this document will serve as the semantic interoperability framework component as part of a domain specific interoperability framework 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

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

the Domain specific Interoperability Framework (DIFs). It remains to be investigated whether or not this concerns being part of the existing DIF, the refined eHealth European Interoperability Framework (ReEIF[3]), or whether it will be a specific 1+MG DIF that might or might not have overlap with the ReEIF.

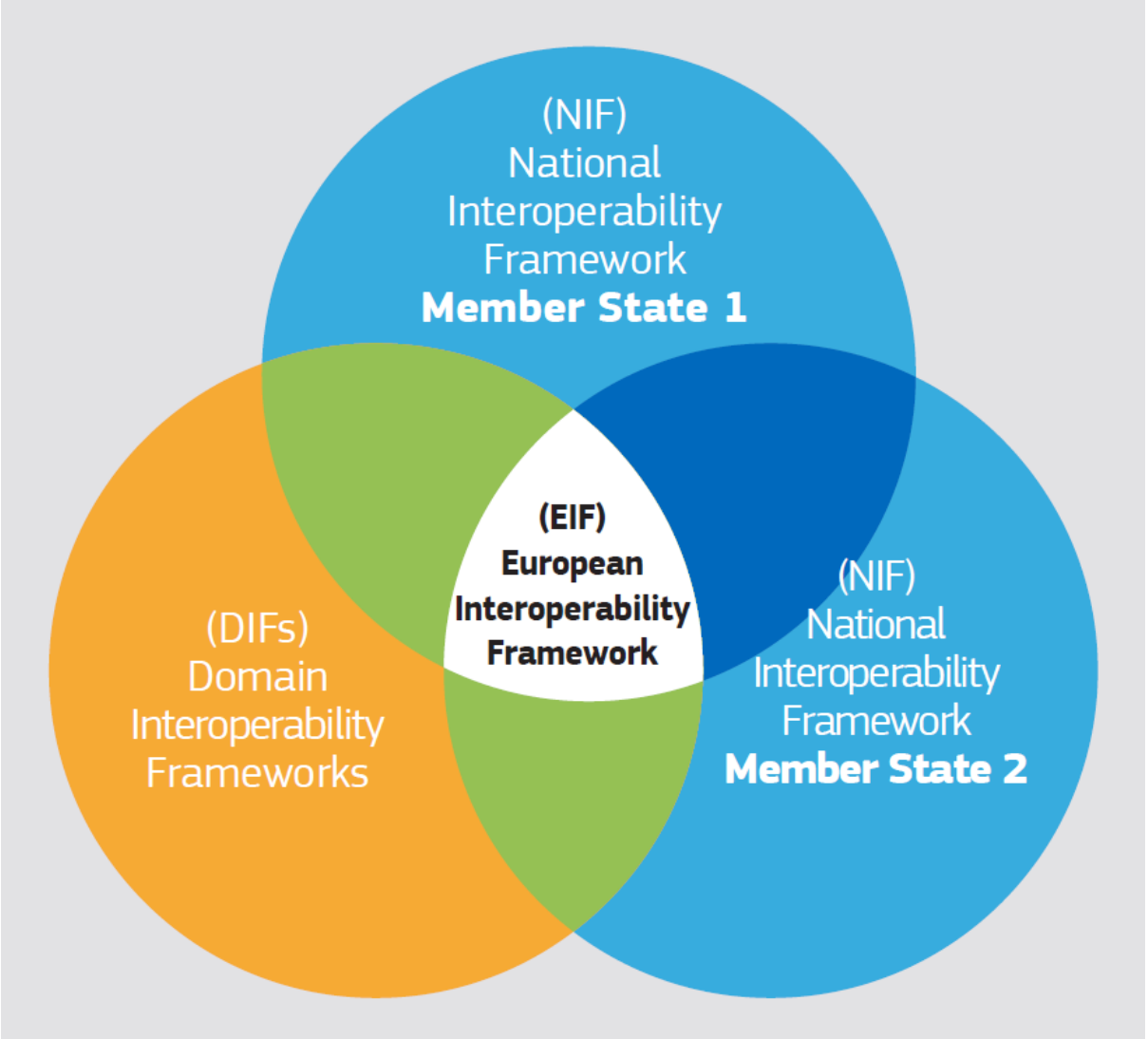

Figure 1: The relationship between EIF, NIFs and DIFs as presented in https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf

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

[3]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

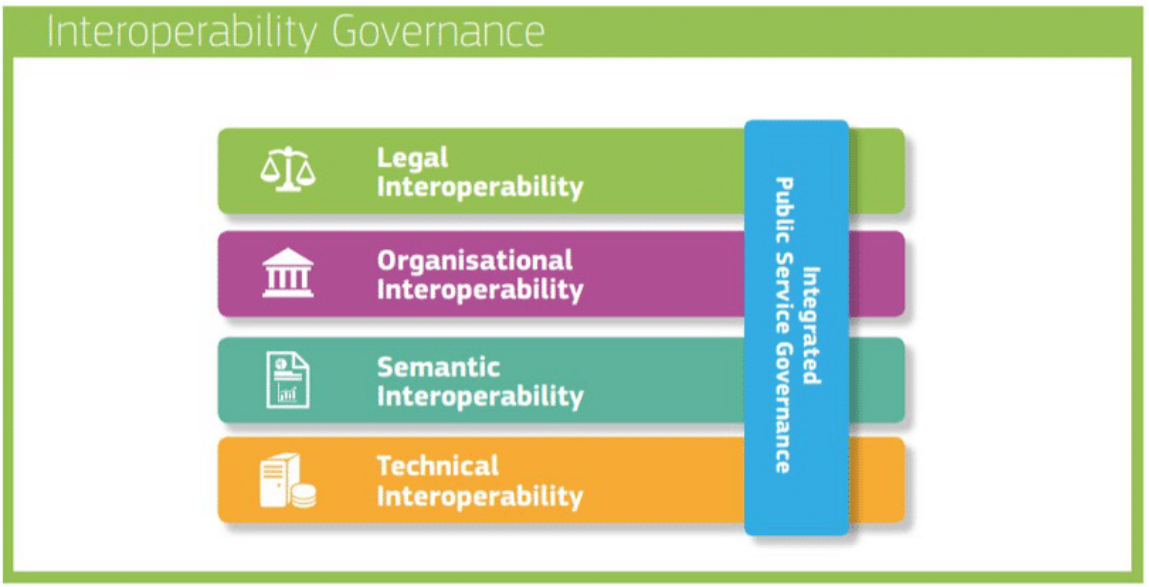

Figure 2: The new European Interoperability Framework model which includes the semantic interoperability model covered by this framework document.

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 (cancer, rare diseases, diagnoses, laboratory data). 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: table XX** Documented best practices in sharing and linking phenotypic and genetic data - live working document[4]).

Besides ultimately aiming for semantic interoperability, identifying and accessing relevant existing datasets and/or agreeing to a minimal 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[5] to understand what existing genomic datasets and corresponding phenotypic/clinical information are effectively available for participation in the 1+MG. The survey results will help identify the challenges and bottlenecks for sharing, as well as making recommendations for design of a European framework for sharing genomic and associated clinical data (this framework document). Initial results show that most genomic datasets have linked clinical data, but that further detailed analysis is needed 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 Accessible Genome Dashboard (B1MG new task 3.6).

[4] https://docs.google.com/document/d/1GOvcR3l3t8T4cJLDVx7kVxbYa9F2oo7deQveMG6Og6c/edit
[5] 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

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.

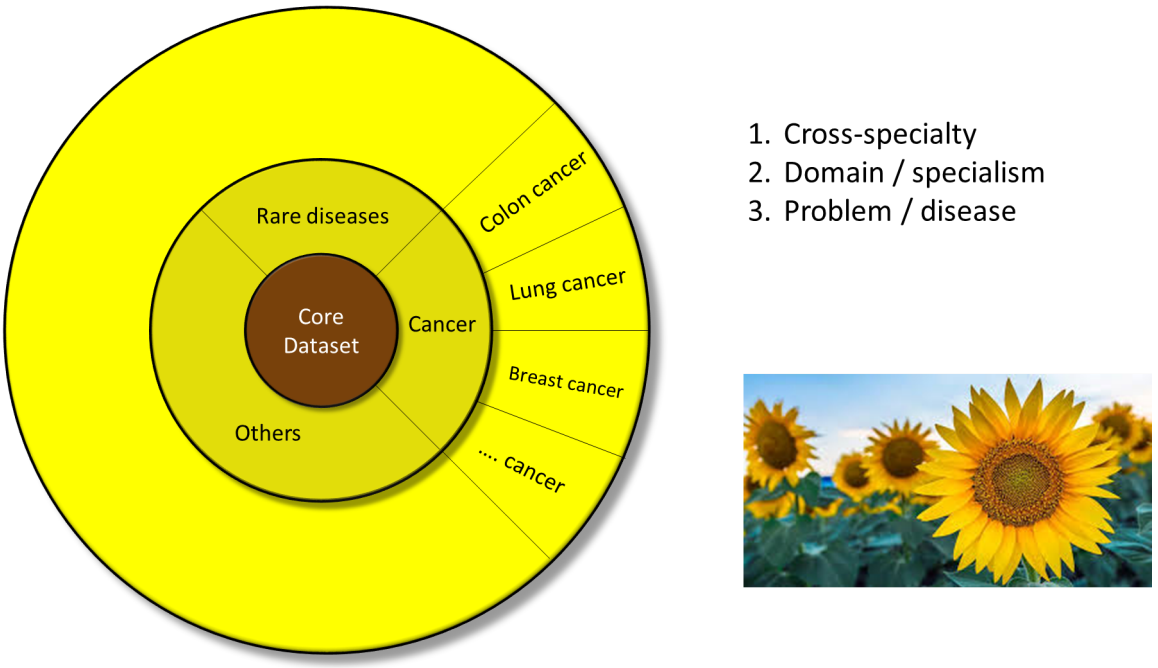

Figure 3: Multi-level standardisation. The core minimal dataset contains cross-specialty mandatory fields, the next layer inherits the mandatory fields of the core and adds mandatory fields specific for a domain or specialism. The outer layer (petals of sunflower) adds the problem/disease specific mandatory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

Guidance is needed in order to reach interoperability. Which standard is prioritised in which context? See for example the recent scientific report in *Health Technology* by de Mello et al. which describes SNOMED CT as a "valuable medical dictionary bringing a clinically validated, semantically rich, controlled vocabulary." SNOMED CT facilitates evolutionary growth in expressivity to meet emerging requirements. [ref: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8791650/pdf/12553_2022_Article_639.pdf] Therefore, it is important to define a preferred clinical terminology for common use, and in case of specific needs, which standards are recommended.

As part of the next versions of this framework document (deliverables 3.5 and 3.6), 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 will aim to provide:

- an overview on relevant code systems (dictionaries) and their characteristics, including details on which systems are mapped to others, and where their domain of interest overlaps. The initial version of this Table XXX has already been added as Appendix X.
- advice on which standards are of preference or mandatory.

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

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 standards. The steps outlined below describe the steps that can be taken to establish such a set of standards.

Measure the current state of maturity
The MLM (Maturity level model) 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 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 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[6]), 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 should be contacted to further document the use case.

[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

- Use case harmonisation

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 standards

If the use case cannot be mapped to existing standards, it can be applicable to create 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 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 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 Appendix X)

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

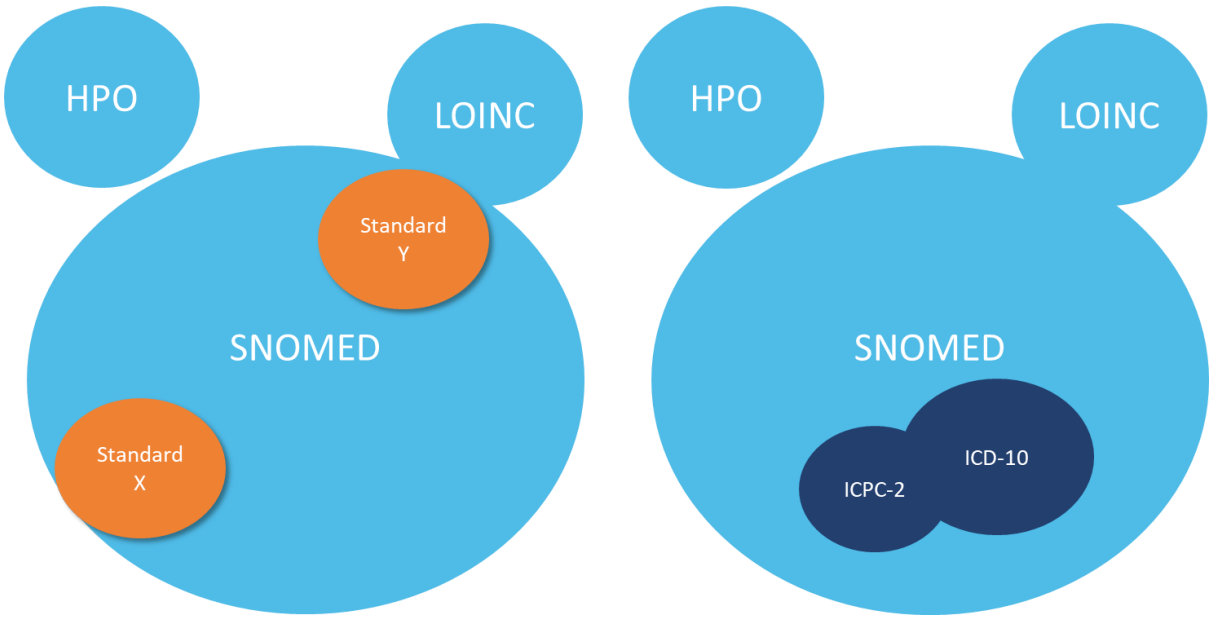

Figure 4: Core terminology overview on (inter)national level: standardisation and mapping

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[7] - *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
- User of the data standard: implements standards into applications. For example, an application vendor
- 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.

[7] 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

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, 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.

Potential frameworks
ART-DECOR, in use in various standardizations in multiple countries and identified as a best practice[8]. For more information see paragraph 4.2 A potential tool for in part operationalizing the framework: ART-DECOR.

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

The above approach closely follows the Maelstrom data harmonisation guidelines, see https://www.maelstrom-research.org/page/maelstrom-guidelines.

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

[8] https://zenodo.org/record/4819149#.YK_bKZNKj0o

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

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 to identify accessible genomes with, if available, linked, clinical data (https://ec.europa.eu/eusurvey/runner/1plusMG_Survey2020). In addition, potentially relevant minimal datasets (domains: research, industry, existing international networks) were identified and workshops were/will be held with domain experts to define the relevant minimal dataset for specified use cases. These datasets, as well as those (to be) defined by all the use-case WGs of 1+MG, were or will be shared with relevant WPs/WGs. The targeted end-user of these datasets is a researcher, MD (Medical Doctor) or patient. The process towards developing interoperable minimal datasets with relevant clinical data follows the Maelstrom data harmonisation guidelines[9].

4.1.2 Operationalising genotype and phenotype data

Which phenotype data 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 (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.
For each 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 medicine and research.

4.1.4 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 are evaluating 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 will be published as part of B1MG Deliverable 3.8.

For now, in summary, the following standards/terminologies are preliminary advised by B1MG-WP3/1+MG WG3 experts:

- For cancer: SNOMED CT (preferred) or ICD10-O
- For rare diseases: ORPHAcodes (part of Orphanet)
- For phenotypic abnormalities: HPO
- For common and complex diseases: 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[10] is recommended
- For exposure ascertainment , the following validated questionnaires are recommended (just a snapshot (some examples), this list is extensive; see also ICHOM):

[9] https://www.maelstrom-research.org/page/maelstrom-guidelines
[10] 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

- 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:

- SNOMED CT, of which the European Commission currently contributes 60% of the annual base licence fee.

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 supports translation and language preferences[11])

The interoperability framework should conform with 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[12]
- 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.

4.1.5 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.

4.1.6 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

[11] https://confluence.ihtsdotools.org/pages/viewpage.action?pageId=26837136#:~:text=Today%2C%20SNOMED%20CT%20is%20available,being%20done%20by%20member%20countries
[12] 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[13] as well as drafted the central elements of the recommended 1+MG data access governance framework for research purposes[14]. 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.

4.1.7 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 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. 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.

[13] https://zenodo.org/record/6363119#.YjIQGxD7T0o
[14] https://zenodo.org/record/6363157#.YjIVMhD7T0o

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

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[15])).
Since March 2021, two cross-border services[16], the 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, see also https://www.hl7.org/fhir/). FHIR is gaining momentum for its ease of implementation via API (Advanced Programming Interface) technology.
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[17] and FHIR to mCODE[18]. 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 second one might be of interest for WG9-cancer use case, as part of the minimal dataset as currently 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 phenotypic and clinical metadata framework.

[15] https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5b56dffdc&appId=PPGMS
[16] https://ec.europa.eu/health/ehealth-digital-health-and-care/electronic-cross-border-health-services_en
[17] https://omoponfhir.org/
[18] https://github.com/HL7/fhir-mCODE-ig

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

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 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.

7. Conclusions

This document provides the first 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 versions 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. 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. We will also initialise a 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.

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

Appendix X - Initial version of Table XXX

Table XXX. Generic overview of common terminology standards and their characteristics

Standard	SNOMED	LOINC	Orphanet	ICD-10	ICD-O	HPO	UCUM	IDMP	TNM	...		
Owner	IHTSDO	Regenstrief Institute										
Health care domain	primary care											
	secondary care											
	tertiary care											
Scope		Standardising lab data (lab requests, lab reports, clinical observables). Chemistry, haematology, serology, microbiology, toxicology, parasitology and virology are domains within the LOINC LAB domain.										
Hierarchy	polyhierarchy											
Code characteristics	Preferred term and synonyms											
code example	74770008 |Exploratory laparotomy (procedure)|											
Release management	twice a year											
browser	https://browser.ihtsdotools.org/?	https://loinc.org/downloads/; https://decor.nictiz.nl/art-decor/loinc										

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

mappings	**Mappings with ICD-10, ICD-O, GMDN, Meddra, Orphanet, CPT**	**LOINC and SNOMED (2-way)**										
mapping details		**1. LOINC Parts that are used in the Cooperative Areas and SNOMED CT Concepts that are used in the "Observables Model" will be mapped to one another to the extent necessary to enable convergence towards a common semantic foundation. 2. Existing SNOMED CT Concepts that are subtypes of Observable Entity or Evaluation Procedure (and fall with the scope of Cooperative Areas) will be mapped to LOINC Terms. 3. LOINC Terms that are not already represented by SNOMED CT Concepts will be associated with post-coordinated expressions. 4. Where appropriate, LOINC Terms that represent observables with ordinal or nominal answer values will be mapped to SNOMED CT Concept names and codes in the LOINC Answer record.**										
Links to mappings	**https://www.snomed.org/snomed-ct/Use-SNOMED-CT/maps**	**https://confluence.ihtsdotools.org/display/DOCLOINC/Using+LOINC+with+SNOMED+CT; https://loinc.org/collaboration/snomed-international/**										
B1MG evaluation	**prioritised medical terminology, because of its rich nature, high adoption grade in EU (although not yet implemented on a structural basis)**											
Results from reviewers of the minimal dataset workgroups:												
Results from national endorsement by recognized institutio	**https://assess-ct.eu/final-brochure/:**											

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

ns (evaluated by each individual country):												

Table XXX describes common standards in the field of coding medical data and how they are interrelated or mapped.

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