Common Infrastructure for National Cohorts in Europe, Canada, and Africa
- CINECA -

D3.1 - Cohort minimal metadata model

Work Package:	WP3 - Cohort Level metadata Representation
Lead Beneficiary:	European Molecular Biology Laboratory
WP Leaders:	Fiona Brinkman (SFU), Melanie Courtot (EMBL-EBI)
Contributing Partner(s):	SFU, UCT, SIB, HES-SO, EMBL-EBI, SickKids, UHN
Contractual Delivery Date:	31st December, 2020
Actual Delivery Date:	17th December, 2020
Authors of this Deliverable:	Vivian Jin, Fiona Brinkman.
Contributors:	Melanie Courtot, Isuru Liyanage, Gurinder Gosal
Reviewed by:	Jonathan Dursi (SickKids/UHN), Helen Parkinson (EMBL-EBI)
Approved by:	Thomas Keane (EMBL-EBI)
Dissemination Level:	Public
Type of Deliverable:	Other
Grant agreement:	No. 825775 Horizon 2020 (H2020-SC1-BHC-2018-2020)
Type of action:	RIA
Start Date:	1 Jan 2019
Duration:	48 months

Table of contents:

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

1. Executive Summary

To support human cohort genomic and other “omic” data discovery and analysis across jurisdictions, basic data such as cohort participant age, sex etc (termed “minimal metadata”) needs to be harmonised. Developing a key “minimal metadata model” of the basic set of attributes that should be recorded with all cohorts is critical to aid initial querying across jurisdictions for suitable dataset discovery. We describe here the creation of a minimal metadata model, the specific methods used to create the minimal metadata model, and this model’s utility and impact. A first version of the metadata model was built based on a review of Maelstrom research data standards and a manual survey of cohort data dictionaries, which identified and incorporated overlapping core variables across CINECA cohorts. The model was then converted to Genomics Cohorts Knowledge Ontology (GECKO) format and further expanded with additional terms. The model was extensively reviewed and has been utilised successfully for demonstrating federated querying of CINECA cohorts at the last CINECA annual general meeting.

To aid development of subsequent methods for data exchange, we went beyond the initial objectives for this deliverable and constructed synthetic datasets for select cohorts, that are based on this minimal metadata model and cohort data. Such datasets enable downstream application development with less concerns about data identifiability and privacy associated with real data. These synthetic data resources are proving to be of high interest by other EuCan projects, in addition to being of use in facilitating appropriate tool development within other CINECA work packages.

While the COVID-19 pandemic has had an impact on many CINECA project participants, particularly some members of WP3 directly involved in the pandemic response, we are pleased that these resources - both the minimal metadata model and synthetic datasets - have been created and are proving to be useful. This work is being written up in a publication that will also describe best practices for such development, and commentary on benefits of such resources. The minimal metadata model is being made broadly available to aid any project or projects, including those outside of CINECA interested in facilitating cross-jurisdictional data discovery and analysis.

2. Project objectives

WP3 Task 3.1 objectives:

1. To define the project’s metadata representation needs
2. To define a cross cohort minimal metadata model
3. To deliver best practice for cross cohort metadata representation
4. To populate the minimal metadata model for cohort metadata

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

3. Detailed report on the deliverable

3.1 Background

The goal of CINECA is to enable federated queries and analyses of the varying and wide-ranging datasets from the 10 CINECA cohorts. The role of the minimal metadata model is to be an agreed upon and machine readable standard so that the wide ranging and differing variables from each cohort's dataset can map to standardised and ontologised variables, and thus be harmonised to enable federated querying. The model must be flexible and its variables must be sufficient to cover (1) common CINECA cohort data, (2) any requirements for CINECA use cases, and (3) key cohort metadata commonly collected by renowned data catalogues. To test the usefulness of the model, several synthetic datasets have been created to validate the minimal metadata model and demonstrate ability to harmonise and query data from the CINECA cohorts. Real cohort datasets are not easily usable as there are strict data governance policies around sharing of any sensitive data. Thus, synthetic datasets were created to be used in place of real datasets. The synthetic datasets are de-identified and obscure sensitive data to allow for public use, but they must also reliably represent the real data types of the original datasets.

3.2 Work Done

3.2.1 Creation of minimal metadata model

The first draft of a minimal metadata model was created by collecting and structuring key variables from the CINECA cohorts. To begin the process of choosing variables for the model, variables from the 10 CINECA cohorts were gathered either by web scraping publicly available cohort data, and/or by directly obtaining data dictionaries from cohort contacts. Each variable was grouped into a broad category; examples of categories include socio-demographic and economic characteristics, diseases, and lifestyle and behaviours. Maelstrom Research data standards[1] were used as a basis when developing the model - the Maelstrom Research group is known for their focus on data harmonisation methodology and has developed a standard approach to documenting and disseminating epidemiological study metadata. The Maelstrom Research Catalogue[2] is a collection of comprehensive study metadata from numerous collaborative and international projects. Four CINECA cohorts have already been integrated into the Maelstrom Catalogue and thus, the structure of the metadata model was pragmatically based on Maelstrom's structuring of metadata variables into categories under broad 'Areas of Information'. The majority of categories have been used in the model, and the categories used depend on how many CINECA cohorts have

[1] Fortier, I., Raina, P., Van den Heuvel, E. R., Griffith, L. E., Craig, C., Saliba, M., Doiron, D., Stolk, R. P., Knoppers, B. M., Ferretti, V., Granda, P., & Burton, P. (2017). Maelstrom Research guidelines for rigorous retrospective data harmonization. International journal of epidemiology, 46(1), 103–105. https://doi.org/10.1093/ije/dyw075

[2] https://www.maelstrom-research.org/maelstrom-catalogue

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

collected data for that category. There is much cross project compatibility between CINECA and Maelstrom Catalogue - and by using the same categories, more CINECA cohorts may be easily integrated into the Maelstrom Catalogue in future.

Through categorisation, the variable overlap between cohorts (Appendix 8.1) was determined, and a list of most commonly collected variables was taken to form the basic variable set of the minimal metadata model. Next, a list of use cases from CINECA WP4/5 - Federated Joint Cohort Analysis/ Clinical Applications (Appendix 8.2) was compiled (as these WP are the primary users of the minimal meta data model) and any missing variables were added to the minimal metadata variable set to ensure the model is sufficient to cover all use case requirements. As an example, one CINECA use case is conducting federated eQTL analyses. In order to conduct these analyses, researchers require variables for describing sequencing data and sequencing metadata (e.g. RNAseq data, associated cell/tissue type, genotype data, and available data formats), variables for describing the traits, diseases, and medications associated with genotype, and variables for describing socio-demographic and economic characteristics (e.g. gender, age). It is also necessary to have general cohort metadata variables which describe the population of the cohort (for example cohort participant age, ethnicity, etc).

Lastly, a list of 30 major data catalogues (e.g. BBMRI-ERIC Directory, Maelstrom Research Catalogue) (Appendix 8.3) was compiled and the most commonly collected metadata variables from these catalogues were also added to the minimal metadata model. This was done in an effort to follow best practice for cross cohort metadata representation and to ensure the minimal metadata model is populated for general cohort metadata as well as relevant for CINECA participating cohorts.

After the variable set of the minimal metadata model was determined, the structure of each variable was then further refined into broad categories and subcategories (Figure 1), many of which were based on standard categories (used by the major data catalogues such as Maelstrom), with additional categories created to fulfill specific CINECA use case requirements. For example, the variable for ‘blood’ belongs to the subcategory of ‘sample type’, which falls under the broad category of ‘biosample’ (Figure 2). Then, additional columns were added to include the variable description, expected answer type, ontology label/ID/definition, number of known CINECA cohorts having this variable, and the applicable CINECA use case requirement. The variables were ontologised - each variable was compared to ontology terms and the most suitable definition was assigned. The most terms were taken from the National Cancer Institute Thesaurus ontology (NCIT)[3], as this ontology has terms which encapsulated the majority of the minimal model variables. Usage of ontologies is helpful as they are standardised definitions and having cohort variables associated with ontology labels/IDs allows machine readability - they can be easily converted to machine readable formats such as JSON and therefore can be simply implemented in harmonisation tools or to support queries.

[3] https://bioportal.bioontology.org/ontologies/NCIT

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

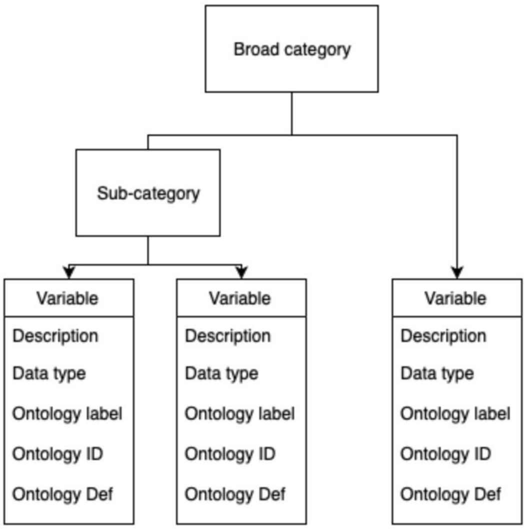


Fig 1. The structuring of minimal metadata model variables into subcategories and categories (generic example).

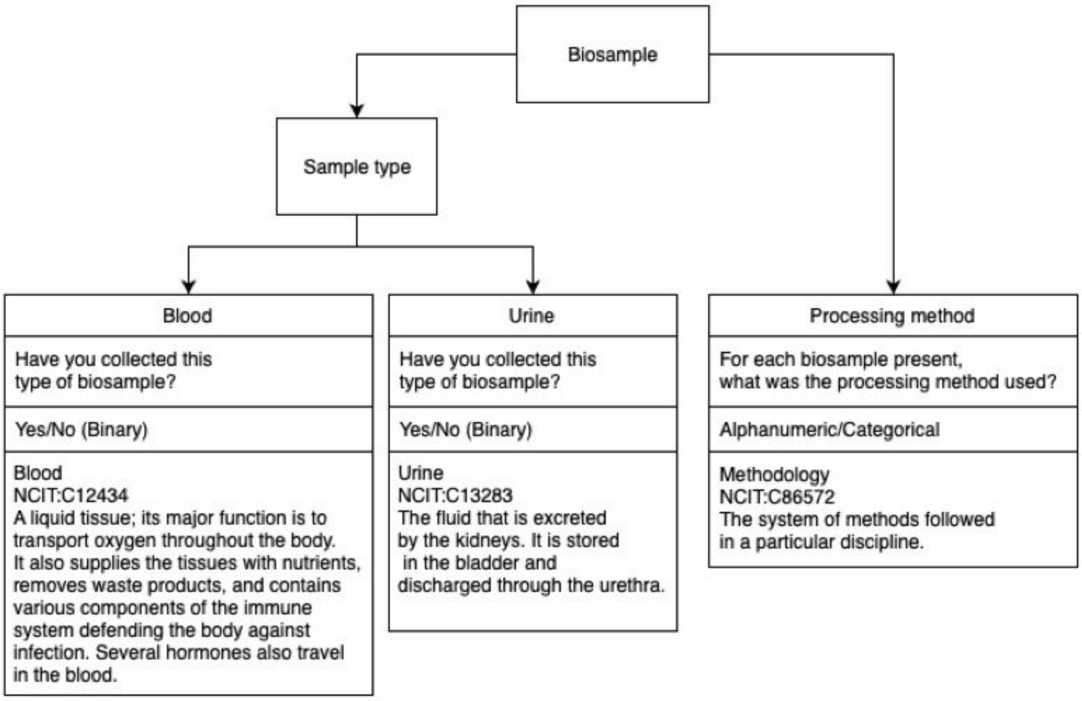

Biosample
Sample type
Blood
Have you collected this type of biosample?
Yes/No (Binary)
Blood
NCIT:C12434
A liquid tissue; its major function is to transport oxygen throughout the body. It also supplies the tissues with nutrients, removes waste products, and contains various components of the immune system defending the body against infection. Several hormones also travel in the blood.
Urine
Have you collected this type of biosample?
Yes/No (Binary)
Urine
NCIT:C13283
The fluid that is excreted by the kidneys. It is stored in the bladder and discharged through the urethra.
Processing method
For each biosample present, what was the processing method used?
Alphanumeric/Categorical
Methodology
NCIT:C86572
The system of methods followed in a particular discipline.

Fig 2. A more concrete example depicting the structuring of variables ‘blood’, ‘urine’, and ‘processing method’ in the broad category of ‘biosample’

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

In summary, the variables selected for the model either are (1) commonly collected variables among the 10 CINECA cohorts (5 or more cohorts), or (2) variables needed to fulfill CINECA use case requirements provided by WP4/5, or (3) typical metadata variables collected by major catalogues which fit common data categories. These variables are tabulated in the Base material for the cohort minimal metadata model (Appendix 8.4). There are currently 80 variables included in the minimal metadata model, structured into 7 broad categories. Approximately 50 variables are commonly collected by the majority of CINECA cohorts, ~50 variables are use case requirements, and ~70 variables have been ontologised. Note these numbers are subject to change as the minimal metadata model is further refined and expanded. These variables should reflect what would be an ideal minimal set of data that should be collected, based on major catalogues, cohort comparisons, and analysis of major use-case requirements.

3.2.2 Initial applications of the minimal metadata model

3.2.2.1 Model validation and demonstration of federated querying using the model

The model has been validated within CINECA through a demonstration of federated querying of 3-4 CINECA cohorts by WP1 (Federated Data Discovery and Querying) presented at the CINECA March 2020 annual general assembly. Synthetic datasets having basic variables were created for demo use (see further description of initial and subsequent synthetic data below). These datasets were mapped to the minimal metadata model variables and the mappings converted into JSON format so that it could be queried programmatically by WP1. The queries demonstrated by WP1[4] ranged from simple questions such as “which cohorts have the data that is necessary for this use case?” with more complicated questions possible with the synthetic dataset such as “in this dataset, what fraction of patients with disease X and variants in gene Y have outcome Z?”. The minimal metadata model was shown to be robust as a mapping standard so that these queries can be made simultaneously to all datasets via the model. This initial implementation demonstrated feasibility of the pipeline and helped inform other WPs of the practicalities of data retrieval in CINECA. Further evaluation is also possible through our cross-H2020-project “EU-CAN Harmonize” meetings so this model may be more broadly used.

3.2.2.2 Generation of synthetic data using the minimal metadata model, as a resource for development of federated data discovery and analysis

The decision was taken to develop synthetic datasets as a valuable resource for further development of federated data discovery and analysis applications. Though this was not a planned objective it enabled non secured testing and therefore more rapid iteration over the model during development. Therefore these synthetic datasets were expanded and refined for more advanced testing of the CINECA infrastructure, and consequently, the development of synthetic datasets for select Canadian, European and African cohorts was initiated. The

[4] Dursi, J., Rambla de Argila, J., de la Torre, S., Tanzer, R., Naderi, N., Mbiyavanga, M., & Agarwal, S. (2020). CINECA_Discovery Service Catalogue_D1.1. Zenodo. https://zenodo.org/record/3908397

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

synthetic datasets must represent the real data types without revealing any sensitive information, the variables chosen must represent the diversity of CINECA's cohorts, and techniques are used to achieve anonymity and strong privacy guarantees. These datasets must also sufficiently reflect the minimal metadata model, be mappable to the model, cover CINECA use cases, and be diverse in displaying variables representative of the respective cohorts. This synthetic data will be used by WP1 for further development of searching/querying and authorisation through WP2 (Interoperable Authentication and Authorisation Infrastructure), and by WP4/5 (Federated Joint Cohort Analysis and Clinical Applications) to carry out their work on CINECA use cases.

The cohorts that have created synthetic datasets are CHILD[5], CoLaus[6], H3Africa[7], and UK Biobank[8] - these were identified as exemplar cohorts which have rich and diverse datasets, a variety of data types, and are representative of the 3 continents participating in the CINECA project. Sourcing real data, each cohort chose a subset of variables that either mapped to the minimal metadata model, or were relevant for conducting future COVID-19 research. Using more advanced data synthesiser tools Tofu[9] and Data Synthesizer[10], synthetic data was generated from subsets of real de-identified data. Using these specialised tools helped to optimise the synthetic data generation process, and the Data Synthesizer has the added functionality of allowing users to choose the degree to which the synthetic data adheres to the original imported dataset statistically. For example, the generated data could be entirely randomised, or conserve basic statistical measures such as median/mode, or conserve correlations between variables using Bayesian networks. The synthetic datasets were further refined and linked to open source human genotype data from 1000genomes[11]. This enabled querying of genotypes and gene variant information with a freely available dataset that are required for development and testing of many CINECA use cases.

Synthetic data creation as reported by each contributor:

1. UCT: A comprehensive version of the synthetic data using a modified version of Tofu with fields, encodings and stats based on the H3Africa Core phenotypes was delivered. The overlap of the synthetic data with the CINECA minimal metadata model is documented together with ontology mapping. haring and usage guidelines delivered together with other CINECA synthetic data generators ensures responsible use of this synthetic data and that no biological inference is made based on this data. UCT is deploying a Beacon v2 instance that will be part of the CINECA Beacon network. This H3Africa Beacon will also host this synthetic data for discovery as well real H3Africa cohort data.

[5] https://childstudy.ca
[6] https://www.colaus-psycolaus.ch/professionals/colaus
[7] https://h3africa.org
[8] https://www.ukbiobank.ac.uk
[9] https://github.com/spiros/tofu
[10] https://github.com/DataResponsibly/DataSynthesizer
[11] https://www.internationalgenome.org

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

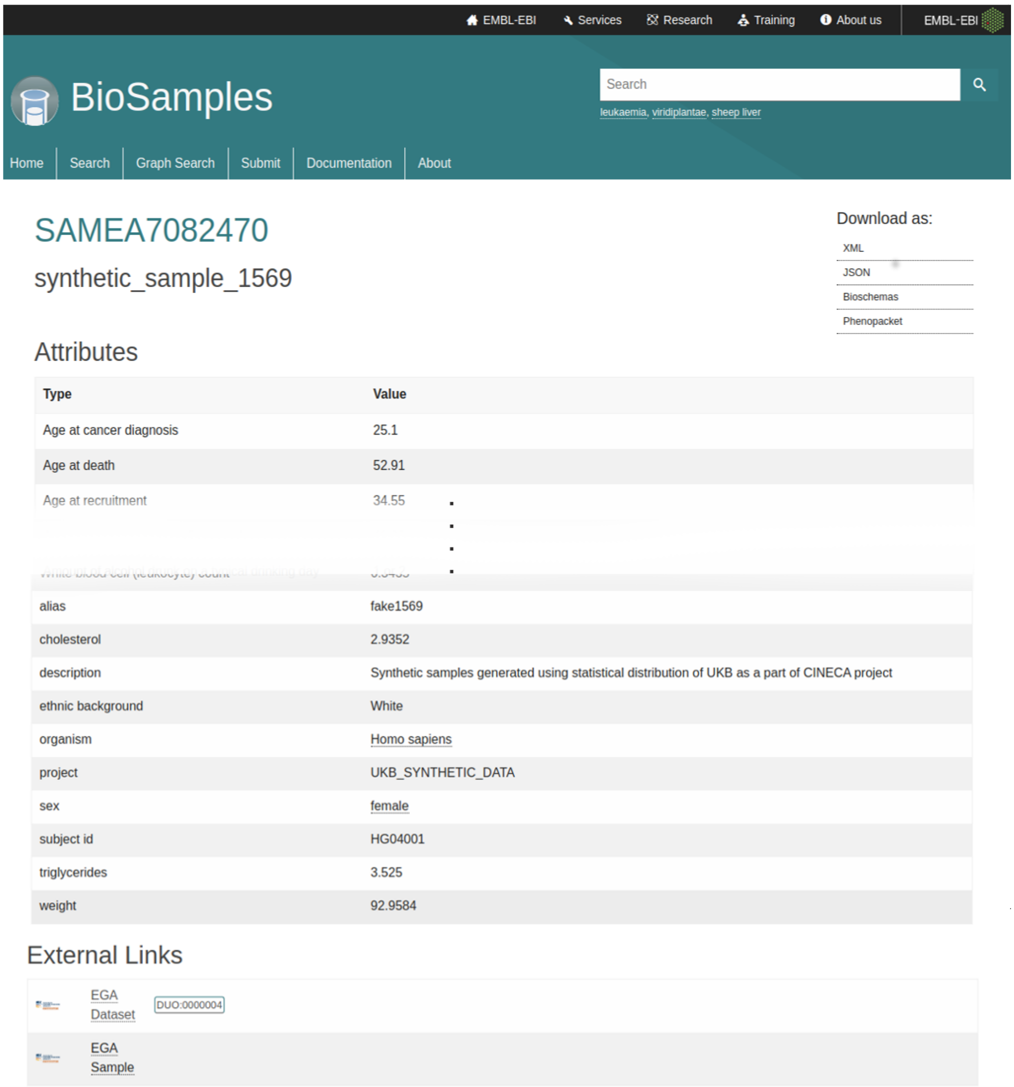
2. HES-SO - We have generated synthetic data for CoLaus/PsyCoLaus using the Data Synthesizer tool. This tool is specifically designed for privacy-preserving datasets.[12] An initial set of 21 data attributes out of 191 were selected based on their relevance to the project and availability in the minimal metadata model. These attributes encode demographics, phenotypic, life style, and clinical information, such as diagnoses, medication, weight, age, smoking status, etc. We focused only on integer, float and semi-structured string data types as they enable easy generation of privacy-preserving synthetic data. Data available in the CoLaus/PsyCoLaus synthetic data was then mapped to the minimal metadata model to ensure its coverage. The phenotypic data is randomly generated (not correlated to the original distribution). Due to privacy reasons, for genotypic data, we rely on the 1000 Genome data. Finally, this data was loaded to a Beacon v2 endpoint to be searchable and findable by the project partners and to be used by other work packages.
3. SFU - We have generated synthetic data for CHILD using the Data Synthesizer tool. Around 100 variables were chosen which (1) covered minimal metadata model, (2) covered COVID-specific use cases and (3) exhibited particular variables that are key to the CHILD study. Real data for these variables were imported into the data synthesizer tool and synthetic data was generated for 150 fake subjects. As there are 3 populations in the CHILD study (mother, father, child), there are 4 synthetic datasets in the form of Excel workbooks (each with 150 synthetic subjects): 1 dataset of correlated anthropomorphic variables concerning children, and 3 datasets of uncorrelated variables for children, mothers, and fathers. The CHILD synthetic data was mapped to the minimal metadata model to demonstrate coverage of the model - this mapping will also enable federated querying of the data through Beacon.
4. EMBL-EBI - We have generated a synthetic dataset to allow the use-cases from WP4/5 to demonstrate their tools which link phenotypic data with genetic data. The phenotypic data is available in EBI BioSamples with associated links to genotypic data hosted by EGA[13]. There are a total of 2504 synthetic BioSample entries available with each sample having approximately 60 attributes, covering the intersection between the minimal metadata model and the UKBiobank. The phenotypic data was generated using the Tofu tool, which generates a set of attributes based on the UKBiobank Data Showcase, for a set number of samples with each attribute following the frequency distribution of the attribute within UKBiobank. A majority of attributes are based on the reported distribution of values within the UKBiobank data showcase. Attributes that may contradict each other, such as date of birth and age / date of death, were curated to ensure that these attributes were

[12] https://github.com/DataResponsibly/DataSynthesizer/blob/master/docs/cr-datasynthesizer-privacy.pdf

[13] https://www.ebi.ac.uk/ega/home

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

consistent with each other, and correlated attributes, for example height, weight and BMI, were calculated from the core generated attributes (height and weight). The genetic data were derived from the 1000 Genomes Phase 3 release, and consists of plink, vcf, bed, and ped files for the called variants of the 2504 samples, plus raw data for a subset of these samples in bam, cram, and FASTQ format. The genetic data and phenotypic data are unrelated, except for the gender of the subject, which is consistent. With the genetic data in EGA, users can test authentication and authorisation processes, and streaming protocols such as htsget to obtain the genetic data.



SAMEA7082470

synthetic_sample_1569

Attributes

Type	Value
Age at cancer diagnosis	25.1
Age at death	52.91
Age at recruitment	34.55
⋮	
White blood cell (leukocyte) count	[illegible]
alias	fake1569
cholesterol	2.9352
description	Synthetic samples generated using statistical distribution of UKB as a part of CINECA project
ethnic background	White
organism	Homo sapiens
project	UKB_SYNTHETIC_DATA
sex	female
subject id	HG04001
triglycerides	3.525
weight	92.9584

External Links

Figure 3. Screenshot of UKB synthetic data available in EMBL-EBI BioSamples database

3.3 Next steps

We have completed our tasks for this deliverable, through defining the project's metadata representation needs, and defining this initial cross cohort minimal metadata model. Future

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

work will seek to iteratively add to the minimal metadata model as needed for the duration of the project to address use cases. Additional work has been ongoing to map synthetic dataset variables to minimal metadata variables in order to 1) further examine coverage of minimal metadata model by the synthetic datasets, 2) create an integrated harmonised table of the mapping of variables in different cohorts with minimal metadata model by incorporating the GECKO ontology[14], and 3) enable synthetic data to be hosted on Beacon[15], a platform used by WP1 to carry out federated querying.The minimal model is being mapped to other resources such as CanDIG (via a Beacon) and we are contributing to the expansion of the initial demo to encompass authorisation and authentication through WP2, plus are focusing on specific use cases such as eQTL analysis provided by WP4 and COVID-19 applications. WP3 will continue to work with WP1 to ensure the latest metadata standards are supported by WP1 APIs, queries, and portals. This minimal metadata model and associated synthetic datasets will be key enablers for other CINECA work packages, and synthetic data can be further expanded to meet use case needs.

We aim to publish this model and we also aim to publish a paper outlining the methods/tools used for synthetic data generation, challenges encountered in the construction of synthetic datasets, and the utility of such data as a resource to facilitate trans-jurisdictional data exchange. This landmark paper will aim to provide the first published best practices regarding synthetic data generation, which will hopefully be of wide utility and interest.

4. References

[1] Fortier, I., Raina, P., Van den Heuvel, E. R., Griffith, L. E., Craig, C., Saliba, M., Doiron, D., Stolk, R. P., Knoppers, B. M., Ferretti, V., Granda, P., & Burton, P. (2017). Maelstrom Research guidelines for rigorous retrospective data harmonization. *International journal of epidemiology*, *46*(1), 103–105. https://doi.org/10.1093/ije/dyw075
[2] Maelstrom Research. (2020). Maelstrom Catalogue.
https://www.maelstrom-research.org/maelstrom-catalogue
[3] National Cancer Institute Thesaurus Ontology. (2020).
https://bioportal.bioontology.org/ontologies/NCIT
[4] Dursi, J., Rambla de Argila, J., de la Torre, S., Tanzer, R., Naderi, N., Mbiyavanga, M., & Agarwal, S. (2020). CINECA_Discovery Service Catalogue_D1.1. Zenodo.
https://zenodo.org/record/3908397
[5] CHILD cohort study. (2020). https://childstudy.ca
[6] Colaus Study. (2020). https://www.colaus-psycolaus.ch/professionals/colaus
[7] Human Heredity and Health in Africa (H3Africa). (2020). https://h3africa.org
[8] UK Biobank. (2020). https://www.ukbiobank.ac.uk
[9] Spiros Denaxas. (2020). spiros/tofu: Updated release for DOI (Version v1.1). Zenodo. http://doi.org/10.5281/zenodo.3634604

[14] http://www.obofoundry.org/ontology/gecko.html
[15] https://github.com/EGA-archive/beacon-2.x

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

[10] Howe, B., Stoyanovich, J., Ping, H., Herman, B., and Gee, M. (2017). Synthetic Data for Social Good. arXiv:1710.08874
[11] 1000 Genomes. (2020). https://www.internationalgenome.org
[12] Genomic Cohorts Knowledge Ontology. (2020). http://www.obofoundry.org/ontology/gecko.html
[13] EGA Archive Beacon v2.x (2020). https://github.com/EGA-archive/beacon-2.x
[14] European Genome-phenome Archive. (2020). https://www.ebi.ac.uk/ega/home

5. Abbreviations

AGM	Annual general meeting
CHILD	CHILD cohort study
DUO	Data Use Ontology
eQTL	Expression quantitative trait loci
EUCAN	EUCAN Connect (Europe and Canadian consortium)
GA4GH	The Global Alliance for Genomics and Health
GECKO	Genomics Cohorts Knowledge Ontology
JSON	JavaScript Object Notation
NCIT	National Cancer Institute Thesaurus
OWL	Web Ontology Language
WP	Work Package

6. Work Packages in CINECA

WP1 - Federated Data Discovery and Querying
WP2 - Interoperable Authentication and Authorisation Infrastructure
WP3 - Cohort Level Meta Data Representation
WP4 - Federated Joint Cohort Analysis
WP5 - Healthcare Interoperability and Clinical Applications
WP6 - Outreach, training and dissemination
WP7 - Ethical and legal governance framework for transnational data-sharing
WP8 - Project Management and coordination
WP9 - Ethics requirements

7. Delivery and schedule

The delivery is on time.

This project has received funding from the European Union's Horizon 2020
Research and Innovation Programme under grant agreement No. 825775.

8. Appendices

8.1 CINECA_Maelstrom Overlap_April2020

8.1.1 General Overlap

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

8.1 CINECA_Maelstrom Overlap_April2020

8.1.1 General Overlap

	Number of participants	Location	Longitudinal	Diseases	Gender	Participant description	WGS	WES	RNASeq	Epigenetics	Genotyping
CHILD	3.5k	CA	X	Population based developmental health and disease	M & F	Children (birth to 8 years) and mother/father	X		X	X	X
CARTaGENE	43k	CA	X	Population based cohort	M & F	men and women aged between 40 and 69 years residing in metropolitan areas of Quebec.	X		X		X
CLSA	50k	CA	X	Population based cohort	M & F	canadians aged 45-85					X
H3Africa	75k	SA		Multiple communicable and non-communicable diseases in multiple African countries	M & F		X	X			X
BIOS	4k	NL		Population based cohort	M & F		X		X	X	X
Estonian Biobank	51k	EE	X	Population based cohort	M & F	>= 18 years. Estonians represent 83%, Russians 14%, and other nationalities 3% of all participants.	X	X	X	X	X
CoLaus	6.1k	CH	X	Cardiovascular diseases	M & F	middle-aged			X		X
PsyCoLaus	3.6k	CH	X	Mental disorders	M & F				X		X
EGA	700k	UK+ES		Multiple diseases and healthy cohorts	M & F		X	X	X	X	X
UK Biobank	500k	UK	X	Population cohort and disease; cancer, heart disease, stroke, diabetes, arthritis, osteoporosis, eye disorder, depression and form of dementia	M & F	aged 40-69 years and who lived within ~25 miles of a UK assessment centre	X	X			X

Biosamples:	urine	breast milk	blood	meconium	stool	saliva	nasal swab	viral swab	lipid panel		Other samples:	dust from home
CHILD	x	x	x (venous)	x	x	x	x	x				x
CARTaGENE												
CLSA	x		x (nonfasting)									
H3Africa												
BIOS												
Estonian Biobank			x (venous)									
CoLaus/PsyCoLaus	x		x (venous + fasting)		x				x			
EGA												
UK Biobank	x		x			x						

8.1.2 Maelstrom categories

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

8.1.2 Maelstrom categories

Maelstrom Catalogue: Areas of Information

green = maelstrom catalogue cohorts

orange = filled in with supplementary metadata

grey = these cohorts are comprised of sub-projects with widely varying metadata, or are non-questionnaire based

Socio-demographic and economic characteristics

	Age/birthdate	Sex/gender	Family and household structure	Citizenship and immigrant status	Education	Ethnicity, race and religion	Twin	Residence	Language	Marital/partner status	Birthplace	Labour force and retirement
CHILD	x	x	x		x	x		x	x	x	x	x
CARTaGENE	x	x	x	x	x	x	x	x	x	x	x	x
CLSA	x	x	x	x	x	x		x	x	x	x	x
H3Africa		x										
BIOS		x										
Estonian Biobank			x		x	x	x	x	x		x	x
CoLaus/PsyCoLaus	x	x	x		x	x		x		x	x	
EGA		x (min requirement for submission)										
UK Biobank	x	x	x	x	x	x	x	x			x	x

Lifestyle and behaviours

													Other
	Tobacco	Alcohol	Drugs	Breastfeeding	Sleep	Physical activity	Sexual behaviours and orientation	Transportation	Leisure activities	Nutrition	Personal hygiene	Technological devices	Fatigue
CHILD	x			x	x	x		x		x		x	
CARTaGENE	x	x	x	x	x	x		x		x			
CLSA	x	x	x		x	x	x	x	x	x	x	x	
H3Africa													
BIOS													
Estonian Biobank	x	x			x	x			x	x			
CoLaus/PsyCoLaus	x	x	x		x	x				x (FFQ data)			x
EGA													
UK Biobank	x	x		x	x	x	x	x	x	x		x	

Birth, pregnancy and reproductive health history

	Puberty, menstruation, menopause and andropause	Contraception	Pregnancy, delivery and birth	Fertility and sexual health
CHILD			x	
CARTaGENE	x	x	x	x
CLSA	x		x	
H3Africa				
BIOS				
Estonian Biobank	x	x	x	
CoLaus/PsyCoLaus				
EGA				
UK Biobank	x	x	x	

Perception of health, quality of life, development and functional limitations

						Other
	Perception of health	Quality of life	Life course development	Functional limitations	Use of assistive devices	Places been to
CHILD		x	x	x	x	
CARTaGENE	x			x	x	
CLSA	x	x		x	x	x
H3Africa						
BIOS						
Estonian Biobank	x					
CoLaus/PsyCoLaus	x					
EGA						
UK Biobank	x					

Diseases

													Other			
	Certain infectious and parasitic diseases	Neoplasms	blood and blood-forming organs and certain disorders involving the immune mechanism	Endocrine, nutritional and metabolic diseases	Mental and behavioural disorders	nervous system	eye and adnexa	ear and mastoid process	circulatory system	respiratory system	digestive system	skin and subcutaneous tissue	joint-related	bone-related	reproductive system	prostate
CHILD	x	x	x	x	x	x	x	x	x	x	x	x				
CARTaGENE	x	x	x	x	x	x	x	x	x	x	x	x				
CLSA	x	x	x	x	x	x	x	x	x	x	x					
H3Africa																
BIOS																
Estonian Biobank			x	x												x

CoLaus/PsyCo Laus			x	x	x	x			x	x	x	x	x	x	x
EGA															
UK Biobank		x		x	x	x	x	x	x	x	x	x			

	Symptoms and signs								
	circulatory and respiratory systems	digestive system and abdomen	skin and subcutaneous tissue	nervous and musculoskeletal systems	urinary system	cognition, perception, emotional state and behaviour	speech and voice	General symptoms and signs	Symptoms related to multiple categories
CHILD	x	x	x					x	x
CARTaGENE	x				x				
CLSA	x	x			x	x		x	
H3Africa									
BIOS									
Estonian Biobank									
CoLaus/PsyCo Laus	x								
EGA									
UK Biobank	x	x						x	x

	Medication and supplements			Other
	Medication and supplement intake	Posology and protocol of administration	Other and unspecified pharmacological interventions	Side effects
CHILD	x	x	x	
CARTaGENE	x		x	
CLSA	x			
H3Africa				
BIOS				
Estonian Biobank	x			x
CoLaus/PsyCo Laus	x			
EGA				
UK Biobank	x		x	

	Non-pharmacological interventions							Other	
	Surgical interventions	Radiological interventions	Physical therapy interventions	Cognitive, psychological and sensory interventions	Educational and health promotion interventions	Laboratory diagnosis interventions	Other and unspecified non-pharmacological interventions	Hormone treatment	Chemotherapy
CHILD	x	x				x	x		
CARTaGENE	x	x				x	x		
CLSA	x		x	x	x		x		
H3Africa									
BIOS									
Estonian Biobank	x	x					x	x	x
CoLaus/PsyCo Laus	x								
EGA									
UK Biobank	x	x		x		x	x		

	Health and community care services utilization			
	Visits to health professionals	Hospitalizations	Community and social care	Other health and community care
CHILD	x	x		
CARTaGENE	x			
CLSA	x	x	x	x
H3Africa				
BIOS				
Estonian Biobank	x			
CoLaus/PsyCo Laus	x	x?		
EGA				
UK Biobank	x	x		x

	Death	
	Vital status	Cause of death
CHILD		
CARTaGENE	x	x
CLSA		
H3Africa		
BIOS		
Estonian Biobank	x	x
CoLaus/PsyCo Laus	x?	x
EGA		

UK Biobank	x	x

	Physical measures and assessments										Other											
	Physical characteristics	Anthropometry	Circulation and respiration	Muscles, skeleton and mobility	Sensory and pain	Brain and nerves	Skin and subcutaneous tissue	Speech and voice	Digestion	Reproduction	Allergy skin prick	BMR	Head and Neck exam	Apgar test	Lean mass	Impedance	Whole body composition	Bone densitometry	Polysomnography	Grip strength	Baldness/age of onset	Visual acuity
CHILD		x	x				x				x		x	x								
CARTaGENE	x	x	x	x			x					x				x						
CLSA	x	x	x	x	x		x								x							
H3Africa																						
BIOS																						
Estonian Biobank	x	x	x															x				x
CoLaus/PsyCoLaus		x	x		x (pain questionnaire)											x	x	x	x	x	x	x
EGA																						
UK Biobank	x	x	x	x	x							x					x	x		x		x

	Laboratory measures							
	Hematology	Biochemistry	Microbiology	Virology	Immunology	Toxicology	Histology	Genomics
CHILD	x	x (FFQ data)	x (derived from stool)					x
CARTaGENE	x	x						x
CLSA	x	x						x
H3Africa								x
BIOS								x
Estonian Biobank	x							x
CoLaus/PsyCoLaus		x	x (derived from stool)					x
EGA								x
UK Biobank	x							x

	Cognition, personality and psychological measures and assessments		
	Cognitive functioning	Personality	Psychological distress and emotions
CHILD	x	x	x
CARTaGENE			x
CLSA	x	x	x
H3Africa			
BIOS			
Estonian Biobank			x (MINI and SSP interview)
CoLaus/PsyCoLaus	x?		x
EGA			
UK Biobank	x	x	x

	Life events, life plans, beliefs and values		
	Life events	Life plans	Beliefs and values
CHILD	x		
CARTaGENE	x		
CLSA		x	
H3Africa			
BIOS			
Estonian Biobank			
CoLaus/PsyCoLaus			
EGA			
UK Biobank	x		

	Preschool, school and work life		
	Preschool life	School life	Work life
CHILD	x	x	
CARTaGENE			x
CLSA			x
H3Africa			
BIOS			
Estonian Biobank			x
CoLaus/PsyCoLaus			
EGA			
UK Biobank			x

	Social environment and relationships				Other	
	Social network	Social participation	Social support	Parenting and familial environment	Own a pet	Voted in last election
CHILD		x	x	x	x	
CARTaGENE			x			
CLSA	x	x	x		x	x
H3Africa						
BIOS						

Estonian Biobank								
CoLaus/PsyCoLaus								
EGA								
UK Biobank		x	x					

	Physical environment						**Other**	
	Housing characteristics	Built environment/neighbourhood characteristics	Workplace characteristics	Radiation exposure	Chemical exposure	Biological exposure	Various questions	Loud music exposure frequency
CHILD	x	x	x		x	x	x (see HENV questionnaires)	
CARTaGENE				x	x			
CLSA	x	x			x			
H3Africa								
BIOS								
Estonian Biobank								
CoLaus/PsyCoLaus								
EGA								
UK Biobank	x	x	x	x	x			x

	Administrative information					**Other**		
	Identifiers	Date and time-related information	Questionnaire and interview-related information	Physical and cognitive measure and biosample-related information	Data and sample collection center-related information	Consent	Reason for withdrawal	Willing to be contacted
CHILD	x	x	x	x		x	x	x
CARTaGENE	x	x	x	x	x	x		x
CLSA	x	x	x	x	x	x		
H3Africa	x		(depends on subproject)					
BIOS	x							
Estonian Biobank			x					
CoLaus/PsyCoLaus			x					
EGA	x							
UK Biobank	x	x	x	x	x			x

8.2 Initial Use Case Metadata Requirements

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

8.2 Initial Use Case Metadata Requirements

Category	Sub-category	Requirement	Use Case	Use Case description	Owner(s)
Genomics	Sequencing data	Expression data (RNAseq) and data formats are available in relevant cohorts, e.g. RNAseq in fastq format	4.1	Federated eQTL analysis	Kaur Alasoo, Harm-Jan Westra, Will Rayner
Genomics	Sequencing data	Genotype data and data formats are available in relevant cohorts e.g. VCF format, BCF, BGEN, etc.	4.1	Federated eQTL analysis	Kaur Alasoo, Harm-Jan Westra, Will Rayner
Genomics	Sequencing metadata	Associated cell/tissue type	4.1	Federated eQTL analysis	Kaur Alasoo, Harm-Jan Westra, Will Rayner
Genomics	Sequencing metadata	eQTL format: Cis eQTLs or Trans eQTLs or both	4.1	Federated eQTL analysis	Kaur Alasoo, Harm-Jan Westra, Will Rayner
Derived data	Statistics	Summary statistics for additional data from federated sources	4.1	Federated eQTL analysis	Kaur Alasoo, Harm-Jan Westra, Will Rayner
Questionnaire	Diseases/Medication	Trait/Disease/Medication (use as disease proxy) associated with genotype	4.1	Federated eQTL analysis	Kaur Alasoo, Harm-Jan Westra, Will Rayner
Questionnaire	Socio-demographic and economic characteristics	Covariates e.g. gender, age, etc.	4.1	Federated eQTL analysis	Kaur Alasoo, Harm-Jan Westra, Will Rayner
Questionnaire	General cohort metadata	Population	4.1	Federated eQTL analysis	Kaur Alasoo, Harm-Jan Westra, Will Rayner
Genomics	Sequencing data	Genotype data	4.2	Polygenic risk score analysis	Mark McCarthy, Will Rayner
Questionnaire	Diseases,Symptoms and signs,Medication	Trait/Disease/Medication associated with genotype	4.2	Polygenic risk score analysis	Mark McCarthy, Will Rayner
Derived data	Statistics	Summary statistics for additional data from federated sources	4.2	Polygenic risk score analysis	Mark McCarthy, Will Rayner
Biosamples	Biosample metadata	Tissue/liquid sample information	5.1	Query of biobank catalogue data	Morris Swertz (UMCG), Esther van Enckevort (UMCG)
Biosamples	Biosample metadata	Tissue type and processing	5.1	Query of biobank catalogue data	Morris Swertz (UMCG), Esther van Enckevort (UMCG)
Biosamples	Biosample metadata	Storage, availability	5.1	Query of biobank catalogue data	Morris Swertz (UMCG), Esther van Enckevort (UMCG)
Questionnaire	Diseases,Symptoms and signs,Medication	Clinical data on relevant patients	5.1	Query of biobank catalogue data	Morris Swertz (UMCG), Esther van Enckevort (UMCG)
Genomics	Sequencing data	Genomic and transcript data	5.2	FAIR Data analysis for (pancreatic) cancer diagnostics	Andrew Stubbs (EMC)
Genomics	Sequencing data	NGS (whole genome, whole exome, panel-based)	5.2	FAIR Data analysis for (pancreatic) cancer diagnostics	Andrew Stubbs (EMC)
Genomics	Sequencing data	SNP arrays	5.2	FAIR Data analysis for (pancreatic) cancer diagnostics	Andrew Stubbs (EMC)
Genomics	Sequencing data	Methylation	5.2	FAIR Data analysis for (pancreatic) cancer diagnostics	Andrew Stubbs (EMC)
Genomics	Sequencing data	Metagenomics	5.2	FAIR Data analysis for (pancreatic) cancer diagnostics	Andrew Stubbs (EMC)
Questionnaire	Diseases,Symptoms and signs,Medication	Clinical information, eg survival analysis, stratification based on treatment, toxicity	5.2	FAIR Data analysis for (pancreatic) cancer diagnostics	Andrew Stubbs (EMC)
Genomics	Sequencing data	DNA, RNA and other molecular data for test and reference sets	5.3.1	Analytical sandbox for diagnostic services	Morris Swertz (UMCG)
Questionnaire	Diseases,Symptoms and signs,Medication	Associated symptoms, prescriptions etc.	5.3.1	Analytical sandbox for diagnostic services	Morris Swertz (UMCG)
Genomics	Sequencing data	Sequences and variants such SNPs, CNV	5.3.2	Scoring service to assess pathogenicity scales of variants	Patrick Ruch (HES), Morris Swertz (UMCG)
Questionnaire	Diseases,Symptoms and signs,Medication	Structured clinical data (demographics, diagnosis, prescription)	5.3.2	Scoring service to assess pathogenicity scales of variants	Patrick Ruch (HES), Morris Swertz (UMCG)
Documentation	Documentation	General research data documents (pdf, ppt, images etc)	5.3.2	Scoring service to assess pathogenicity scales of variants	Patrick Ruch (HES), Morris Swertz (UMCG)
Genomics	Sequencing data	NGS	5.3.3	GDPR/FAIR compliant Diagnostic reporting	Thomas Binsl, Alex Michie (CGO)
Questionnaire	Diseases, Symptoms and signs, Medication, Non-pharmacological interventions	Associated oncological data	5.3.3	GDPR/FAIR compliant Diagnostic reporting	Thomas Binsl, Alex Michie (CGO)
Genomics	Sequencing data	NGS	5.4	Curation support for care planning	Patrick Ruch (HES)
Genomics	Sequencing data	SNPs	5.4	Curation support for care planning	Patrick Ruch (HES)
Questionnaire	Diseases,Symptoms and signs,Medication	Diagnosis/symptoms	5.4	Curation support for care planning	Patrick Ruch (HES)
Questionnaire	Diseases,Symptoms and signs,Medication	Pathology information	5.4	Curation support for care planning	Patrick Ruch (HES)
Questionnaire	Diseases,Symptoms and signs,Medication	Drug responses	5.4	Curation support for care planning	Patrick Ruch (HES)

8.3 List of catalogues used in development of the minimal metadata model

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

8.3 List of catalogues used in development of the minimal metadata model

Initiative	url	Contents	Country	Source
BBMRI-ERIC Directory	http://directory.bbmri-eric.eu			David
B.R.I.D.G.E. TO DATA	https://www.bridgetodata.org/	Various study designs	International	Source: https://doi.org/10.1371/journal.pone.0200926
Biological and BioMolecular resources Research InfrastructureLarge Prospective Cohorts (BBMRI-LPC)	http://www.bbmri-lpc-biobanks.eu/catalogue.html			David
Biomarker for Cardiovascular Risk Assessment in Europe (BiomarCaRE)	http://www.biomarcare.eu/	Cohorts and clinical trials	International	Source: https://doi.org/10.1371/journal.pone.0200926
Birthcohorts.net	http://www.birthcohorts.net/	Cohorts	International	Source: https://doi.org/10.1371/journal.pone.0200926
Cancer Epidemiology Descriptive Cohort Database (CEDCD)	https://cedcd.nci.nih.gov/	Cohorts	International	Source: https://doi.org/10.1371/journal.pone.0200926
Centre for Longitudinal Studies (CLS)	http://www.cls.ioe.ac.uk/	Cohorts	United Kingdom	Source: https://doi.org/10.1371/journal.pone.0200926
Cohort and Longitudinal Studies Enhancement Resources (CLOSER)	http://www.closer.ac.uk/	Cohorts	United Kingdom	Source: https://doi.org/10.1371/journal.pone.0200926
EU Joint Programme—Neurodegenerative Disease Research Global Cohort Portal (JPND)	http://www.neurodegenerationresearch.eu/	Cohorts	International	Source: https://doi.org/10.1371/journal.pone.0200926
Inter-university Consortium for Political and Social Research (ICPSR)	https://www.icpsr.umich.edu/icpsrweb/	Various study designs	International	Source: https://doi.org/10.1371/journal.pone.0200926
InterConnect	http://www.interconnect-diabetes.eu		Europe	David
Lifecycle	http://catalogue.lifecycle-project.eu		Europe	David
Lifelines	http://catalogue.lifelines.nl		Netherlands	David - currently offline due to tech difficulties
Maelstrom Research catalogue	https://www.maelstrom-research.org	Various study designs	Canada	David
Maternal, Infant, Child & Youth Research Network (MICYRN)	http://micyrn.ca/	Various study designs	Canada	Source: https://doi.org/10.1371/journal.pone.0200926
Medical Research Council Research Data Gateway	https://www.mrc.ac.uk/research/facilities-and-resources-for-researchers/mrc-researchdata-gateway/	Cohorts	United Kingdom	Source: https://doi.org/10.1371/journal.pone.0200926
National Archive of Computerized Data on Aging (NACDA)	http://www.icpsr.umich.edu/icpsrweb/NACDA/	Various study designs	International	Source: https://doi.org/10.1371/journal.pone.0200926
ONTOFORCE	https://www.ontoforce.com/	Various databases	International	Source: https://doi.org/10.1371/journal.pone.0200926
Portail Epidemiologie France	https://epidemiologie-france.aviesan.fr/	Various study designs	France	Source: https://doi.org/10.1371/journal.pone.0200926
RAND Survey Metadata Repository	https://www.rand.org/labor/data.html	Various study designs	International	Source: https://doi.org/10.1371/journal.pone.0200926
RD-Connect sample catalogue	https://samples.rd-connect.eu/	Various study designs	Europe	David
RD-connect registry finder	http://catalogue.rd-connect.eu/			
Registry of Research Data Repositories (re3data.org)	http://www.re3data.org/	Various databases	International	Source: https://doi.org/10.1371/journal.pone.0200926
Swedish National Data Service (SND)	https://snd.gu.se/en	Various databases	Sweden	Source: https://doi.org/10.1371/journal.pone.0200926
The Gateway to Global Aging Data	https://g2aging.org/?	Cohorts	International	Source: https://doi.org/10.1371/journal.pone.0200926
The Global Alzheimer's Association Interactive Network (GAAIN)	http://www.gaain.org/	Cohorts	International	Source: https://doi.org/10.1371/journal.pone.0200926
UK Data service	https://www.ukdataservice.ac.uk/	Various study designs	United Kingdom	Source: https://doi.org/10.1371/journal.pone.0200926
HBP catalogue	https://kg.ebrains.eu/search/?facet_type[0]=Dataset	Cohorts	Europe	
SwissEGA	http://candy.hesge.ch/SwissEGA/index.jsp	Cohorts	Switzerland	Source:https://doi.org/10.1093/database/bax083

8.4 Base material for cohort minimal meta data model

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 825775.

8.4 Base material for cohort minimal meta data model

Base material for cohort minimal meta data model									
Summary:	This is a first pass of gathering and structuring key data to be used to build the CINECA WP3 minimal metadata model - current under development.								
Content:	The chosen sub-categories are shaped by common data categories collected by major catalogues such as Maelstrom. The chosen variables are either CINECA use case requirements, or variables that can be mapped to by the majority of CINECA cohorts, or typical metadata variables collected by major catalogues. We have applied ontology terms to variables (mostly from NCIT). We are also looking into cross referencing with existing minimal data models.								
Attribution:	If you use this work, please attribute CINECA and the primary contacts listed below.								
Primary Contacts:	Vivian Jin and Fiona Brinkman (SFU, Canada); Melanie Courtot (EBI, UK).								
Possible next steps:	Create application ontology and structure terms in a hierarchy for easier search. For each variable, specify values and associated ontology terms (eg. for gender - specify male/female/transgender/non-binary).								
9 CINECA cohorts	**stand-alone cohorts**	**consortiums**							
	CHILD	H3Africa							
	CARTaGENE	BIOS							
	CLSA	EGA							
	Estonian Biobank								
	CoLaus/PsyCoLaus								
	UK Biobank								
					NOTE: use blank response if answer to question description is N/A				
Broad category	**Sub-category**	**Sub-category/variable**	**Variable**	**Question Description**	**Expected Answer Type**	**Ontology label**	**Ontology ID**	**Ontology definition**	**Known num. of cohorts with this data (if greater zero)**	**Use Cases Requirement**
basic cohort attributes	aims and objectives			What are the aims and objectives of your cohort?	Alphanumeric	Study Objective	NCIT:C93415	The reason for performing a study in terms of the scientific questions to be answered by the analysis of data collected during the study.	9	
	timeline			What is the timeline for your cohort study?	Alphanumeric	Timeline	NCIT:C54576	A chronological schedule of when activities or events occurred or will occur.	9	
	study design (eg. longitudinal)			What is the study design for your cohort study?	MIABIS codes (Categorical)	Study Design	NCIT:C15320	A plan detailing how a study will be performed in order to represent the phenomenon under examination, to answer the research questions that have been asked, and defining the methods of data analysis. Study design is driven by research hypothesis being posed, study subject/population/sample available, logistics/resources: technology, support, networking, collaborative support, etc.…	9	
	population data (Population Group? NCIT: C17005)	location		What location is your cohort's population based in?	Alphanumeric	Location	NCIT:C25341	A position, site, or point in space where something can be found.	9	4.1
		criteria for enrollment and recruitment procedures		What is your cohort's criteria of enrollment or recruitment procedures?	Alphanumeric	Inclusion Criteria	NCIT:C25532	Medical and/or social characteristics which are necessary for a subject to be allowed to participate in a clinical study, as outlined in the study protocol. Meeting inclusion criteria is not a sufficient condition for entry or recruitment of a subject into the study. Characteristics limiting the eligibility of a subject for the clinical study must be considered	6	4.1
		num. participants		What are the number of participants in your cohort?	Numeric	Planned Subject Number	NCIT:C49692	The number of subjects entered in a clinical trial.	9	4.1, 5
	demographic data (NCIT:C142508)	sex(es) studied in cohort	** these are also captured in the survey category	How many of each sex are studied in your cohort?	Numeric	Sex	NCIT:C28421	The assemblage of physical properties or qualities by which male is distinguished from female; the physical difference between male and female; the distinguishing peculiarity of male or female.	9	
		gender(s) studied in cohort		How many of each gender are studied in your cohort?	Numeric	Gender	NCIT:C17357	The assemblage of properties that distinguish people on the basis of the societal roles expected for the two sexes.		
		age range		What is the age range of your cohort's participants?	Alphanumeric				6	5.3
	data collection events			If data was collected over several events or time	Alphanumeric					
biosample (NCIT:C43412)	sample type (Body fluid or substance NCIT: C13236)	urine		Have you collected this type of biosample?	Yes/No (Binary)	Urine	NCIT:C13283	The fluid that is excreted by the	4	5.1
		blood	venous or arterial			Blood	NCIT:C12434	A liquid tissue; its major function is to transport oxygen throughout the body. It also supplies the tissues with nutrients, removes waste products, and contains various components of the immune system defending the body against infection. Several hormones also travel in the blood.		

			fasting or non-fasting						5	
		stool				Feces	NCIT:C13234	The material discharged from the bowel during defecation. It consists of undigested food, intestinal mucus, epithelial cells, and bacteria.	2	
		saliva				Saliva	NCIT:C13275	The watery fluid in the mouth made by the salivary glands. Saliva moistens food to help digestion and it helps protect the mouth against infections.	2	
		other								
	availability			For each biosample present, is it available/accessible for research?	DUO code(s)	Availability	NCIT:C25429	The quality of being obtainable or accessible and ready for use or service.	TBD	5.1
	sample size			For each biosample present, what is the sample size?	Numeric	Sample Size				
	processing method			For each biosample present, what was the processing method used?	Alphanumeric (or Categorical?)	Methodology	NCIT:C86572	The system of methods followed in a particular discipline.	TBD	5.1
	storage method			For each biosample present, what is the storage method?	MIABIS codes (Categorical)	Defined Specimen Storage	NCIT:C93374	An administrative activity defined at a global library level that is an action of storing samples.	TBD	5.1
laboratory measures (Laboratory Procedure NCIT:C25294)	microbiology (Microbiology Test NCIT:C49188)	microbial data		Have you collected microbial data?	Yes/No (Binary)				TBD	
		biosample source/anatomical location		From where (specimen source site) have you collected microbial data?	Alphanumeric (or Categorical?)	Specimen Source Site	NCIT:C159256	A request to identify the specimen source site.		
		available data format(s)		For each specimen source site, what are the available data formats?	Alphanumeric (or Categorical?)	Format	NCIT:C42761	The organization of information according to preset specifications.	TBD	
	genomics (Molecular Analysis? NCIT:C19770)	data type	DNA/Genotyping	Have you collected this type of data?	Yes/No (Binary)	DNA Sequencing/Nucleic Acid Sequencing	NCIT: C153598/NCIT: C18881		9	4.1,4.2,5.2,5.3
			WGS			Whole Genome Sequencing	NCIT:C101294		7	5.2,5.4
			WES			Whole Exome Sequencing	NCIT:C101295		4	5.2,5.4
			Sequence variants (CNV, SNP arrays)			Single Nucleotide Polymorphism Profile	NCIT:C129888	The analysis of all of the single nucleotide polymorphisms in the genome of a biological sample.	TBD	5.2,5.3,5.4
			Epigenetics			Epigenetic Profile	NCIT:C129887	The analysis of all of the epigenetic DNA modifications in the genome a biological sample.	4	5.2
			Metagenomics			Metagenomics analysis	ERO:0000657 (***ERO is obsolete)	A molecular assay that is used to analyze metagenomic data; genetic material recovered directly from environmental samples for genomic research.	TBD	5.2
			Microbiome markers (rRNA, etc)							
			RNAseq/gene expression			mRNA sequencing	NCIT:C129432	A procedure that can determine the RNA sequences for all or part of the poly-A tail-containing messenger RNA transcripts in an individual.	6	4.1,5.3
			eQTL (Cis eQTLs and/or Trans eQTLs)			Expression Quantitative Trait Locus	NCIT:C113415	A stretch of DNA at a particular chromosomal location that is able to regulate the expression of a specific mRNA or protein.	TBD	4.1
			other							
		sample size		For each data type present, what is the sample size?	Numeric	Sample Size	NCIT:C53190	A subset of a larger population, selected for investigation to draw conclusions or make estimates about the larger population.	TBD	all?
		available data format(s)		For each data type present, what are the available data formats?	Alphanumeric (or Categorical?)	Format	NCIT:C42761	The organization of information according to preset specifications.	TBD	4.1
		availability		For each data type present, is it available/accessible for research?	DUO code(s)	Availability	NCIT:C25429	The quality of being obtainable or accessible and ready for use or service.	TBD	all?
		processing method		For each data type present, what was the processing method used (eg. sequencer/software)?	Alphanumeric	Methodology	NCIT:C86572	The system of methods followed in a particular discipline.	TBD	
		associated cell type/tissue type/biosample		For each data type present, what are the associated cell/tissue/biosample types?	Alphanumeric (or Categorical?)	Specimen Source Site	NCIT:C159256	A request to identify the specimen source site.	TBD	
		associated phenotype		For each data type present, what are the associated phenotype(s)?	Alphanumeric	Trait/Phenotype	NCIT: C985496/NCIT: C16977	Any genetically determined characteristic/The assemblage of traits or outward appearance of an individual. It is the product of interactions between genes and between genes and the environment.		
		main diagnosis		For each data type present, what is the main diagnosis ?	Alphanumeric	Diagnosis				
		associated disease		For each data type present, what are the associated diseases?	Alphanumeric	Diagnosis	NCIT:C15220	The investigation, analysis and recognition of the presence and nature of disease, condition, or injury from expressed signs and symptoms; also, the scientific determination of any kind; the concise results of such an investigation.		5
		associated medication		For each data type present, what are the associated medications?	Alphanumeric	Medication	NCIT:C459		TBD	4.1,4.2,5.1,5.2,5.3
survey administration (NCIT:C64252)	date and time-related information			Have you collected the date/time of survey assessments?	Yes/No (Binary)	Assessment Date	NCIT:C93511	The date (and time) on which an assessment is completed.	4	
	consent/accessibility			Is there consent for your survey data to be accessible for research?	DUO code(s)	Informed consent	NCIT:C16735	Consent by a patient to a surgical or medical procedure or participation in a clinical study after achieving an understanding of the relevant medical facts and the risks involved.	TBD	
	unique identifiers			Are there unique identifiers for subjects/participants?	Yes/No (Binary)	Identifier	NCIT:C25364	One or more characters used to identify, name, or characterize the nature, properties, or contents of a thing.	6	all?

questionnaire/survey data (NCIT: C17176)	socio-demographic and economic characteristics	age/birthdate		If this data type is present, what are the associated variables from your cohort's survey/questionnaire? For each associated variable, what is the data type (eg. categorical/binary/numeric/alphanumeric/datetime), the data collection event or time point (if applicable), the population type (if there are different participants sharing one identifier; eg. child, mother, father)?	.csv format	Age /Birth Date	NCIT: C25150/NCIT: C68615		5	4.1, 5
		biological sex				Sex	NCIT:C28421	The assemblage of physical properties or qualities by which male is distinguished from female; the physical difference between male and female; the distinguishing peculiarity of male or female.	8	4.1, 5
		gender				Gender	NCIT:C17357	The assemblage of properties that distinguish people on the basis of the societal roles expected for the two sexes.	8	5
		ethnicity/race				Ethnicity/Race	NCIT: C16564/NCIT: C17049	A social group characterized by a distinctive social and cultural tradition that is maintained from generation to generation. Members share a common history and origin and a sense of identification with the group. They have similar and distinctive features in their lifestyle habits and shared experiences. They often have a common genetic heritage which may be reflected in their experience of health and disease...	6	4.1, 5
		genealogy							TBD	4.1
		birthplace							6	4.1?
		residence				Residence	NCIT:C25273	Any address at which a person dwells more than temporarily.	6	4.1
		education				Education	NCIT:C16529	The activities of educating or instructing or teaching; activities that impart knowledge or skill.	6	
		family and household structure							6	
	lifestyle and behaviours (lifestyle NCIT:C16795)	tobacco				Tobacco Smoking History	NCIT:C29719	A record of an individual's background in regard to smoking tobacco. This would include such factors as start date, end date (if applicable), number of cigarette smoked, attempts to quit, and others.	6	
		alcohol				Alcohol Use History	NCIT:C81229	A description of an individual's current and past experience with alcoholic beverage consumption.	5	
		physical activity				Physical Activity Measurement	NCIT:C120914	A measurement of a subject's physical activity or movement.	6	
		sleep				Sleep	NCIT:C73425	A natural and periodic state of rest during which consciousness of the world is suspended.	6	
		nutrition				Nutrition	NCIT:C28294	That which is consumed to fuel necessary life processes of an organism.	6	
	physician/practitioner info					Admitting Physician	NCIT:C51798	The physician responsible for the admission of a patient to a hospital or other inpatient health institution. The admitting physician evaluates patients, makes admitting decisions and assesses diagnostic and treatment plans	TBD	
	diseases (Disease or Disorder NCIT:C2991)	diagnosis	blood-related disorders			Hematopoietic and Lymphoid System Disorder	NCIT:C35814	Any deviation from the normal structure or function of the blood or lymphatic system that is manifested by a characteristic set of symptoms and signs.	5	4.1,4.2,5.1,5.2,5.3,5.4
			endocrine/nutritional/metabolic disorders			Endocrine System Disorder	NCIT_C3009	A non-neoplastic or neoplastic disorder that affects the endocrine system. Representative examples of non-neoplastic disorders include diabetes mellitus, hyperthyroidism, and adrenal gland insufficiency. Representative examples of neoplastic disorders include carcinoid tumor, neuroendocrine carcinoma, and pheochromocytoma.…	6	
			mental and behaviour disorders			Psychiatric Disorder/Behavioural Disorder	NCIT:C2893/NCIT: C35470	A disorder characterized by behavioral and/or psychological abnormalities, often accompanied by physical symptoms. The symptoms may cause clinically significant distress or impairment in social and occupational areas of functioning. Representative examples include anxiety disorders, cognitive disorders, mood disorders and schizophrenia…./A specific behavioral problem that occurs in persistent patterns and characteristic clusters and that causes clinically significant impairment.	5	
			nervous system			Nervous System Disorder	NCIT:C26835	A non-neoplastic or neoplastic disorder that affects the brain, spinal cord, or peripheral nerves.	5	
			digestive system			Digestive System Disorder	NCIT:C2990	A non-neoplastic or neoplastic disorder that affects the gastrointestinal tract, anus, liver, biliary system, and pancreas.	5	

			respiratory system			Respiratory System Disorder	NCIT:C26871	A non-neoplastic or neoplastic disorder that affects the respiratory system. Representative examples include pneumonia, chronic obstructive pulmonary disease, pulmonary failure, lung adenoma, lung carcinoma, and tracheal carcinoma.	5	
			circulatory system			Cardiovascular Disorder	NCIT:C2931		5	
			oncological			Cancer-Related Condition	NCIT:C8278		TBD	
	signs and symptoms (Sign or Symptom NCIT: C100104)								5	
	physiological measurements	anthropometry	weight			Weight	NCIT:C25208	The vertical force exerted by a mass as a result of gravity.	6	
			height			Height	NCIT:C25347	The vertical measurement or distance from the base to the top of an object; the vertical dimension of extension.		
		circulation and respiration	blood pressure			Blood Pressure	NCIT:C54706	The pressure of the circulating blood against the walls of the blood vessels.	6	
			heart rate (HR)			Heart Rate	NCIT:C49677	The number of heartbeats per unit of time, usually expressed as beats per minute.		
	non-pharmacological interventions	surgical interventions				Vasculature Mechanical or Surgical Intervention	NCIT_C119212		6	
	medication (Medication NCIT:C459)	associated disease(s)							6	4.1,4.2,5.1,5.2,5.3,5.4
		prescription				Prescription	NCIT:C28180	A verbal or written order given by an authorized person instructing a patient to obtain and use a medical device, prescription or undergo a procedure.		
		drug response(s)								
		posology				Dosage Regimen	NCIT:C142516	The specific way a therapeutic drug is to be taken, including formulation, route of administration, dose, dosing interval, and treatment duration.		
		administration method								
	life stage/time point					Life Stage	NCIT:C89335	A designation assigned to a particular period during a life cycle, generally defined by chronological parameters.	TBD	
general research data documents (pdf, ppt, images etc)						Research Material	NCIT:C84338	Any item with which a scientist works.	TBD	5.3
statistics	summary statistics for additional data from federated sources					Statistical Analysis Documentation	NCIT:C115732	Records pertaining to the statistical analyses and reports of a clinical trial.	TBD	4.1,4.2