19 | 10 | 2020

Jean-François **GORSE**, Xavier **GANSEL**, MAEL **LE GALL,** Emmanuelle **THERON**

VALUE-Dx WP3 – Task 3.4

OHDSI-OMOP

A proposed standard for AMR data Modelling : Readiness and Recommendations

VALUE-Dx

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 820755. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and bioMérieux SA, Janssen Pharmaceutica NV, Accelerate Diagnostics S.L., Abbott, Bio-Rad Laboratories, BD Switzerland Sàrl, and The Wellcome Trust Limited.



University
of Antwerp

BIOMÉRIEUX

wellcome

UNIVERSITY OF
OXFORD

UMC Utrecht

Penta
Child Health Research

UNIVERSITÀ
di VERONA

THE UNIVERSITY
of EDINBURGH

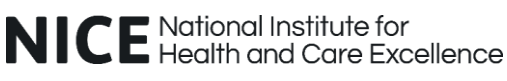
University Medical Center Groningen

RAMBAM
Health Care Campus

UNIVERSIDAD
DE LA RIOJA

BOSTON
UNIVERSITY

BIOASTER
MICROBIOLOGY TECHNOLOGY INSTITUTE

FIND

NICE National Institute for
Health and Care Excellence

Gesundheit Österreich
GmbH

ESCMID
EUROPEAN SOCIETY
OF CLINICAL MICROBIOLOGY
AND INFECTIOUS DISEASES

ERS
EUROPEAN
RESPIRATORY
SOCIETY
every breath counts

Berry Consultants
Statistical Innovation

IBBL
INTEGRATED BIOBANK OF LUXEMBOURG
FOR NEXT GENERATION HEALTHCARE

ZonMw

janssen

ACCELERATE
DIAGNOSTICS

Abbott

BIO-RAD

BD

Table of contents

1. Background of the study

1.1. Purpose

It is essential in the fight against Antimicrobial Resistance (AMR) that all relevant data (e.g. diagnostic test results, Antibiotic Susceptibility Testing (AST) data, patient records and treatment regimens) can be collected from diverse sources and systems in a timely, accurate and efficient manner, and shared with concerned stakeholders, in a standardised and easily accessible way.

The purpose of this study is to assess how modern standards-based connectivity and interoperability solutions can be used to allow information (i.e. anonymised AMR case data) to be automatically connected from diagnostic devices and or Health Information Systems and subsequently shared across laboratories and partners.

In the course of VALUE-Dx Task 3.4, two main use-cases were proposed to be studied:

- Ability to profile an AMR data provider in terms of testing resources (equipment and associated diagnostic systems), testing protocols (nature of specimen able to be processed), and local ecology
- Ability to report aggregated (across AMR data providers) micro-organisms occurrence and Antibiotic resistance

In the long-term, the intent would be to leverage such enabling technologies within the ECRAID[1] network and therefore allowing for the digitalisation of clinical trials.

1.2. A stepwise approach for the study

In order to conduct this study a series of analysis and implementation steps were defined.

Step1: Identify relevant technologies in this application field, including similar experiences conducted by other teams in EU or elsewhere in the world.

Step2: Conduct a survey on existing laboratory capabilities in order to determine connectivity capabilities, compliance to existing information interchange standards. This survey was conducted through a connectivity questionnaire.

Step3: Select one technology and implement a Proof of Concept (POC) using simulated data in a limited network infrastructure

Step4: Prepare a second Proof of Concept (if project budget permits) where the technology is deployed on a limited number of clinical trial sites of VALUE-Dx.

Step5: Build an evaluation report and establish technology implementation guidelines.

1 European Clinical Research Alliance on Infectious Diseases

1.3. Numerous challenges to be overcome

Five main challenges can easily be identified.

1. Where should the data be coming from within the complex set of healthcare information systems?
2. Which kind of data has to be exchanged?
3. How this data should be aggregated at the network level?
4. How can we make the data interoperable among different IT systems (languages, codes…)?
5. How can we protect privacy and ensure security (within European GDPR[2])?

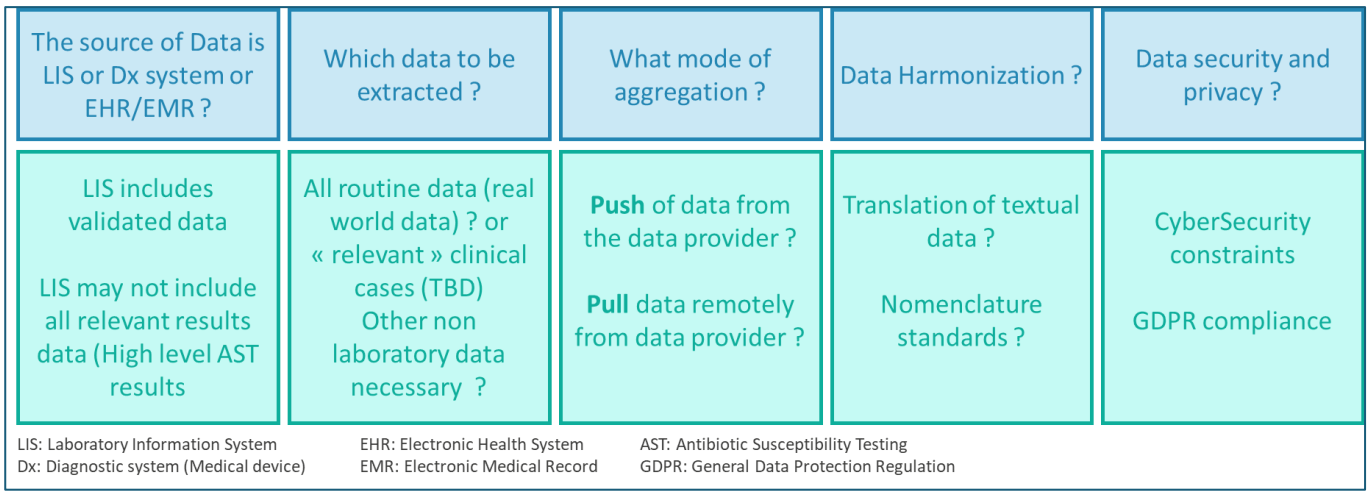


Figure 1: The five main challenges

1.3.1. Data source challenge

Diagnostic systems (Dx) provide for very precise data when it comes to release results to other Information systems, however this level of precision may not always be able to be integrated into the next IT system in the process as the Laboratory Information System (LIS). This may lead to some data loss, such as resistance phenotypes for example. On the other hand, some patient related information is generally not available at the bench where the Dx operates.

Domain Middleware (M/W) may sometimes be in charge of connecting multiple Dx of the same domain (Microbiology, Immunology …) in order to ensure workflow between these systems and also to provide for additional data analytics capabilities specific to the domain that cannot be handled by the LIS.

Laboratory Information System (LIS) is usually connected to the Dx and/or M/W and in charge of delivering the results to the clinician either via connectivity to Electronic Medical Records (EMR) of the Hospital Information System (HIS) or Electronic Health Records (EHR) systems.

Depending on the level of data precision required, the data source can be any of these systems.

2 GDPR : General Data Protection Regulation

1.3.2. Data Nature Challenge

Depending on the type of study to be run, all routine data could be used for real world data studies (or observational study) or only data relevant to specific clinical cases making the data selection process more difficult.

It has to be established what patient clinical data is necessary for the studies that are intended, since they may be stored in different IT systems.

1.3.3. Data aggregation challenge

Data aggregation is necessary in the context of a network of data providers. Two main options exist:

1| Push of the data from the data provider to a centralized data store (warehouse or Data Lake).
2| Pull data query results from a central location.

The option 1 is called a centralised architecture, option 2 is called a federated architecture.

A few standards exist today to administrate relationship between IT systems. They address different levels of exchange such as communication protocols as well as communication messages.

Communication protocols regulate the exchanges between Dx and LIS as well as LIS and HIS. The messaging formats are based upon the HL7 (Health Level 7) format which includes message identifiers and codes in precise message locations for the data which has to be embedded. However, many Dx systems may not yet comply to all these standards and possibly still use proprietary communication protocols and messages…

1.3.4. Data harmonisation challenges

In order to be able to aggregate data from different data providers located potentially in different countries, speaking different languages, using different IT systems, a common data vocabulary has to be used. This is the basis for interoperability.

For laboratory testing such as Micro-organism identification or detection and Antibiotic Susceptibility Testing, standardised vocabularies are already in use in quite a few countries. Moreover, certain regulations are imposing on the use of specific standards.

Once again, depending on the variety of the data to be retrieved (lab results, clinical data…), and on the aggregation scheme, the standardisation may need to be pushed beyond alignment on a set of standardised vocabularies but also on data organisation through data modelling.

1.3.5. Data Security and privacy challenges

The European GDPR[3] imposes strict constraints on how to handle patient sensitive health data. On the other hand, when building a data network additional cyber security measures have to be put in place.

Depending on the data aggregation scheme (centralised or federated) the field of constraints is different.

3 General Data Protection Regulation

2. Addressing the challenges for POC1

2.1. Selection of Data sources

For the POC1, three types of data will be sourced:

- Microbiology middleware sample data from bioMérieux MYLA® provided as a large text file
- Detection panel data from bioFire FilmArray® as a set of sample XML files
- AMR data from the WHONET example data

The data will be static, meaning that it will be prepared once and not challenged in real-time.

2.2. Nature of data selected for POC1

Limited patient data may be used, when available, in order to trace admission and release as well as the ward in which the patient was located during the testing.

If available, the specimen nature and collection date will be included.

The AMR data will include micro-organism identification results and antibiotic susceptibility results by MIC [4] and clinical categories.

Regarding the detection panels, the micro-organisms detected will be included along with the sample type and sample collection date.

2.3. Data aggregation model

2.3.1. Centralised and Federated architectures

As explained earlier, two schemes are competing:

- The centralised architecture, where all the data itself is pushed into a central location and merged with data coming from all other data providers as a single “warehouse”. The data analysis is conducted on this central location. Results of the analysis can be seen from a remote location.

- The federated architecture where the data still resides at the data provider location and is queried from a central “observatory” location. Each data location is called a node. Only results of the query at the node is pushed to the central observatory. The results of the queries of all nodes are aggregated on the central observatory.

[4] MIC= Minimal Inhibitory Concentration

In order for a centralised architecture to be actionable, data vocabularies have to be harmonised prior to its integration into the “warehouse”. In the following figure, this harmonisation can occur in the box named “Translation services”.

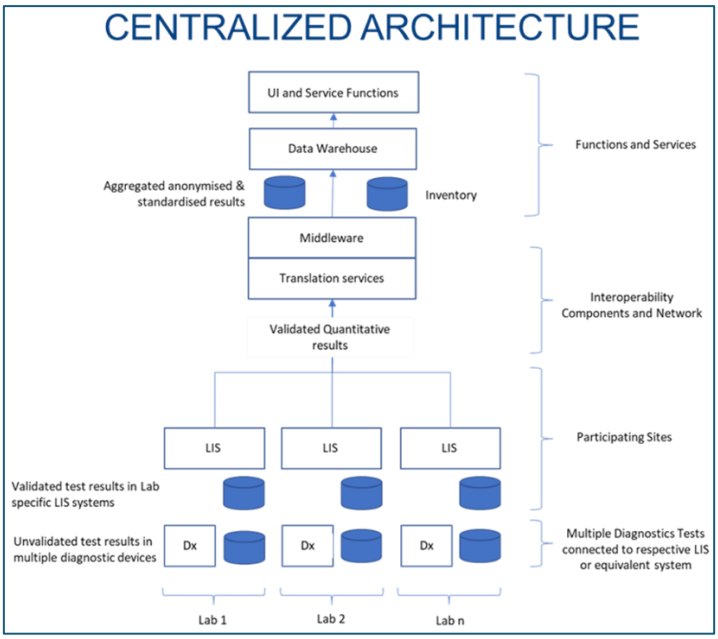


Figure 2: Centralised Architecture

In order for a federated architecture to be actionable, data organisation needs to be harmonised, meaning that all contributing data providers should organise their data storage by following a strict modelling guideline. Therefore, the same data is duplicated and reorganised into the new system. In the following figure it is called “Staged”. Harmonisation of data vocabularies is also required.

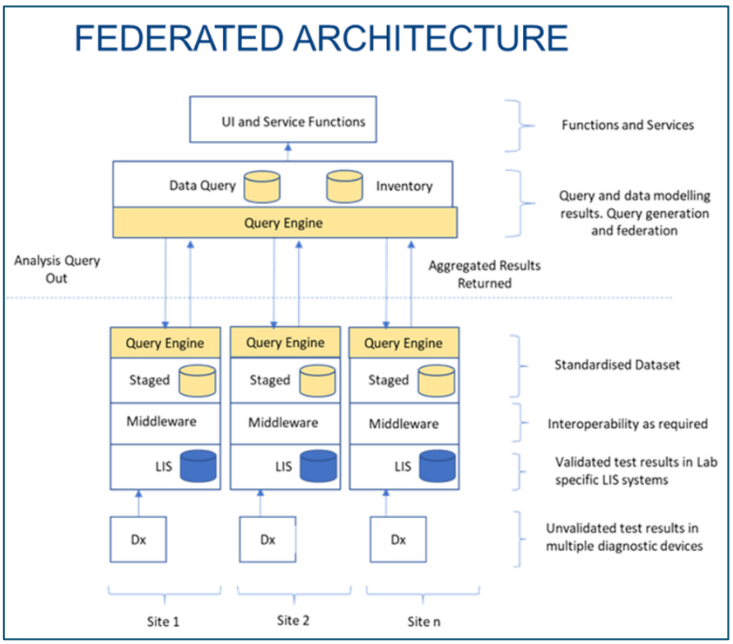


Figure 3: Federated Architecture

Data aggregation can be found in two different types of systems:

- Surveillance systems, such as ECDC-EARS-Net[5], WHO-GLASS[6] where data is collected and aggregated in a central place. It is usually focused on specific organism-drug combinations that are reported. We can consider that it is Isolate-centric.
- Clinical trial or Observational data networks where different application models exist but rely on federated network of data providers. We can consider that it is patient-centric.

When looking at clinical trial data systems, a few application models can be found for each one based on a federated architecture. The advantage of such an architecture resides mainly in the ownership of the data that remains at the data provider level, while avoiding the transfer of highly patient sensitive data to a central repository.

In order to standardise data representation for each of these application models, a Common Data Model (CDM) is enforced, to which every data provider needs to convert its data. Depending on the application model, the data vocabulary harmonisation may be limited or extensive, meaning that every piece of data needs to belong to standardised vocabulary or only a few have to be "mapped" to a standard vocabulary.

The following includes a few examples of Clinical research infrastructures:

- I2B2: Informatics for Integrating Biology & the Bedside

  This is a standardised data model. A set of tools are available for vocabulary "alignment" and for running queries. This system is used in the US and allows some interoperability with other models (OHDSI-OMOP). This system has been recently used in Spain to follow the COVID-19 pandemic.

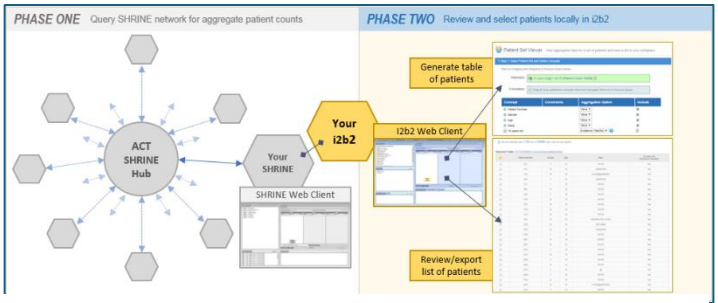


Figure 4: I2B2 Network and process (from community/i2b2.org/wiki)

- FDA-Sentinel

  This is a model promoted and supported by the FDA. It is widely used in the US. Data models and tools are available, data harmonisation is limited to data to be studied.

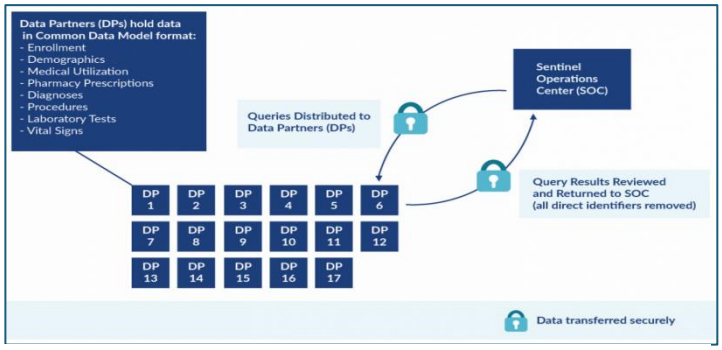


Figure 5: Sentinel network approach (from sentineldevprod.acqui-sites.com)

5 EARS : European Antimicrobial Resistance Surveillance Network

6 GLASS : Global Antimicrobial Resistance Surveillance System

- PCORnet: National Patient-Centered Clinical Research Network
  This a model extension from the previous one, also widely used in the US. A number of data elements have to be mapped to standard vocabularies.

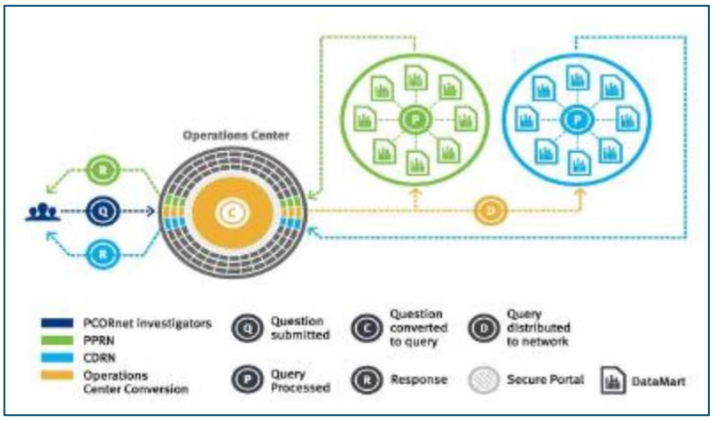


Figure 6: PCORnet Distributed Network
(from PCORnet github)

- OHDSI-OMOP: Observational Health Data Sciences and Informatics – Observational Medical Outcomes Partnership.
  The OMOP-CDM is supported and maintained by a dedicated community. It is used in 17 countries. This model requires the conversion to the CDM as well as a complete mapping of "local" vocabularies to standardised vocabularies. Open source tools are available to build the models, preparing the data vocabulary mappings and preparing data analysis. This model has been selected in previous IMI[7] funded project (EMIF[8]) and is now promoted by another IMI funded project: EHDEN[9].

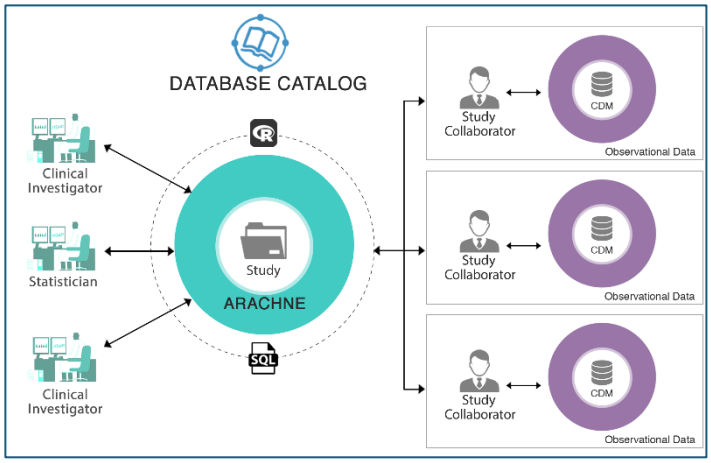


Figure 7: OHDSI Network study Workflow
(from Ohdsi.github.io/TheBookOfOhdsi)

The European Medicines Agency (EMA) has published an interesting review on the different CDM available (European Medicines Agency, 2018).

2.3.2. Experience from EHDEN

The EHDEN project, another IMI funded project, following previous European initiatives in the area of health data infrastructure, has decided to promote the adoption of the OHDSI-OMOP Common Data Model through education and financing Small and Medium Enterprises (SME) to help the Data providers in their efforts to build large data sets according to this model and to undergo data vocabularies mapping.

The OMOP-CDM allows for capturing a large set of data from various sources at the data provider:

7 IMI : Innovative mdeicine Innitaitive

8 EMIF : European Medical Information Framework (www.emif.eu)

9 EHDEN : European health data Evidence Network (www.ehden.eu)

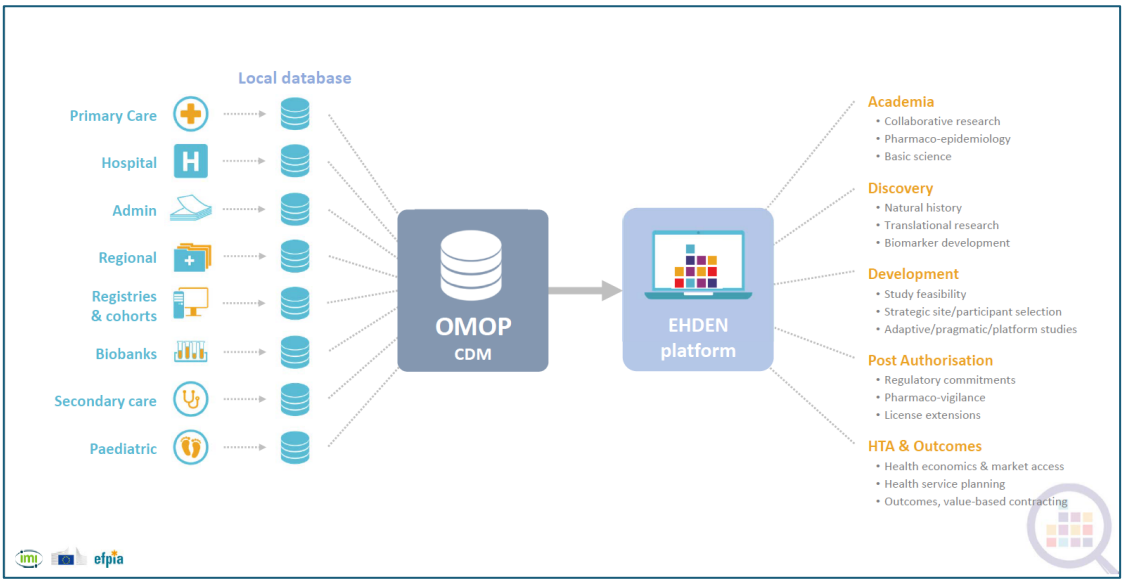


Figure 8: Sources of Data and applications in EHDEN (courtesy of EHDEN Project - N. Hughes)

The data is queried through a federated architecture where software tools reside at the Data provider location, running the query locally and reporting aggregated results to the central observatory.

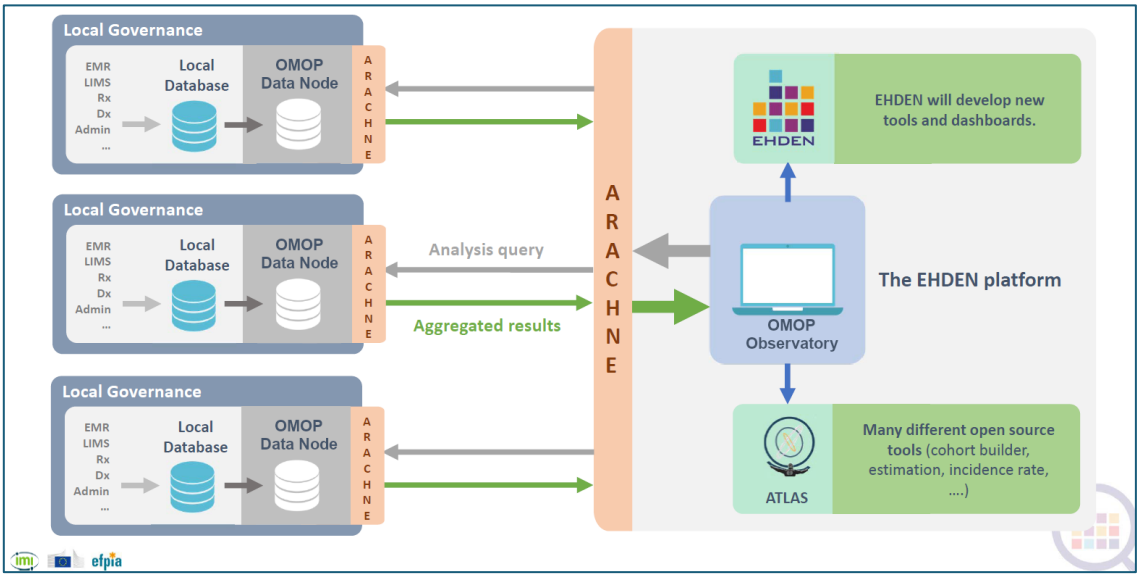


Figure 9: The federated network of EHDEN (courtesy of EHDEN project – N. Hughes)

The data model itself covers a large scope of clinical related data:

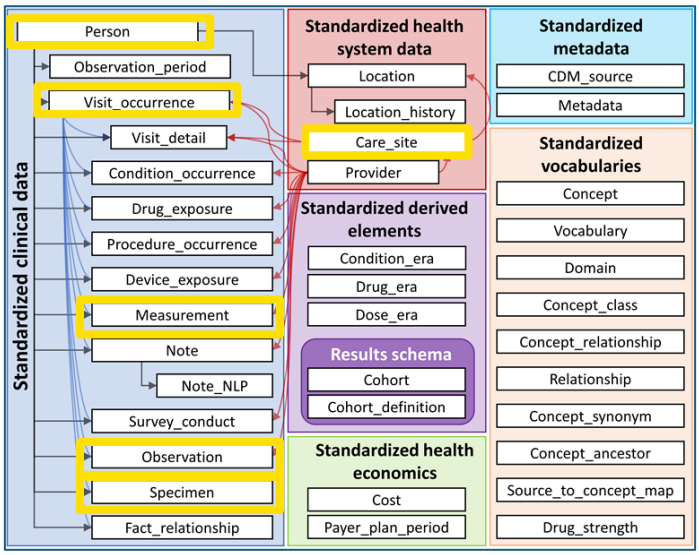


Figure 10: The OHDSI- OMOP Common Data Model
(Yellow squares represent the data that will be collected during POC1)

In the above figure, the yellow squares outline the tables within the model that could be used to represent laboratory results, such as AST data and organism identification or detection.

The data tables named under "Standardised vocabularies" are aimed at mapping all the local vocabularies (text and codes) used by the data provider to a rigorously managed library of codes, themselves leveraging standard vocabularies that had been created to serve clinical domains.

2.4. Data harmonisation and mapping to standardised vocabularies

Throughout the progress made in medicine since Imhotep in Ancient Egypt (considered to be the father of medicine), a number of vocabularies have been established in order to describe clinical symptoms, pathology diagnosis etc. With the introduction of IT systems, many medical codes have flourished, leading to a forest of systems not interoperable However, since a few decades a series of standardisation efforts have been pursued in various domains in order to, not only facilitate digitisation of data (and payments), but also to allow information interchange between different components of the overall IT infrastructure.

The following diagram outlines a few of these vocabularies, such as SNOMED-CT[10], LOINC[11], ICD-9[12], RxNorm[13] etc.

10 SNOMED-CT : Systematized Nomenclature of Medicine – Clinical Terms

11 LOINC : Logical Observation Identifiers Names and Codes

12 ICD : International Classification of Diseases

13 RxNorm : normalized names for clinical drugs (US)

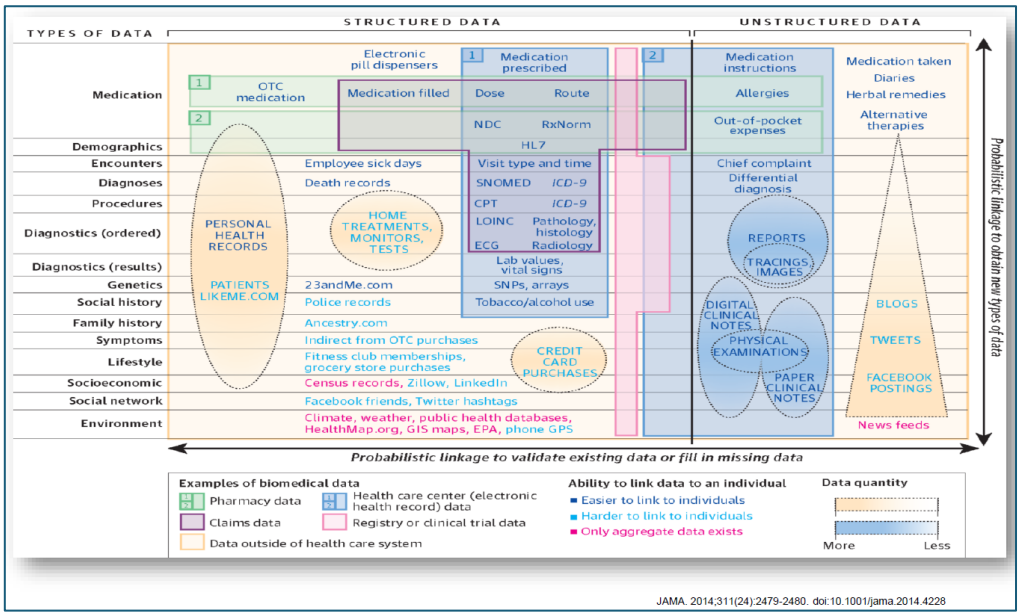


Figure 11: High Value information sources to be linked to an individual
(from JAMA, June 25, 2016, Vol 311, 24)

Within the Lab testing domain, two major vocabularies are found:

- LOINC: which is aimed at coding laboratory requests
- SNOMED-CT: which is aimed at coding test results such as organisms, clinical categories for AST testing. SNOMED CT is not limited to Lab results but is also used to capture clinical observations and other clinically relevant data.

In order to enforce interoperability between IT systems, a few countries have already set regulations imposing the usage of these coding systems. However, the adoption of these coding systems by data providers is slow, and for SNOMED-CT a license is required.

2.4.1. LOINC for test requests

The LOINC codes are maintained by the Regenstrief institute (Indiana University).

Before launching a new IVD test on the market it is mandatory to define the coding of this test and register it, after LOINC curation approval, into the LOINC database in order for IT systems to be able to deploy the code. More information is available at https://loinc.org.

The code is constructed based on 6 semantic parts:

1. Analyte: it describes the molecule the test is measuring, or the organism which is tested
2. Unit: the unit which is used to report the result, or presence absence or value threshold
3. Time: the temporality of the result, end-point measure or kinetics
4. System: the sample upon which the analysis is performed
5. Scale: nature of the result, Ordinal or Numerical
6. Method: protocol used for the testing

The current LOINC database contains more than 90 000 codes and is updated regularly.

A few LOINC code examples:

	LOINC CODE	LOINC TEXT
MALDI-TOF identification test	76346-6	Microorganism identified in Isolate by MS.MALDI-TOF
Automated culture-based identification test	43409-2	Bacteria identified in Isolate by Culture
COVID-19 Test	94565-9	SARS-CoV-2 (COVID-19) RNA [Presence] in Nasopharynx by NAA with non-probe detection
Ampicillin by MIC testing	28-1	Ampicillin [Susceptibility] by Minimum Inhibitory Concentration (MIC)
Ampicillin testing method less	188864-9	Ampicillin [Susceptibility]

Table 1: Examples of LOINC codes applicable for Identification/detection or AST tests

2.4.2. SNOMED-CT for tests results

The SNOMED codes are maintained by SNOMED International which is a non-for-profit organisation. However, a license fee is required before using the coding system. The License can be purchased by a country, in that case the fee is based on the country's wealth, otherwise the license can be purchased for a particular IT product. In August 2020, 39 countries were reported members of SNOMED International.

Its content is updated twice a year.

EU counts 17 member countries: Austria, Belgium, Cyprus, Czech Republic, Denmark, Spain, Estonia, Finland, Ireland, Lithuania, Luxembourg, Malta, The Netherlands, Portugal, Slovak Republic, Republic of Slovenia, Sweden.

Figure 12: Member countries in SNOMED International

SNOMED-CT is not just a coding system but instead an ontology constructed around 19 main domains from which hierarchical concepts are organized, including multiple hierarchical relationships among them. More than 300 000 concepts are modelled using more than 1 000 000 relations.

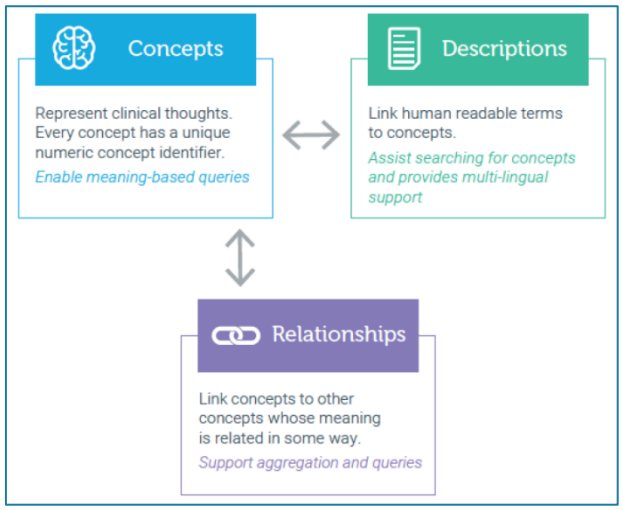


Figure 13: The 3 components of the SNOMED-CT ontologies

Major hierarchies found in SNOMED-CT (out of the 19 groups):

- Clinical finding: disorders, symptoms / signs
- Procedure: surgical procedures, exams, lab tests, nursing care, management procedures
- Body structure: systems, tissues, organs
- Observable: height, weight, blood pressure
- Pharmaceutical/biological products: antibiotics, vitamins, hormones, anesthetics
- Specimen: blood, urine, biopsy specimen
- Organism: bacteria, virus, animal, plant
- Substance: biological / chemical substance, plasma, protein
- Environment or geographical location: countries, languages, hospital, department, clinics, community environment

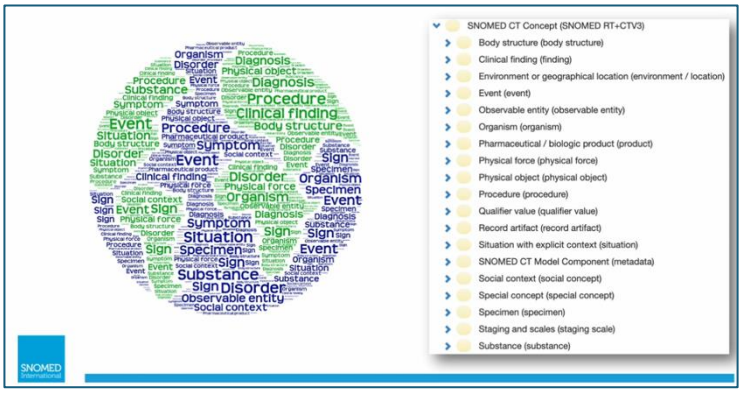


Figure 14: SNOMED 19 Hierarchies of concepts

SNOMED-CT is used to encode lab test results for:

- Culture based results (positive, negative)
- Micro-organisms identification (full name or present/absent)
- Numerical operators (<, <=, =, >=, >)
- Antibiotic susceptibility testing results (sensitive, resistant)

As well as observations derived from test results.

Here are a few examples of codes:

Text to be coded	SNOMED CODE	SNOMED TEXT
Enterococcus faecalis	78065002	Enterococcus faecalis
S	131196009	Susceptible
<=	4171754	<=
Detected ***(from a detection panel)***	260373001	Detected

2.4.3. OMOP Vocabularies and relationship to LOINC and SNOMED

The OMOP dictionary of codes (identifiers) is a meta vocabulary that references a number of standardised vocabularies, such as LOINC vocabulary for laboratory tests, SNOMED vocabulary for Lab results and clinical observations, RxNorm for drugs as prescriptions, etc…

Therefore, each code in the OMOP dictionary is linked to a code in a specific standardised vocabulary; when a code is not considered to be standard, it is registered as non-standard and allows for future standardisation.

Each code (also called concept) belongs to a specific domain, such as condition, gender, measurement, payer, specimen, to list a few, within the 32 current standard domains. The relationships between these concepts is also maintained in the vocabulary, allowing for instance to keep track of hierarchies, as they may exist in the original vocabularies.

To ensure data harmonisation, every code used by the “local” Data provider must be converted to a specific OMOP code. For instance, the code of an antibiotic test should be mapped to the OMOP-CDM code which belongs to the measurement domain; in that case, it is a LOINC code.

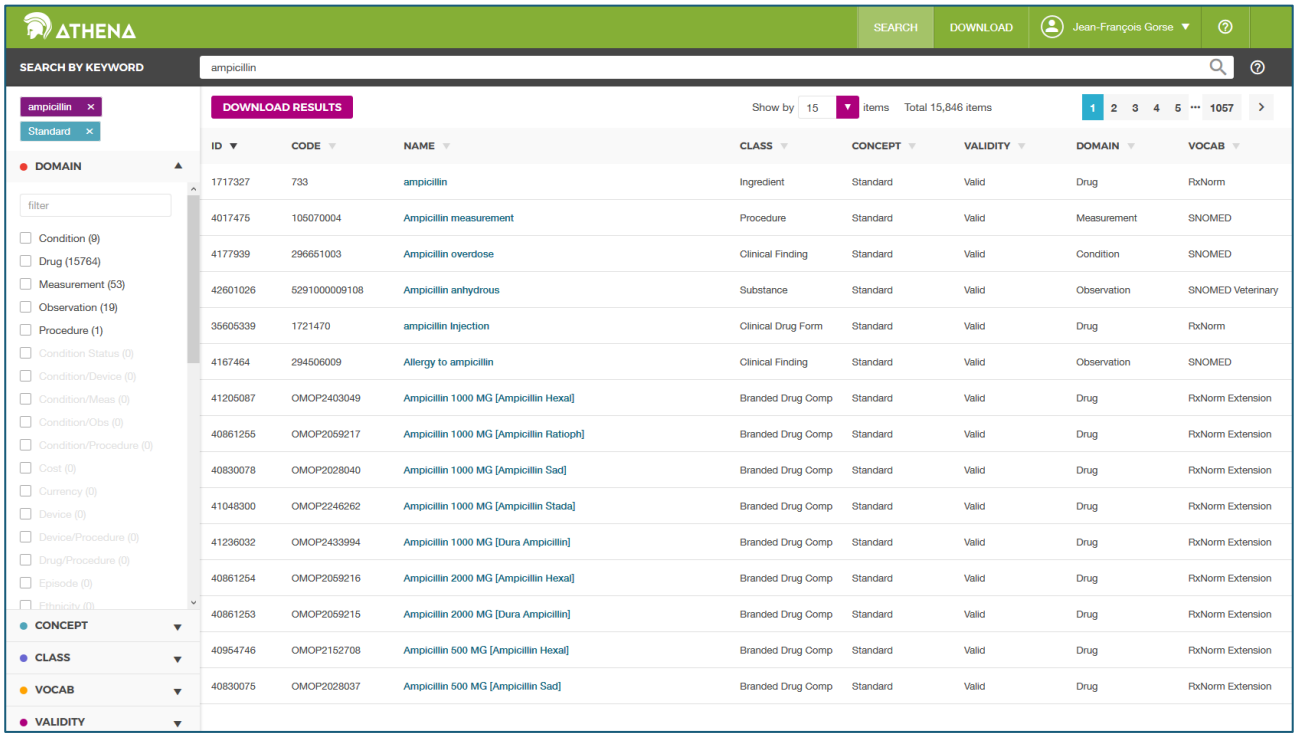
If the data provider has already built a mapping table from his local codes to LOINC, this table can be used to quickly build the mapping to OMOP.

If the data provider does not own a mapping table between its local codes and LOINC codes, the mapping is still possible using OHDSI tools.

The same situation occurs with results; if the user does not benefit from a mapping table between its micro-organisms code and SNOMED, the same OHDSI tool can be used to help the mapping.

A web server is available to get access, navigate and download through the meta dictionary, at https://Athena.ohdsi.org.

In the following example, the text “Ampicillin” can be related to more than 15000 concepts in the OMOP dictionary:



ID	CODE	NAME	CLASS	CONCEPT	VALIDITY	DOMAIN	VOCAB
1717327	733	ampicillin	Ingredient	Standard	Valid	Drug	RxNorm
4017475	105070004	Ampicillin measurement	Procedure	Standard	Valid	Measurement	SNOMED
4177939	296651003	Ampicillin overdose	Clinical Finding	Standard	Valid	Condition	SNOMED
42601026	5291000009108	Ampicillin anhydrous	Substance	Standard	Valid	Observation	SNOMED Veterinary
35605339	1721470	ampicillin Injection	Clinical Drug Form	Standard	Valid	Drug	RxNorm
4167464	294506009	Allergy to ampicillin	Clinical Finding	Standard	Valid	Observation	SNOMED
41205087	OMOP2403049	Ampicillin 1000 MG [Ampicillin Hexal]	Branded Drug Comp	Standard	Valid	Drug	RxNorm Extension
40861255	OMOP2059217	Ampicillin 1000 MG [Ampicillin Ratioph]	Branded Drug Comp	Standard	Valid	Drug	RxNorm Extension
40830078	OMOP2028040	Ampicillin 1000 MG [Ampicillin Sad]	Branded Drug Comp	Standard	Valid	Drug	RxNorm Extension
41048300	OMOP2246262	Ampicillin 1000 MG [Ampicillin Stada]	Branded Drug Comp	Standard	Valid	Drug	RxNorm Extension
41236032	OMOP2433994	Ampicillin 1000 MG [Dura Ampicillin]	Branded Drug Comp	Standard	Valid	Drug	RxNorm Extension
40861254	OMOP2059216	Ampicillin 2000 MG [Ampicillin Hexal]	Branded Drug Comp	Standard	Valid	Drug	RxNorm Extension
40861253	OMOP2059215	Ampicillin 2000 MG [Dura Ampicillin]	Branded Drug Comp	Standard	Valid	Drug	RxNorm Extension
40954746	OMOP2152708	Ampicillin 500 MG [Ampicillin Hexal]	Branded Drug Comp	Standard	Valid	Drug	RxNorm Extension
40830075	OMOP2028037	Ampicillin 500 MG [Ampicillin Sad]	Branded Drug Comp	Standard	Valid	Drug	RxNorm Extension

Figure 15: Ampicillin term found in OMOP dictionary (from https://athena.ohdsi.org)

The concepts can be found in multiple domains such as Condition, Drug, Measurement, Observation, Procedure, and at the same time in 7 different standardised vocabularies (LOINC, RxNorm, SNOMED …). Therefore in our case, for lab testing we must use concepts belonging to the measurement domain which, for antibiotic tests, implies that a concept related to an element belonging to the LOINC vocabulary must be selected.

The same can be observed if we were to look up for *Escherichia coli*, with more than 5000 concepts that could be related in various domains and vocabularies.

At the end of the vocabulary mapping process, the OMOP data base will be populated with fully standardised data representation (model) and data codes (vocabularies). These codes will represent the same concepts whichever OMOP data node is connected to the network.

2.5. Data security and Privacy

For the POC1 implementation, all data will be fake data and therefore transparent to the GDPR.

For the next implementation steps, the risk of privacy will be reduced due to the federated network architecture; it will be the responsibility of the data owner to ensure compliance.

3. POC1 implementation

Three main tasks will be executed during the implementation phase:

1. Experimental process to build the OMOP Node database from the Data source
2. Implement the federated network architecture and tools
3. Implement data queries to the Data nodes from the Data Observatory

This report will detail the first section of the POC1 implementation, which itself can be broken down into four parts (the light blue boxes in the attached figure)

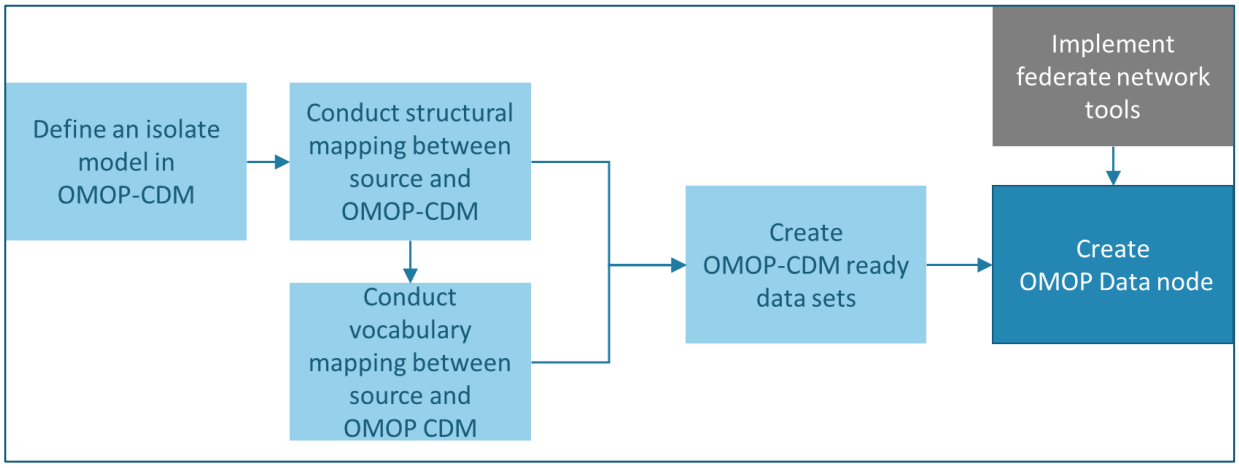


Figure 16: OMOP Data node preparation

3.1. Defining isolate and AMR data modelling into OMOP-CDM

As mentioned in the previous chapter, clinical data networks are patient-centric and therefore all the data is always organized around the patient for whom the laboratory tests data for AMR have to be associated. Depending on the data source for the implementation, the patient data may not be available but instead the data relative to the sample collected from the patient may be accessible; this may become a limitation, since duplicate tests for the same patient may not be recognized at this level.

Modelling an Isolate into OMOP-CDM from and EMR, LIS, or M/W view, a lab results may be represented by the following hierarchy of concepts: Each patient admission in the institution is considered as a visit. During the visit one or more specimen could be sampled for the patient, upon which testing is performed.

Unlike other testing in the lab, a microbiology test is not binary such as a question leading to one answer. A culture may end up being negative (this is binary), but if the culture is positive, additional tests are performed. These include identification of one or more isolates for some, or for all, AST testing with more than a dozen of antibiotics are tested. This tricky situation of cascading testing has been a roadblock for many IT systems in the past.

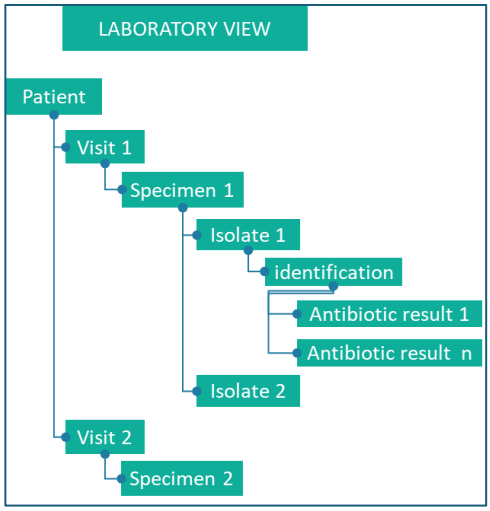


Figure 17: Typical representation of a patient and its AST results

Quite often, multiple isolates may be found during lab work, although not all of them may be reported since they could either be considered as duplicates (same identification and same AST pattern), or considered as contaminants present in the specimen but introduced during sampling at the patient (or at any other subsequent step). The Diagnostic system may keep track of all the testing data, although not all should be considered as relevant for further data analysis. Some of this data may be discarded at the Middleware level (if in use in the lab) or later in the LIS.

If we consider the OMOP-CDM data model, a few database tables can be populated with the Lab results data.

In the OMOP MODEL VIEW figure (figure 18), the links between tables is indicated. Two tables can be populated with test results, the table measurement and the table observation. However, it is noticeable that no explicit link is established between the measurement and observation table and the specimen; measurement and observations are directly linked to the patient.

Here again, the concept of binary coding (one test, one answer) is visible, each measurement (and its associated result) is independent from the other ones.

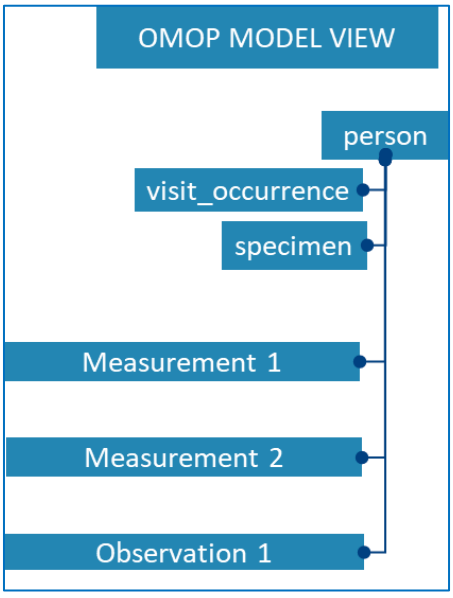


Figure 18:OMOP Tables to be used for Lab Data

In order to be able to transfer the lab data from the "laboratory model" to the OMOP-CDM model, it is necessary to rely on another concept called fact relationship, which establishes the link between the concepts that need to be related to each other. Relationships can be established between specimen and measurement, as well as between measurements themselves.

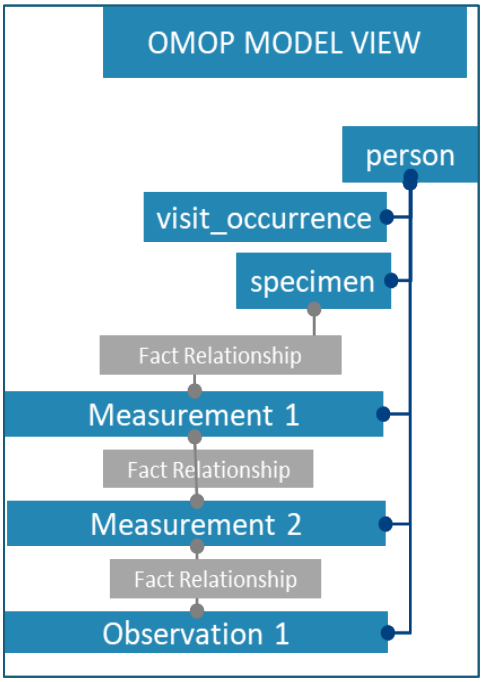


Figure 19: relationship between OMOP-Tables

3.1.1. Selecting the OMOP-CDM tables to collect the results

Since the hierarchy of dependencies can be established, four issues remain to be addressed.

1. Which table should be used between Measurement and Observation?
2. What is the best modelling of the hierarchy?
    a. Modelling of the cultures (allows to code for negative results)?
    b. Modelling of Isolates (only tracks positive results)?
3. Can we combine multiple result types into a single measurement?
   AST tests can generate numerical MIC[14] results along with clinical category results (such as S, I, or R).
4. Can we ensure a consistent modelling between ID/AST testing and detection panel testing?

Measurement and observation

Since OMOP-CDM has to ensure that all data providers of the network provide interoperable data sets, a strict set of semantic rules has to be applied to data coding.

The semantic is enforced by data domains from which the user has to select the vocabulary to be used.

For example, data codes to be used for measurements should come from codes related to the measurement domain. One underlying vocabulary for the measurement domain is LOINC.

14 MIC : Minimum Inhibitory Concentration

Data codes to be used for observations should come from the observation domain. SNOMED is one of the vocabularies included into the observation domain.

This strict semantic rule, aimed at enforcing interoperability, supposedly prohibits the ability to store a test and some of its results in the same measurement, since only codes from the measurement domain can be used. For microbiology results however, organism codes derived from SNOMED belong to the observation domain and must be used to report identification test results. This constraint may only impact microbiology results modelling.

Further discussions with the OHDSI community will occur, with the intent to clarify this constraint.

Modelling scenarios

Three different modelling techniques have been experimented to represent both culture type results (with ID and AST) and results from detection panels (multiplex tests).

1. Model 1: Isolates view

    a. A "root" measurement is used to capture the isolate identification; it includes the code for the identification method and the code of the organism being found.
    b. As many measurements as antibiotic tests are linked to the "root measurement". Each of those contains both the code for the antibiotic and the code for the category result
    c. As many measurements as antibiotic tests are linked to the "root measurement". Each of those contains both the code for the antibiotic and the MIC value.

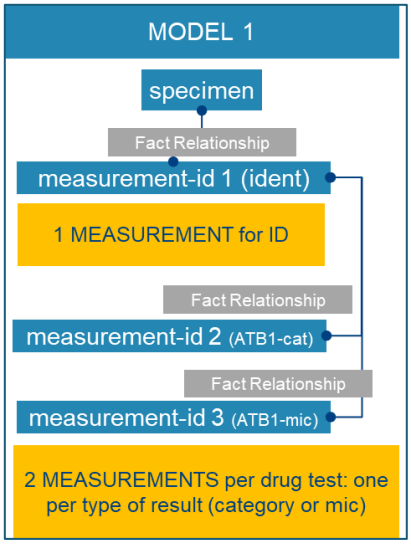


Figure 20: Model 1 representing isolates

NOTE: For measurements listed in b), the associated OMOP code is different from the OMOP code found in the measurements listed in c). This is due to the construction rule of the LOINC code that takes the nature of the result into account, here ordinal value for categories as opposed to numerical values used for MICs. As far as the OMOP code is derived from the LOINC code, the associated OMOP code is different for the two measurements.

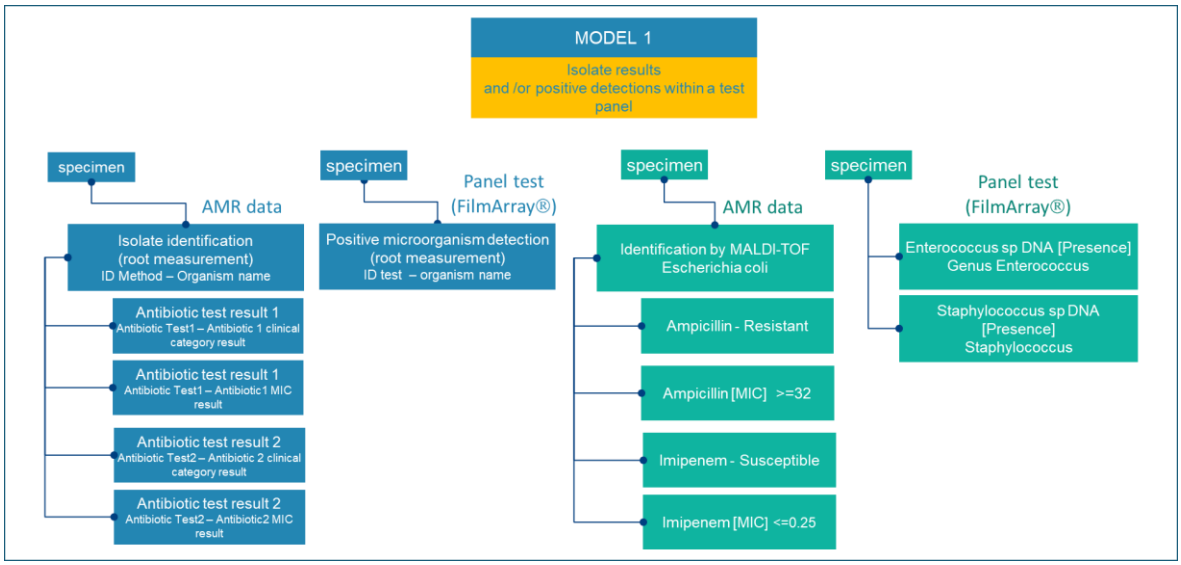


Figure 21: Model 1 and examples for AMR data and detection panel data

2. Model 2: Cultures view
    a. A “root” measurement is created to capture the culture; it includes the code for culture and the result as positive or negative.
    b. An observation is created to capture the result of the identification test. This observation is linked to the “root measurement”. This observation becomes the isolate node.
    c. As many measurements as antibiotic tests are linked to the “root measurement”. Each of those contains both the code for the antibiotic, and the code for the category result.
    d. As many measurements as antibiotic tests are linked to the “root measurement”. Each of those contains both the code for the antibiotic and the MIC value.

NOTE: For measurements listed in c) the associated OMOP code is different from the OMOP code found in the measurements listed in d). This is due to the construction rule of the LOINC code that takes the nature of the result into account, here ordinal value for categories as opposed to numerical values used for MICs. As far as the OMOP code is derived from the LOINC code, the associated OMOP code is different for the two measurements.

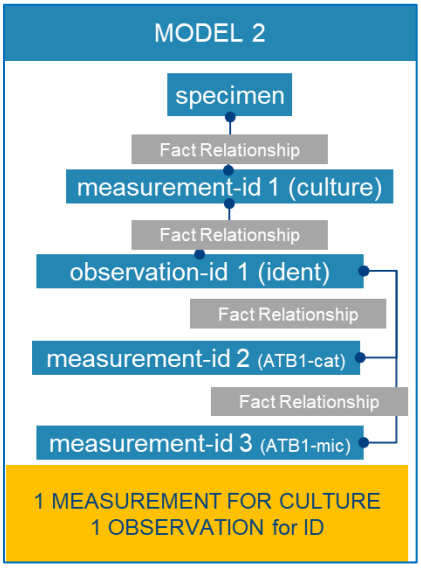


Figure 22: Model 2 representing cultures

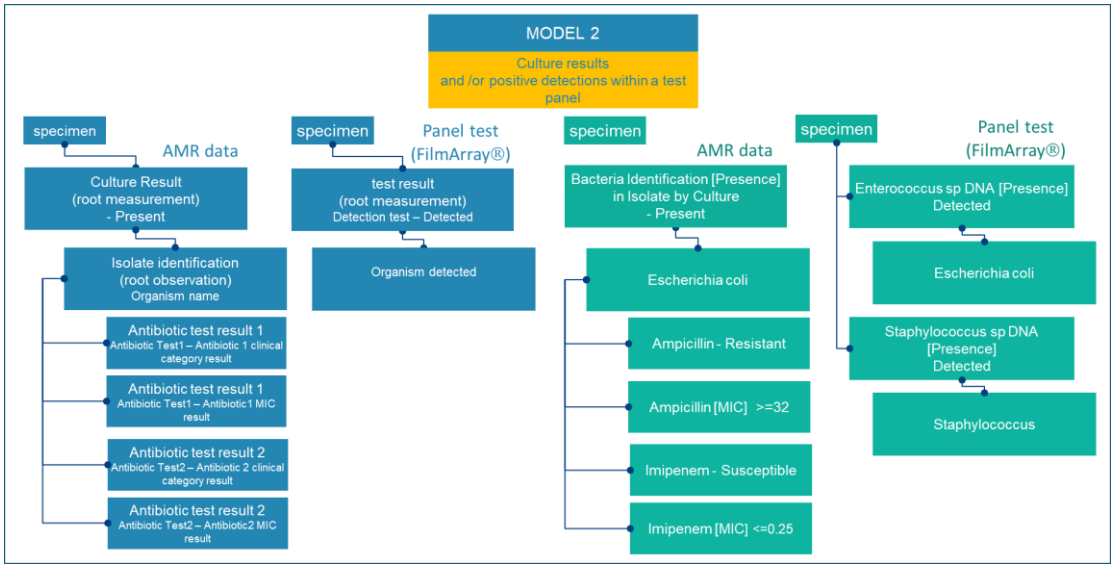


Figure 23: Model 2 and examples for culture results and detection panel

3. Model 3: Simplified Isolates view

a. A “root” measurement is used to capture the isolate identification, it includes the code for the identification method and the code of the organism being found
b. As many measurements as antibiotic tests are linked to the “root measurement”. Each of those contains both the code for the antibiotic and the code for the category result as well as the MIC result (if applicable).

NOTE: For measurements listed in b), the associated OMOP code is either the code corresponding to the LOINC code for MIC results, if MIC results are provided, or the other code if only category results are recorded.

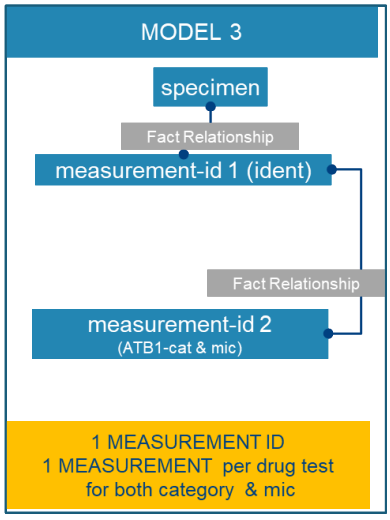


Figure 24: Model 3 representing a simplified isolate model

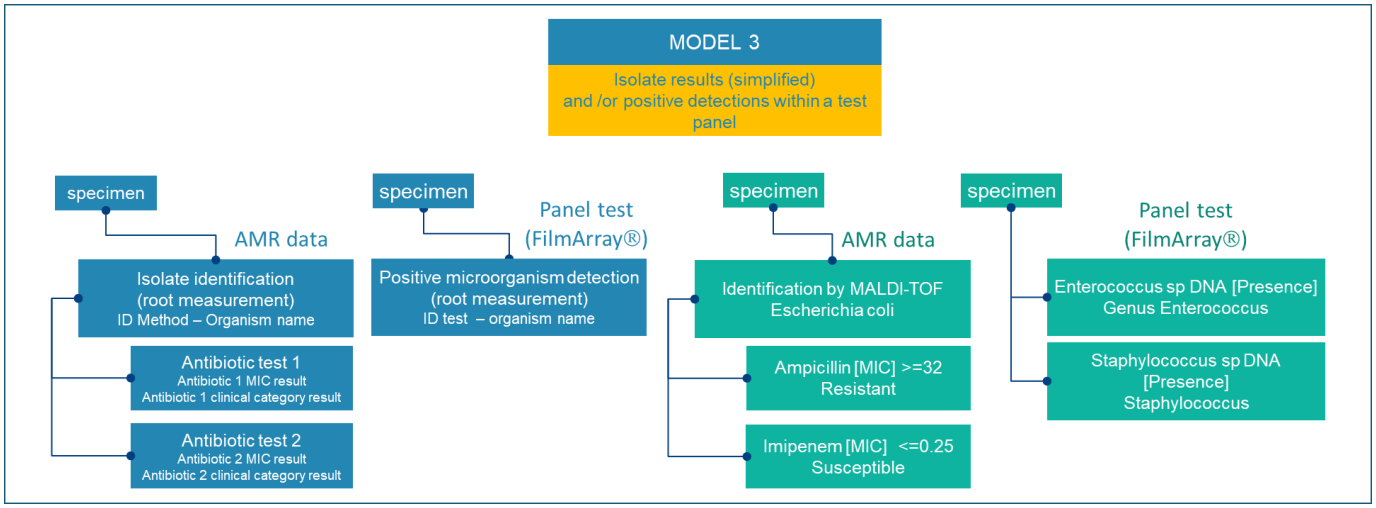


Figure 25: Model 3 with examples for AMR isolate and test panels

3.2. Conducting the Structural mapping

The structural mapping consists of linking source data element structure (table and /or fields) to destination structure (OMOP-Tables and fields).

In order to map the data source model to the OMOP model, OHDSI open source tools are available.

A tool named White Rabbit scans the source of data (database or csv files) and delivers a scan report where all fields of all tables (or files) are inspected.

Following the scan another tool, named Rabbit-in-a-Hat, reads the scan and through a graphical user interface; this tool allows to link the data source fields to the OMOP-CDM table fields.

In the following example, one large data extract (csv text file) from a middleware is used as a data source to be structurally mapped to the relevant OMOP-CDM tables.

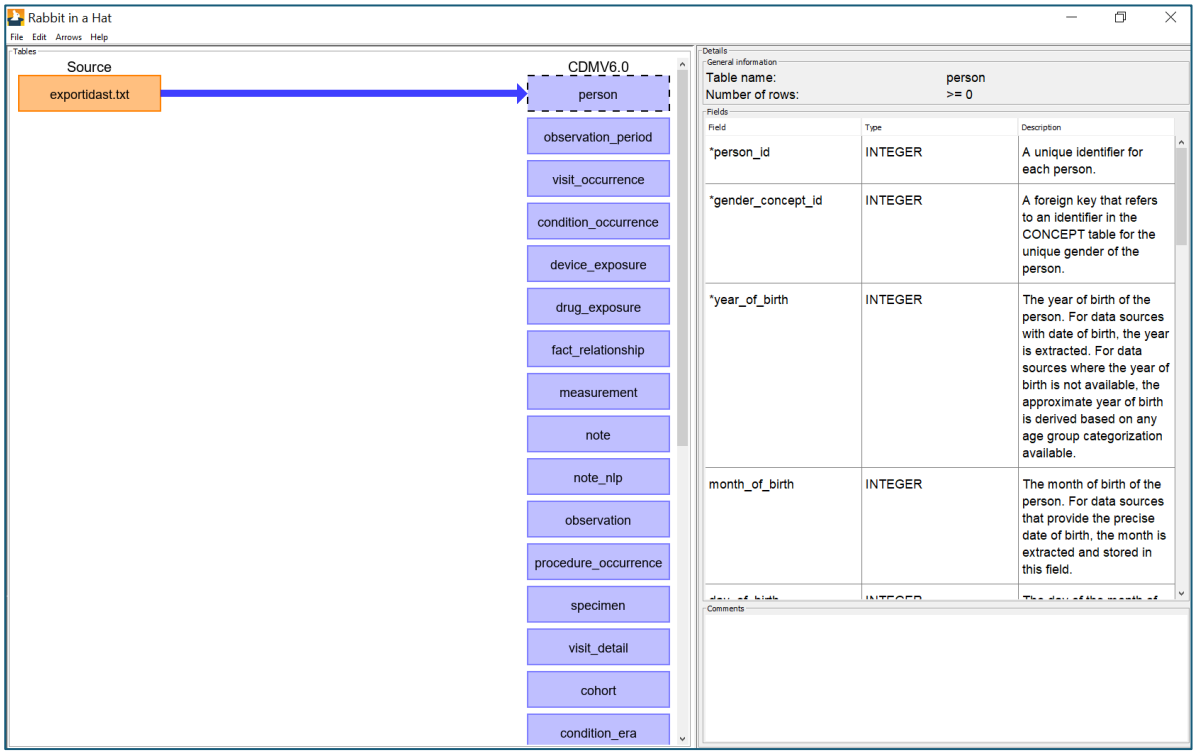


Figure 26: Linking a source data (here a single file) to OMOP Tables

Each column of the source file represents a field that has to be linked to an OMOP-CDM field (or multiple) using the mouse.

In the example below, the field named “sexe” from the source file will be used to populate two fields in the OMOP-person table (corresponding to the patient): the screen displays at the top right part of the screen the actual values for this field that have been found in the source data along with their frequency.

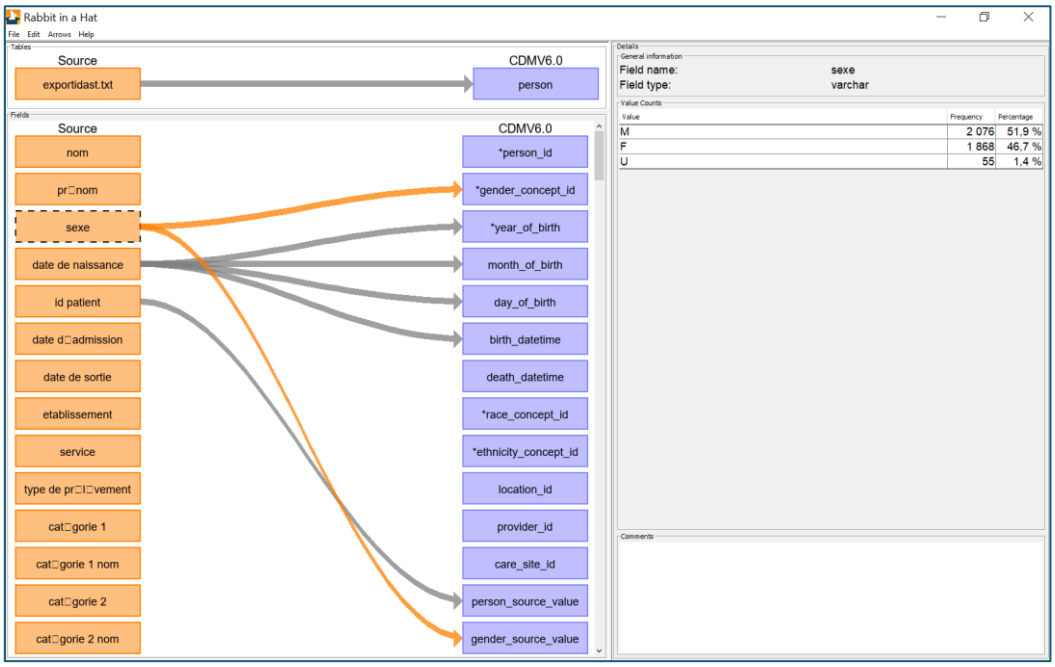


Figure 27: Structurally linking a source data field to an OMOP destination field

By clicking on the OMOP corresponding field, the possible codes to be used are displayed (therefore, avoiding access to Athena for possible codes look-up!).

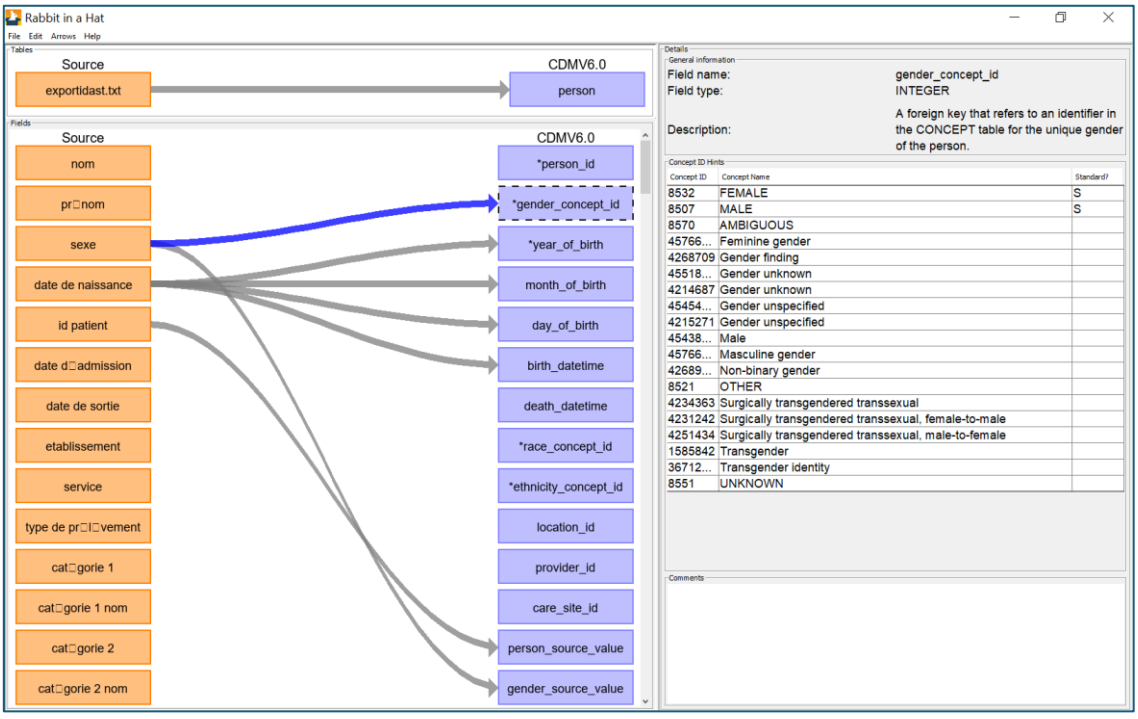


Figure 28: List of possible codes to be used for that field

The comments area at the bottom of the screen is to enter the mapping rules to be used later at the time of converting the data.

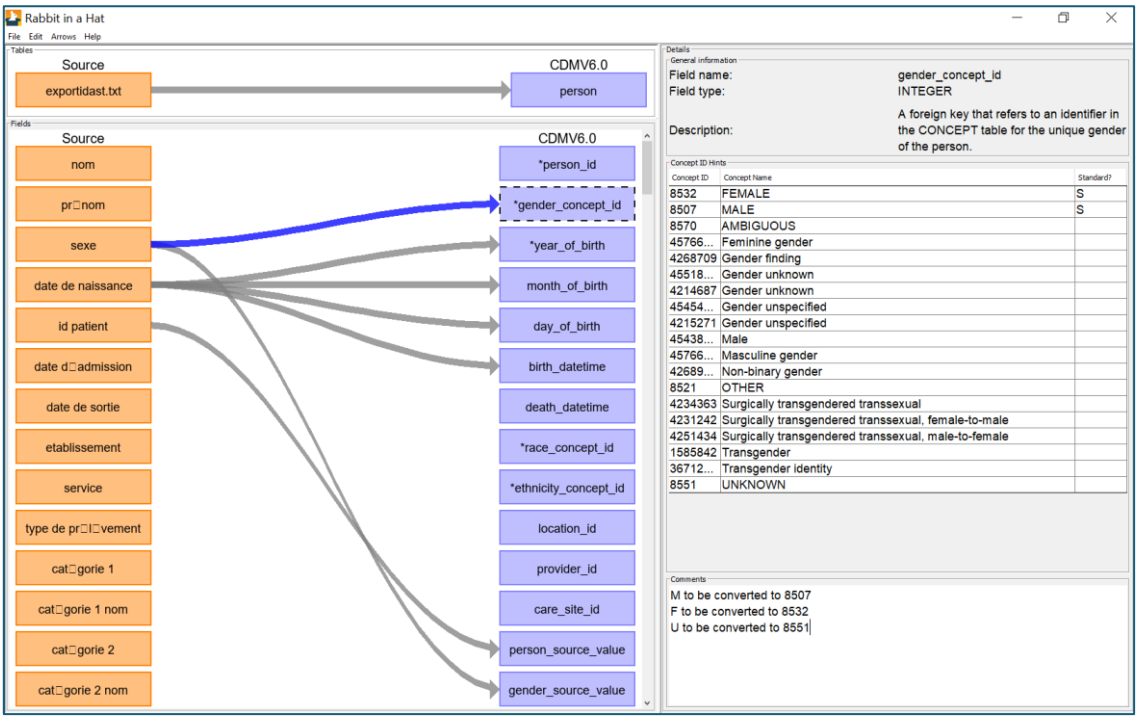


Figure 29: Introducing code mapping rules in Rabbit-in-a-Hat

During our preparation for POC1, we conducted the structural mapping and were able to split the data source packet contained into a single text file, and into the major structures required by OMOP-CDM. A specific data structure was created, ISOLATE TESTS, in order to be in a position to prepare the three potential models that were described in a previous paragraph.

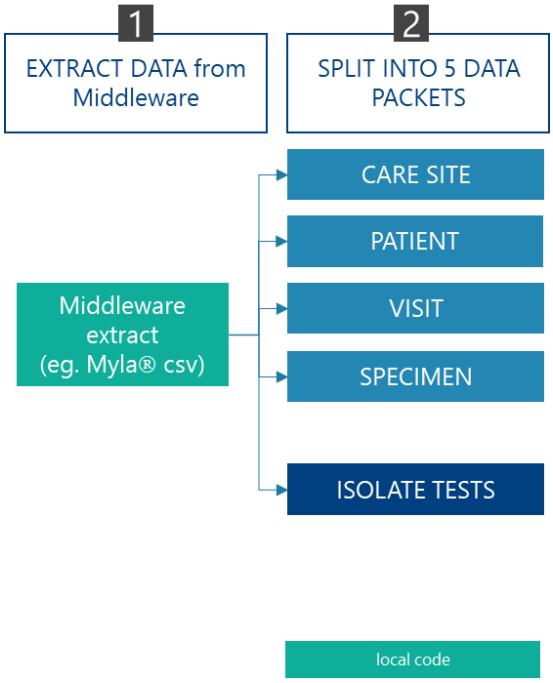


Figure 30: Structural mapping

Thanks to the OHDSI Tool (Rabbit-in-a-hat) we were also able to directly get a few OMOP codes that were required into the OMOP prepared data packets:

PATIENT(=person in OMOP tables)

- gender_concept_id
- race_concept_id (OMOP 8552 : “unknown”)
- ethnicity_concept_id (OMOP 8552 : “unknown”)

VISIT (visit_occurence in OMOP tables)

- visit_concept_id (OMOP 8717: “Hospital In patient”)
- visit_type_concept_id (OMOP 44818518: “visit derived from EHR record”)

SPECIMEN (specimen in OMOP tables)

- specimen_concept_id (default value OMOP 4002873: “Specimen of unknown material”)
- specimen_type_concept_id (OMOP 581378: “EHR detail”)

MEASUREMENT (to be used later in the OMOP table measurement)

- measurement_type_concept_id (OMOP 44818702: “Lab result”)
- operator_concept_id (a set of OMOP codes for {“<”,”<=”,”=”,”>”,”>=”}

OBSERVATION (to be used later in the OMOP table observation)

- observation_type_concept_id (OMOP 581413: “Observation from measurement”)

3.3. Conducting the vocabulary mapping

The OMOP-CDM model requires all data provided by the data provider to be mapped with OMOP-CDM standardised dictionary. On top of this mapping, OMOP-CDM includes additional data fields that may not exist in the data source but are still required to be populated by standard OMOP codes. This data includes types of patient visits, types/origin of measurements, types of observations.

All the tests performed, and their associated results, need to be mapped to OMOP-codes as well.

As mentioned in chapter 2.4.3, multiple means are available to successfully perform all the mapping.

During the structural mapping process, the White Rabbit and Rabbit-in-a-Hat OHDSI tools are preparing the lists of values found in each field of the source data. When source data fields are mapped with OMOP-CDM destination fields, the Rabbit-in-a-Hat is able to propose a list of possible OMOP codes to be used for some of the fields. This applies to data fields where the possible value set is limited. We have seen an example in the previous chapter, with the “sexe” field from the source data.

For fields that can be populated with a large set of values , for instance antibiotic codes or micro-organisms names or specimens, the tools cannot directly propose a list of codes, therefore the mapping needs to be helped with another OHDSI tool, named USAGI, which has the ability to import

lists of local codes and tries to match them automatically with individual concepts in the OMOP dictionary.

In a previous chapter, another OHDSI tool named ATHENA, was presented. This tool allows to browse through the OMOP dictionary and to search for specific terms in order to find the relevant OMOP concept and associated code that were not proposed by Rabbit-in-a-Hat, or that may not belong to a large list of values for instance “Detected” or “Present”.

Three mapping options can be exercised (not exclusive):

1) Getting codes proposed by White Rabbit
2) Getting codes from Athena through searches
3) Semi-automated code mapping with USAGI for larger lists

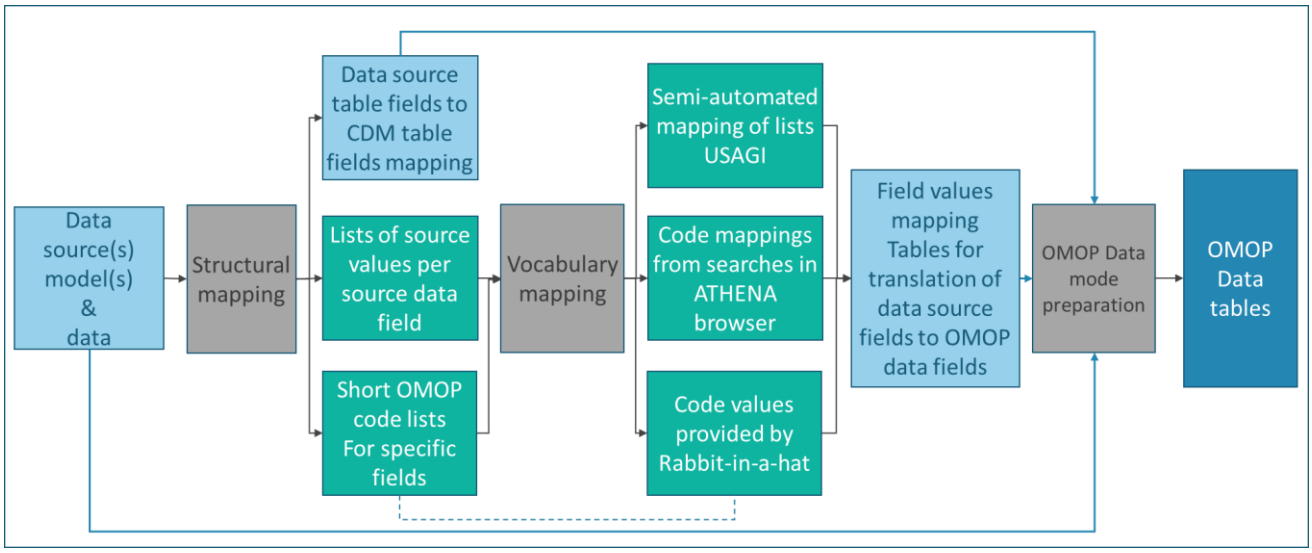


Figure 31: Overview of structural and vocabulary mapping

In order to successfully harmonise AMR related data from multiple sources, a set of concepts related to such lab results have to be mapped to the OMOP dictionary.

- Codes or text for specimen types/nature
- Codes or text for antibiotic tests
- Codes or text for micro-organisms
- Codes for clinical categories {S,I,R} or {Susceptible, Intermediate, Resistant}
- Codes for MIC operators {<,<=,=,>=,>} (already identified during the structural mapping)
- Codes for panel test answers {present, absent} {positive, negative}

- In addition to this minimum set, a number of other concepts may need to be mapped in order to comply to the OMOP-CDM tables layout, since the laboratory results data may be merged in the OMOP Model with other measurements performed at the bed side as well as observations that the clinician can make directly from the patient status. A few codes are therefore necessary to allow for sorting the data during the data analysis. One of these codes has been listed in the

previous section: measurement_type_concept_id, for which the value OMOP 44818702: “Lab result” will be assigned to all measurements that we are going to create from our data set.

3.3.1. Mapping codes provided by Rabbit-in-a-Hat

OMOP codes relative to the main PERSON (patient), such as sex or ethnicity, VISIT (visit_occurence) were obtained directly from Rabbit-in-Hat while the structural mapping was performed.

OMOP codes relative to measurement operators were proposed by Rabbit-in-Hat when making the structural link between our source data MIC results and the OMOP-CDM MEASUREMENT table:

- “<” OMOP 4171756
- “<=” OMOP 4171754
- “=” OMOP 4172703
- “>” OMOP 4172704
- “>=” OMOP 4171755

3.3.2. Mapping codes obtained by search in OHDSI tool ATHENA

When it comes to map codes where the value set is limited (half of a dozen of items), the OHDSI ATHENA browser tool can be leveraged. The codes that are obtained from searches can be entered into a mapping table that will be utilised during the preparation of the remaining OMOP tables related to the isolate results (table MEASUREMENT and OBSERVATION).

OMOP codes relative to clinical category results were obtained through a search using the Measurement Value domain associated to the SNOMED vocabulary.

- S or susceptible = OMOP 4038110: “Susceptible”
- R or resistant = OMOP 4148441: “Resistant”,
- I or Intermediate could be represented by multiple codes
    - OMOP: 4137479: “Intermediately susceptible”
    - OMOP: 4123511: “Moderately resistant”
    - OMOP: 4126676: “Moderately susceptible”
    - OMOP: 4043352: “Intermediate” from Observation domain.

The latter code was used, although another code may be better suited for data analysis.

Since we intend to challenge AMR data modelling by exercising three options for isolate data representation, additional codes were necessary to be found such as OMOP codes relative to presence of bacteria in culture or detection of micro-organism, which are mandatory in order to represent the “culture” step in the Model 2.

These codes were obtained through a search using the measurement Value domain associated to the SNOMED vocabulary:

- “Present”: OMOP 4181412: “Present”
- “Detected”: OMOP 260373001: “Detected”

Model 1: The underlying proposition is to capture only isolate results, meaning that for AST tests a “root measurement” would represent the identification (the isolate), and related measurements would represent antibiotic test results for this isolate. For a detection test panel, there would be as many measurements as positive tests available on the panel.

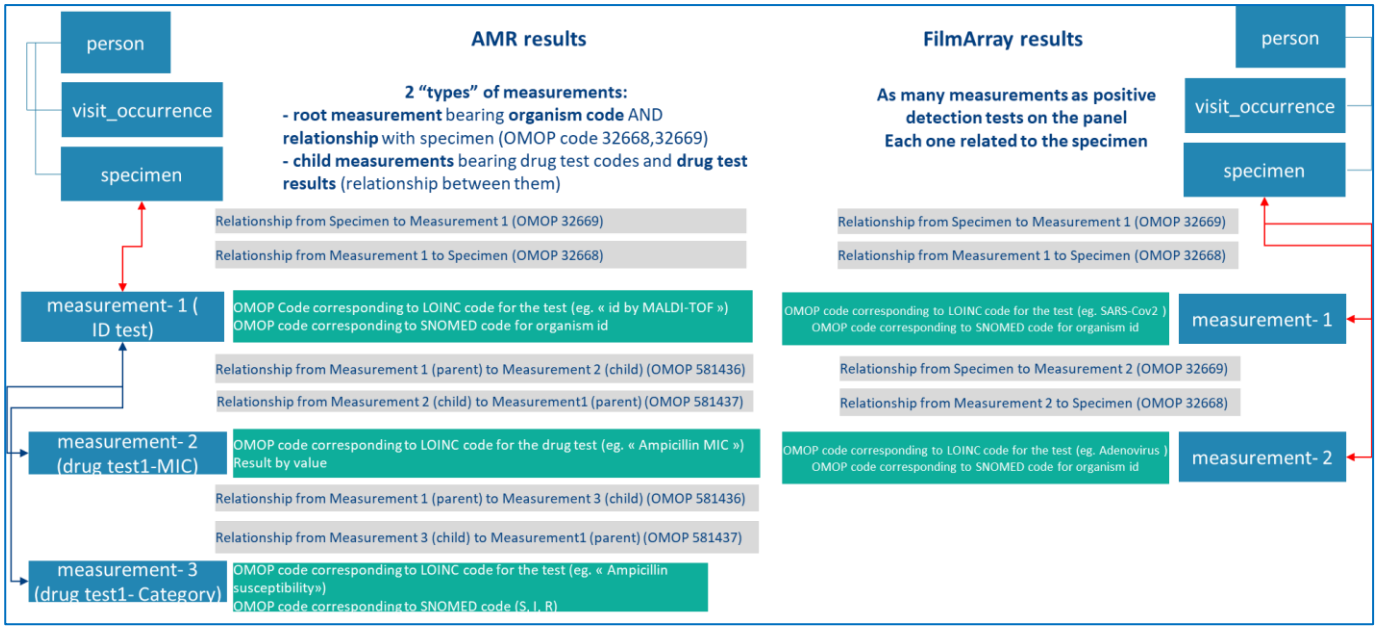


Figure 32: Mapping used for Model 1 (Isolate based)

Coding for AST tests:

- Root measurement linked to the specimen
    - Measurement_concept_id= OMOP 46235829: “Microorganism identified in Isolate by MS.MALDI-TOF”
    - Value_as_concept_id = OMOP code corresponding to a SNOMED code for the micro-organism identified by MALDI-TOF, for that field a large value set will be necessary to be mapped, another tool (USAGI) will be used to identify all relevant OMOP codes) to prepare the mapping table.
- Antibiotic test measurements linked to the “root” (isolate) measurement
    - Measurement_concept_id= OMOP code for the drug test corresponding to a LOINC code, for that field a large value set will be necessary to be mapped, another tool (USAGI) will be used to identify all relevant OMOP codes) to prepare the mapping table.
    - Value_as_concept_id = the OMOP code for the clinical category
    - Operator_concept_id = the OMOP code for the operator associated to the MIC value
    - Value_as_number = the MIC value

      Note: In Model 1, a drug test MEASUREMENT bears either the clinical category OR an MIC with its operator. Therefore two MEASUREMENTS are used for a single drug test, when both results are available.

Coding for Detection tests:

All measurements are structured the same way. There are as many measurements as positive tests, all establishing a direct relationship to the specimen.

- Measurement_concept_id= OMOP corresponding to the LOINC code for the positive test.
- Value_as_concept_id= OMOP code corresponding to the SNOMED code for the organism detected by the test.

Model 2: The underlying proposition is to represent the result of a culture for an AST test, a "root measurement", will capture the culture result as "positive", an observation attached to it will represent the isolate and its identification, a set of measurements will be attached to this observation to capture antibiotic test results. For a detection test, each measurement will be linked to the specimen and would represent the detection test either "Detected" or "Undetected", one observation would be linked to one measurement when then test is positive and would represent the organism detected.

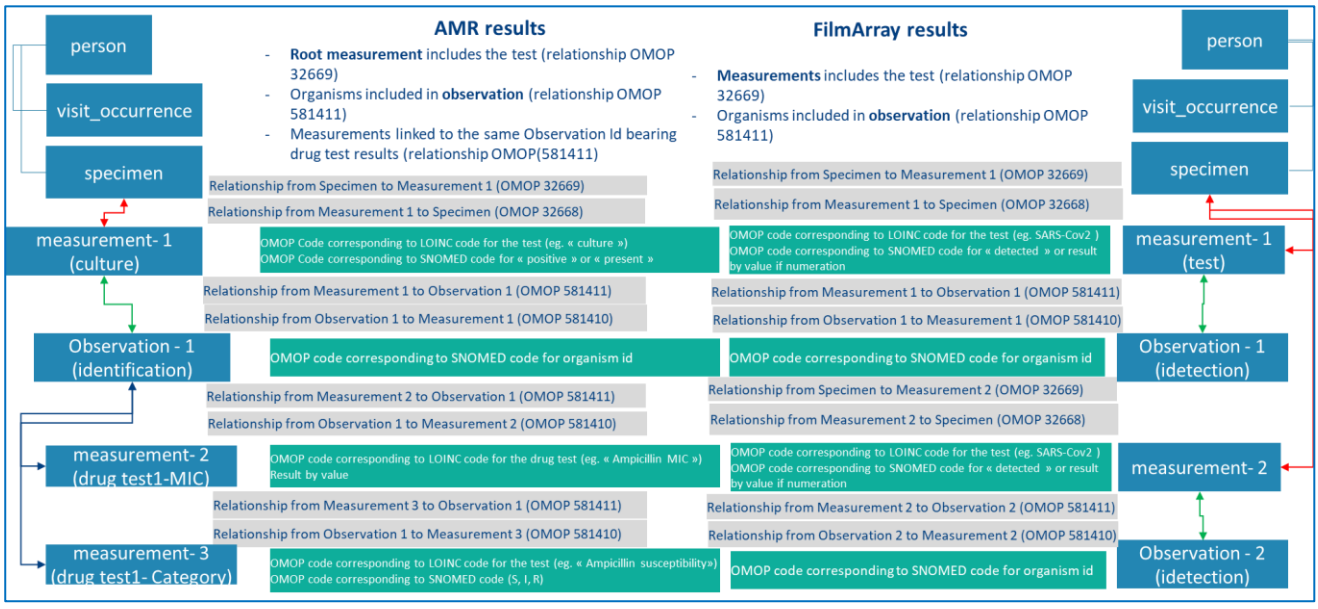


Figure 33: Mapping for Model 2 (culture / test based)

Coding for AST tests:

- Root measurement
  - Measurement_concept_id= OMOP 3044054: "Bacteria Identification [Presence] in Isolate by Culture"
  - Value_as_concept_id = OMOP 4181412: "Present" when present (meaning positive culture), OMOP: 4132135: "Absent" when negative.
- Observation
  - Observation_concept_id= OMOP code for the organism corresponding to a SNOMED code, for that field a large value set will be necessary to be mapped, another tool

(USAGI) will be used to identify all relevant OMOP codes) to prepare the mapping table.

Coding for Detection tests:

- Measurement
  - Measurement_concept_id= OMOP code corresponding to the test performed (mapped to a LOINC code)
  - Value_as_concept_id= OMOP code for "Detected" or "Undetected"
- Observation
  - Observation_concept_id= OMOP code for the organism detected (when the test is positive), mapped to a SNOMED code

Model 3: The underlying proposition is to represent a simplified isolate view compared to Model 1. For an AST test, a "root measurement" will capture the identification of the isolate, a set of measurements will be attached to this root measurement to capture antibiotic test results, there will be a single measurement per antibiotic test including both MIC (if relevant) and category. For a detection test, each measurement will be linked to the specimen and would represent the detected organism, similar to Model 1.

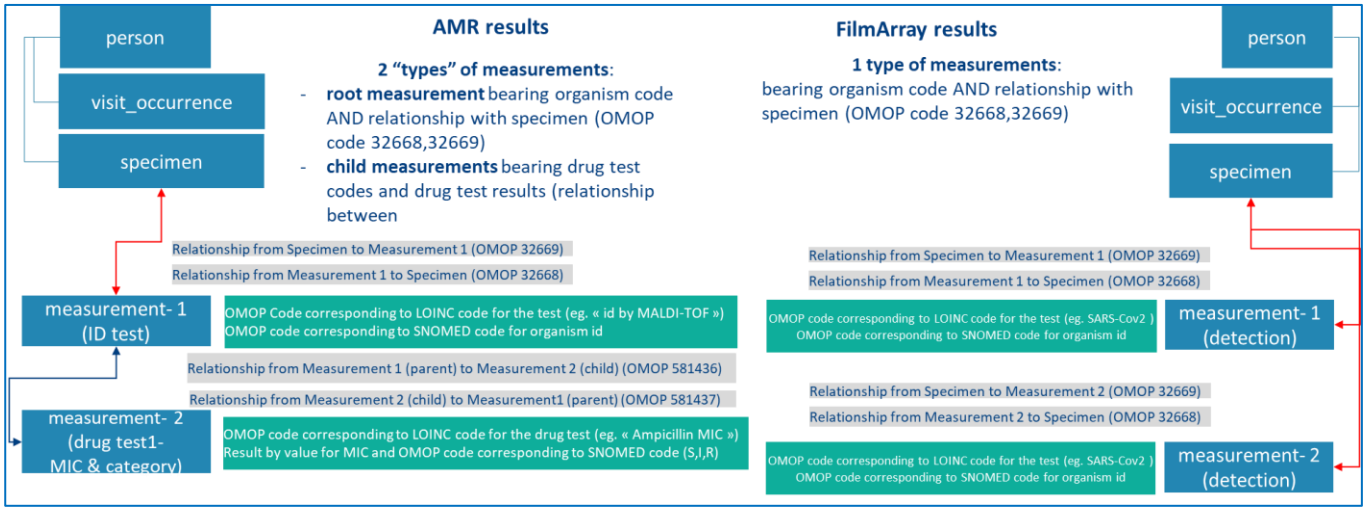


Figure 34: Mapping for Model 3 (Simplified isolate)

Coding for AST tests:

- Root measurement linked to the specimen
  - Measurement_concept_id= OMOP 46235829: "Microorganism identified in Isolate by MS.MALDI-TOF"
  - Value_as_concept_id = OMOP code corresponding to a SNOMED code for the micro-organism identified by MALDI-TOF, for that field a large value set will be necessary to be mapped, another tool (USAGI) will be used to identify all relevant OMOP codes) to prepare the mapping table.
- Antibiotic test measurements linked to the "root" (isolate) measurement
  - Measurement_concept_id= OMOP code for the drug test corresponding to a LOINC code, for that field a large value set will be necessary to be mapped, another tool

(USAGI) will be used to identify all relevant OMOP codes) to prepare the mapping table.

- Value_as_concept_id = the OMOP code for the clinical category
- Operator_concept_id = the OMOP code for the operator associated to the MIC value
- Value_as_number = the MIC value

Note: Both antibiotic test results MIC and category are stored into a single MEASUREMENT per drug test

Coding for Detection tests:

All measurements are structured the same way. There are as many measurements as positive tests, all establishing a direct relationship to the specimen.

- Measurement_concept_id= OMOP corresponding to the LOINC code for the positive test.
- Value_as_concept_id= OMOP code corresponding to the SNOMED code for the organism detected by the test.

3.3.3. Mapping codes obtained by semi-automated process in OHDSI tool USAGI

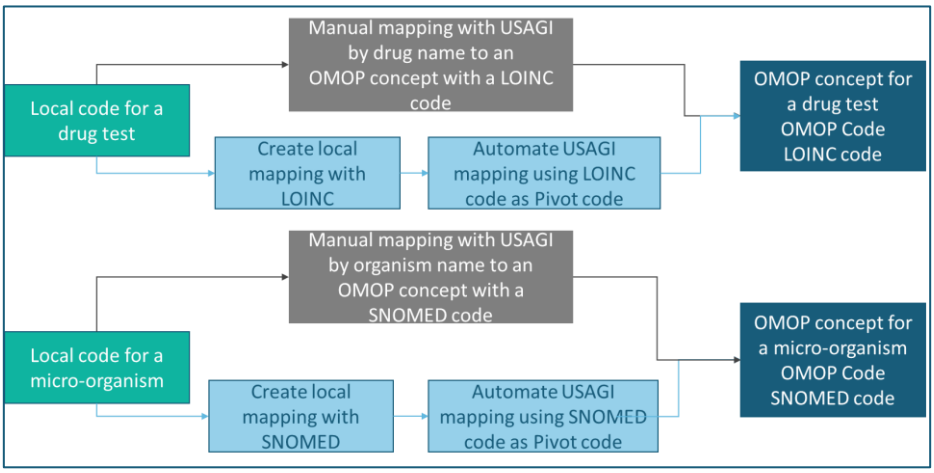


Figure 35: Alternate methods for vocabulary mappings (grey: manual, blue: using pivot codes)

A few fields in AMR can be filled by values belonging to a large list of items. As such, antibiotic tests and micro-organisms fall into this category.

The USAGI tool is used to semi-automatically map a series of local codes into OMOP-CDM codes (or identifiers). The tool performs by matching terms in the local list to their equivalents in the OMOP dictionary. For each match, a score is provided; the highest score, 1, showing the highest probability for an exact match. Nevertheless, it is still recommended to verify each mapping before exporting the final mapping table, which will be used at the time of pushing the data and the OMOP codes into the OMOP database.

Regarding the list of local Antibiotic tests, a previous mapping of the local codes to the LOINC codes will ease the verification process during the OMOP mapping since the OMOP codes for antibiotic tests are themselves linked to LOINC tests.

The mapping process starts with the preparation of the list of local codes and text (the text will be used to be matched to equivalent text in the OMOP dictionary). The text needs first to be translated into English (if applicable).

The lists provided by White_Rabbit can also be used as input.

Mapping Antibiotic tests

If LOINC codes are already available for each local antibiotic test, they should be added as an additional column in the list in order to ease the verification of the automated mapping.

Antibotic label	Local code	LOINC code
Acide fusidique	FA	18927-4
Acide fusidique MIC / Diam	FA	262-6
Acide nalidixique	NA	18952-2
Acide nalidixique MIC / Diam	NA	351-7
Amikacine	AN	18860-7
Amikacine MIC / Diam	AN	12-5

Figure 36: Example of a local Antibiotic test list to be mapped to OMOP

In the above example, it was decided to not translate into English.

USAGI asks for the columns to be used for the text matching as the **source name column** (here the column Antibiotic test), the **source code column** (here the column Local Code) and the Additional. The **Auto concept ID column** which captures pre-existing mapping (not used here). The **additional info column** is used here to display our LOINC mapping.

Before starting the import of the local codes table, it is important to select and filter the concept class as **Lab test**, the vocabulary as **LOINC** and the domain as **Measurement**, by selecting in the lists + check box at the bottom right.

Filtering by the LOINC code vocabulary prevents the mapping of antibiotic tests to drugs present in the RxNorm vocabulary.

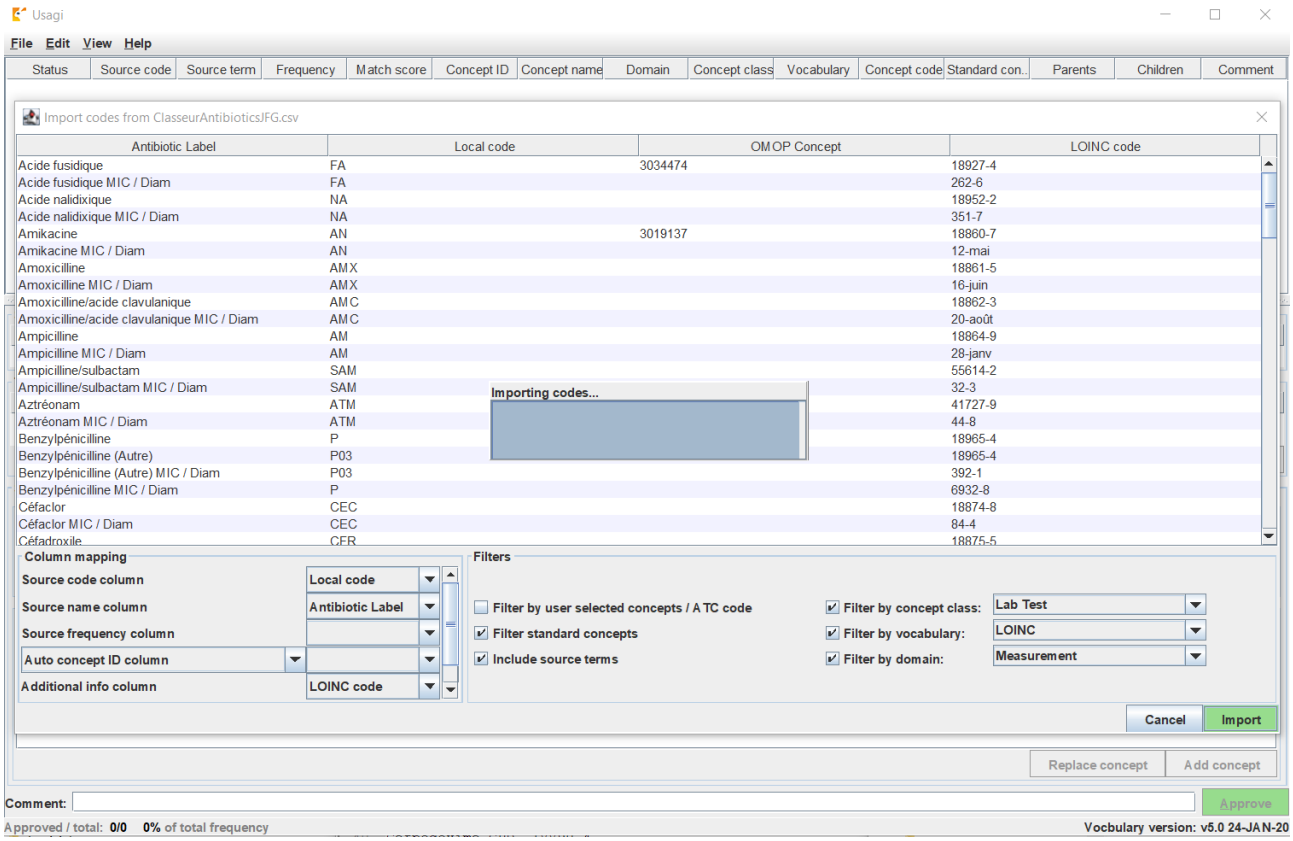


Figure 37: USAGI import of local code table

USAGI then presents the automated mapping that were performed; in case LOINC codes have been provided in the source file, it is easy to verify if the mapping proposed is correct.

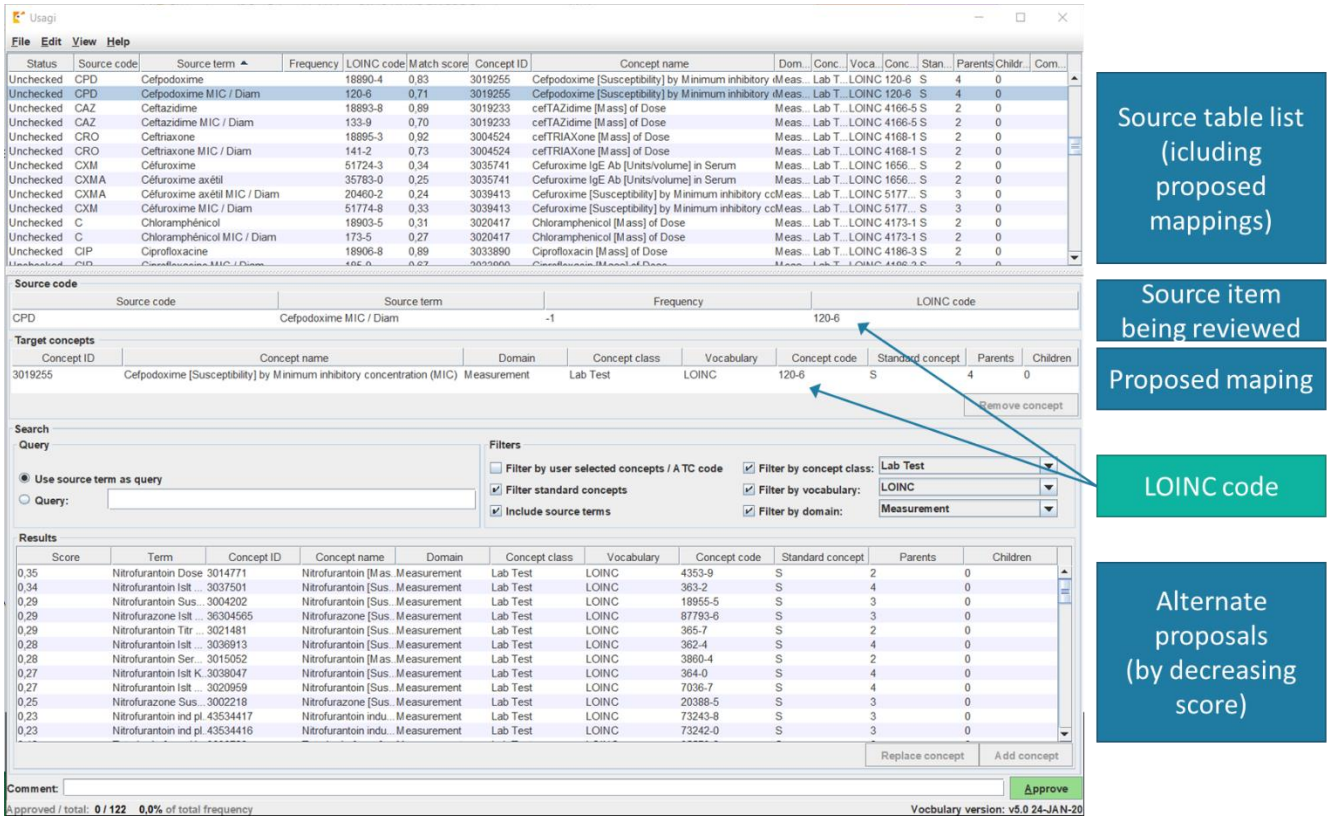


Figure 38: USAGI checking mapping proposed for an antibiotic test

Each mapping can be approved or changed manually; at the end a table is provided in order to integrate the mapping during the database construction.

Mapping Micro-organisms

A similar process is used for mapping micro-organisms code. The species name (mostly in Latin) will be matched with the OMOP vocabulary by USAGI. Including the SNOMED code in the source table, if already known, will ease the verification of the mapping.

During the import the following filtering may be applied:

Filter by concept class = Organism

Filter by vocabulary = SNOMED

Filter by domain = Observation

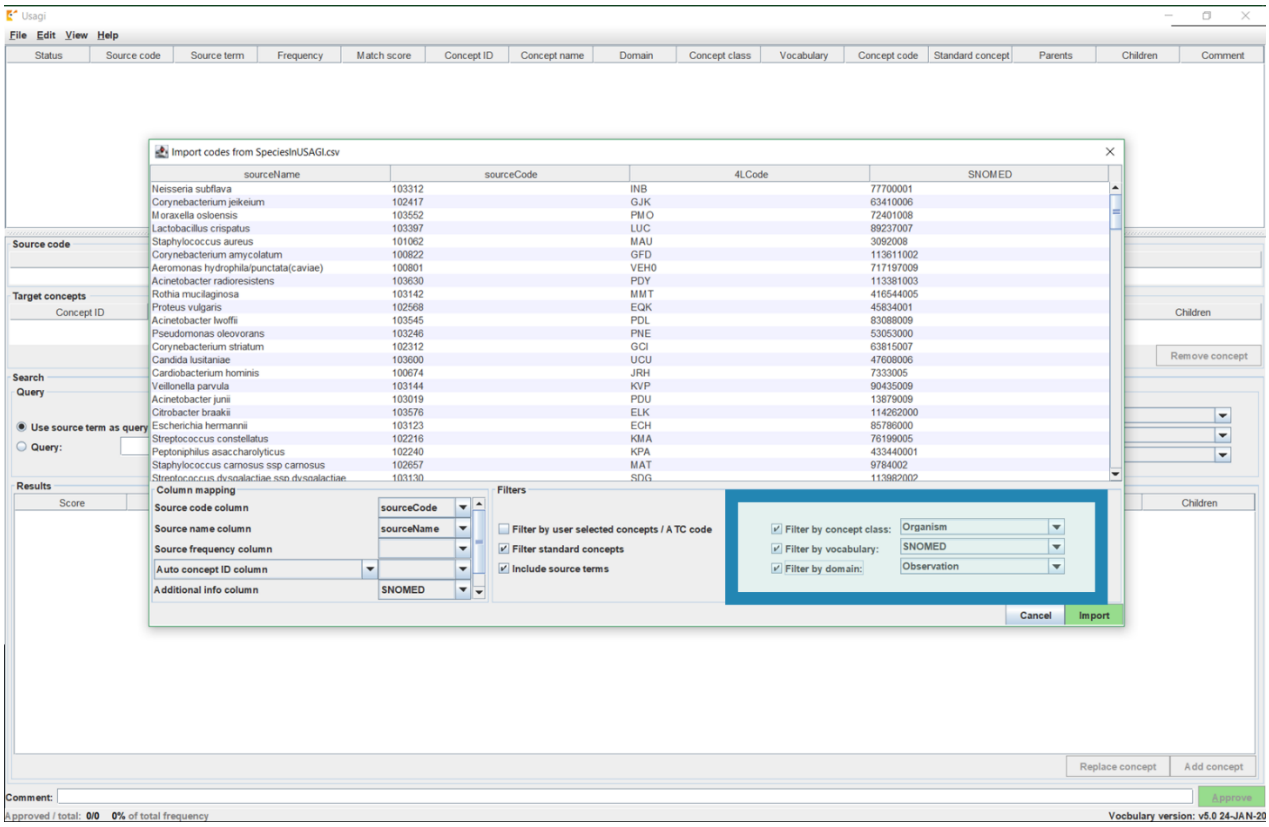


Figure 39: USAGI import of local organism code table

The filtering by SNOMED vocabulary prevents the system to map organisms using the LOINC answer vocabulary, which is considered to be outdated and no longer maintained.

The following figure summarizes all steps performed prior to start the OMOP node database creation.

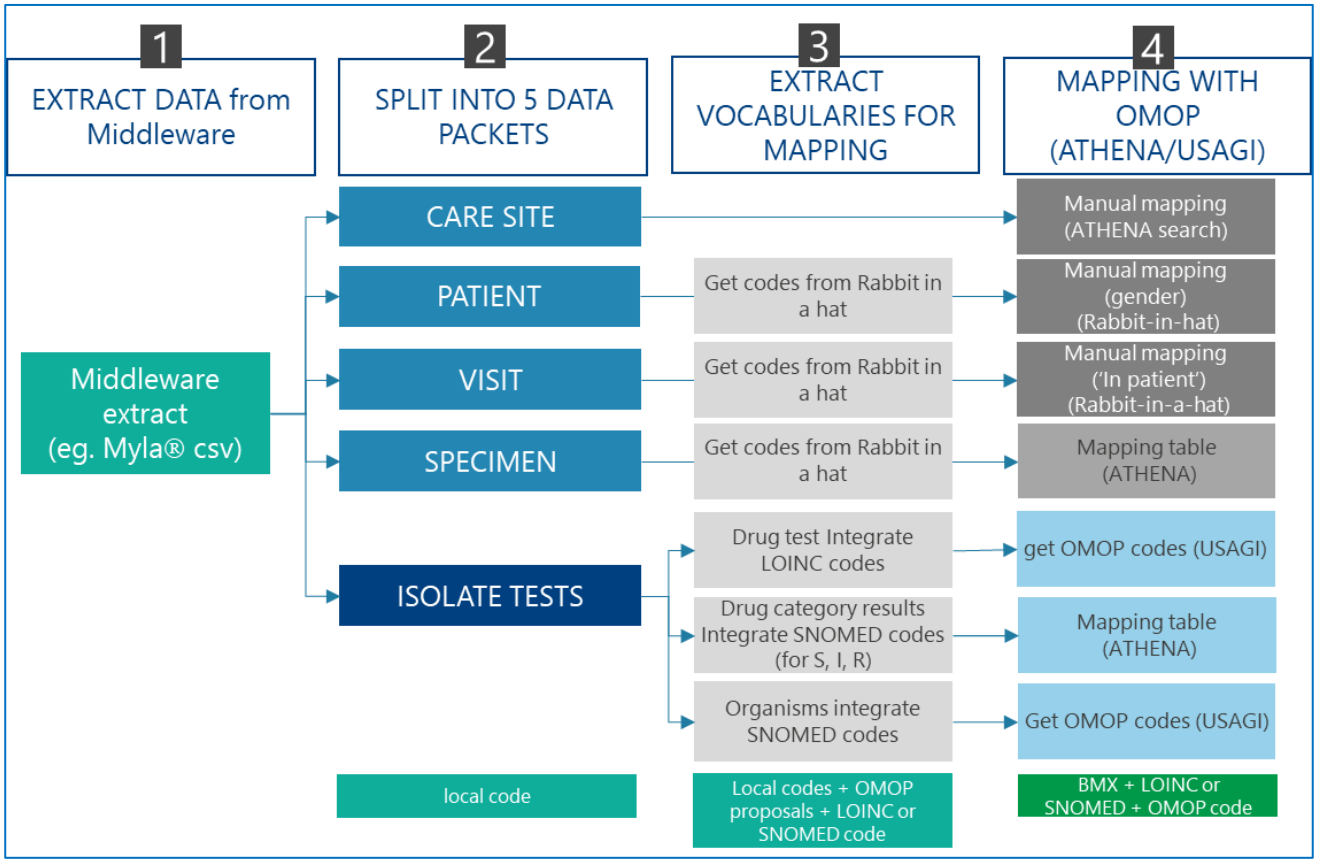


Figure 40: Structural and Vocabulary mapping steps used for an AMR middleware

3.4. Building the OMOP Node database

The OMOP data base is constructed by a series of scripts that prepare the content of each individual relevant table of the data model. The records are generated in a set of csv files, which are structured using the same columns as the destination tables. The identifiers that are required for each of the tables are created on the fly and the local codes are translated into the OMOP *concept_id* codes thanks to the various tables produced during the vocabulary mapping steps.

The **fact relationships** that are aimed at linking the records (specimen to measurement, measurement to observation, measurement to measurement) are calculated on the fly.

The integration of the csv tables into the database is performed by another process which is not described here.

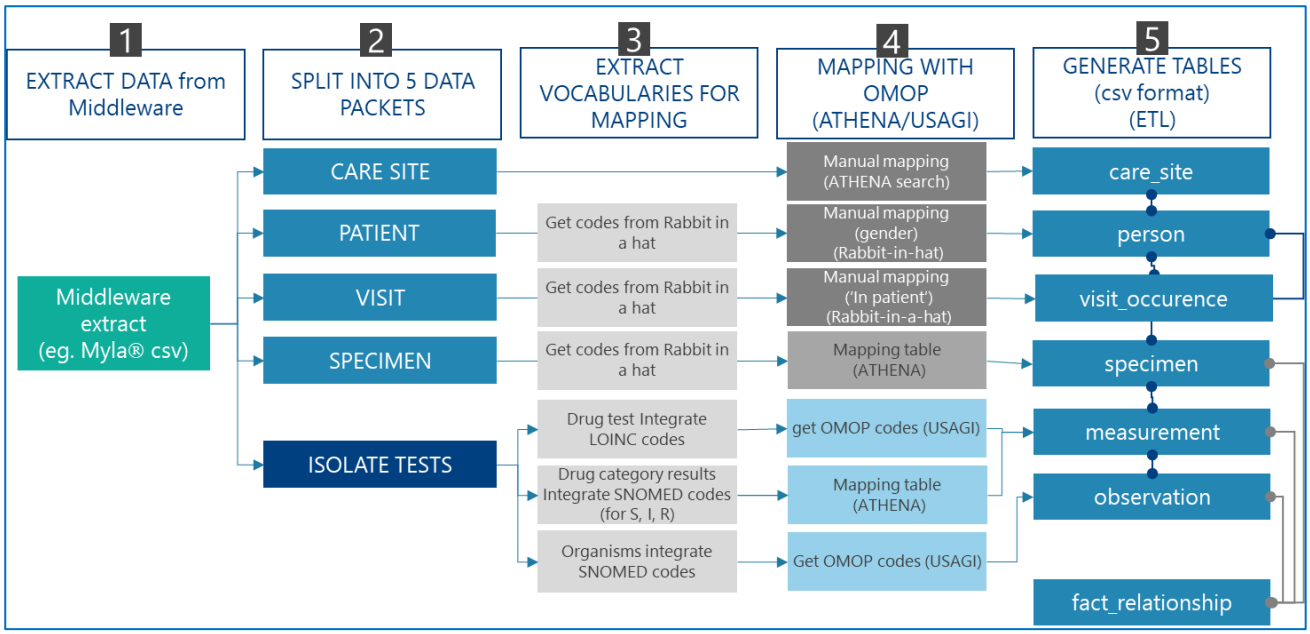


Figure 41: Complete overview of the data preparation process

4. Conclusion on the feasibility of an AMR model into OMOP-CDM

4.1. Limitations of the three models studied

Three modelling options were developed during this study.

However, each one of them advertises limits or do not fully address all the constraints that the OHDSI-OMOP-CDM model imposes.

The following table outlines the main limitations.

	Model 1	Model 2	Model 3
1	Captures only positive tests (isolate centric).	May capture multiple type of tests, but is losing the system which has been used (culture result) instead of testing result for the identification	
2	Two measurements per antibiotic test to strictly comply to LOINC definition of code	Two measurements per antibiotic test to strictly comply to LOINC definition of code	One measurement for each antibiotic test, which may violate a LOINC rule?
3	Organisms identification is captured by an OMOP concept-ID related to a SNOMED code in a measurement which should belong to the observation domain	Organisms captured in an observation which complies to the OMOP constraints	Organisms identification is captured by an OMOP concept-ID related to a SNOMED code in a measurement which should belong to the observation domain

As previously stated, OMOP data model is patient-centric and has not been designed to address the specificity of microbiology data. Therefore, all of the models proposed contravene some of the conventions laid down either by OHDSI, LOINC or SNOMED-CT.

Model 1

This model is Isolate oriented, the “root” measure captures the isolate identification. The result (stored as value-as-concept) corresponds to the OMOP ID of the SNOMED code of the organism name. However, this code belongs to the Observation Domain and OHDSI stipulate that test results in the Measurement table should belong the Measurement Value Domain.

In this Model, MIC and category results of the same AST test are stored separately, in order to respect the units expected by the LOINC code. Using two measurements to store one single test result could lead to inappropriate interpretation of the data. Without in depth knowledge of the data model, it is possible to falsely interpret this two-measurement result as two separate test results.

Model 2

This model is culture oriented. Culture results (positive or negative) are stored in the measurement table and represent the “root” of the model. However, most instruments don’t give culture results. So these results have to be inferred; if a micro-organism is identified, it can be assumed that the culture was positive for micro-organism. This breaks the OHDSI rule to follow the results given by the machine as closely to possible. Moreover it leads to a discrepancy between what is stored as value-source-value (that reflect information of the original database such as the name of the organism identified) and value-as-concept (that, in this case, store OMOP code corresponding to “Present” or “Detected”).

As for Model 1, storing separately MIC and category results could confound statistical analyses of the database.

Model 3

As with Model 1, this model is using SNOMED-derived OMOP code as Value_as_concept_id, violating the Domain restriction imposed for Measurement table.

4.2. Proposal for a new model

None of the proposed models perfectly meet ODHSI's requirements; each of them shows advantages and disadvantages. A new model can be proposed to minimise the number of ODHSI rules broken.

Model 4 Isolate/Culture Mixed model

Model 4 derives from a combination of Model 2 and Model 3.

Instead of storing the result of a culture, the root measurement captures the result of the test for identification or detection (in a test panel), therefore keeping track of the system that has been used for performing the test and producing the identification result. This modelling orientation preserves the traceability to the diagnostic system, which may be of great importance when recording Real World Data (RWD).

This approach also permits the capture of negative results for panel tests, if the interest is demonstrated. Since the positive result of the panel test is stored as the root, and the name of the micro-organism (when detected) stored as an observation, a search by observation will produce a complete report, including both regular identification test results along with micro-organisms detected by panel tests.

Both antibiotic test results (such as MICs) and interpretation (clinical categories), are found in the same measurement, which also permits the traceability between the two pieces of data.

a. Specimen table is used as the “root”
b. A measurement is used to capture the results of the tests as it is present in the data whether it is a micro-organism identification or a binary detection in case of Panel test.
   For AST result
   - value_as_concept_id= OMOP code for the Micro-organism identified (using Meas Value Domain if available)

   For Panel Test
   - value_as_concept_id= OMOP code for “Detected” or “Undetected”

c. An observation capturing the name of the identified organism
   - observation_concept_id= OMOP code for the organism corresponding to a SNOMED code (from the Observation domain)

d. As many measurements as antibiotic tests are linked to the observation. Each of those contains both the code for the antibiotic test and the code for the category result as well as the MIC result. Despite the fact that LOINC code usually prevent from giving both a numerical and a categorical answer, there is a set of LOINC codes (designed for microbiology) that possess an “OrdQn” argument allowing both a numerical and categorical answer.
e. Other measurements present in the data (such as culture result) can be captured in parallel in a separate measurement.

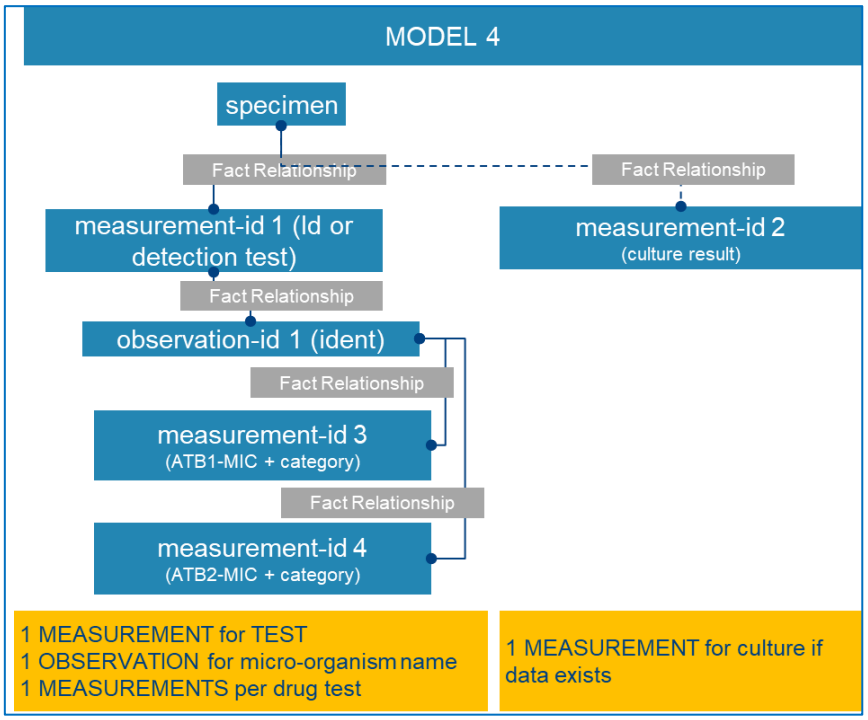


Figure 42: Evolution of Model 2 into Model 4

Compared to Model 2, this model avoids the pitfall of capturing two AST different measurements reflecting the same test result. Moreover, in this Model 4, culture results are not mandatory, preventing the ability to infer them. If present in the data, culture results can still be captured as a measurement. Compared to Model 2, where culture results and identification results were explicitly linked by a fact relationship, in Model 4, the association between a culture result and the corresponding identification results have to be found using specimen ID and date-time. This may only be an issue if the use case is to analyse positive organism identifications in cultures that were considered as negative.

4.3. Conclusions

The OHDSI-OMOP-CDM (V5.3) does not natively support microbiology results, however by leveraging specific features of this version (5.3) the hierarchy of data associated to microbiology results can be represented. This may end-up by making database queries very complex (current OHDSI tools for queries do not yet support this CDM version!).

The constraints of the OHDSI-OMOP vocabularies where each database field needs to be populated by codes associated to a particular domain (such as measurements fields to be populated by codes extracted from vocabularies of the measurement domain), while enforcing interoperability, forces the use of additional tables to ensure compliance to these constraints.

As an outcome to this study, the team will integrate the OHDSI-OMOP-CDM community in order that specific requirements, generated by the support of Microbiology data into the OMOP-CDM model, will be integrated in future evolutions of the data model.

By taking these requirements into account, the OHDSI-OMOP data sets would be in capacity to analyse real world data that would encompass all data captured from preliminary clinical signs up to final patient diagnosis, including all supporting lab testing data.

5. Bibliography

European Medicines Agency *(2018)* A common Data Model for Europe? – Why? Which? How?, Workshop report from a meeting held at the European Medicines Agency, London, United Kingdom.

EDM Forum (2016) Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research - . Gini, M. Schuemie, J. Brown, P. Ryan

PLOS one (2019) Data model harmonization for the All Of Us Research Program!: Transforming i2b2 data into the OMOP common data model. J.G. Klann, M.A. Joss, K. Embree, S.N. Murphy

OXFORD University press (2015) Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. E.A Voss, R. Makadia, A. Matcho, Q. Ma, C. Knoll, M. Schuemie, FJ DeFalco, A. Londhe, V. Zhu, PB. Ryan

The book of OHDSI, *Observational Health Data Sciences and Informatics,* https://ohdsi.github.io/TheBookOfOhdsi/

www.imi.europa.eu www.value-dx.eu