PRECISE4Q

PREDICTIVE MODELLING IN STROKE

DELIVERABLE

Project Acronym: **Precise4Q**

Grant Agreement number: **777107**

Project Title: **Personalised Medicine by Predictive Modelling in Stroke for better Quality of Life**

D3.10 – Paper on Best practices for data sharing in in-silico modelling

Revision: 0.1

Authors and Contributors	Catalina Martínez-Costa (UM); Francisco Abad-Navarro (UM); Stefan Schulz (MUG); Katryna Cisek (TU Dublin); Julia Amann (ETH); Günter Neumann (DFKI); Saadullah Amin (DFKI)			
Responsible Author	Catalina Martínez Costa		**Email**	cmartinezcosta@um.es
	Beneficiary	UM	**Phone**	+34868888432

Project co-funded by the European Commission within **H2020-SC1-2016-2017/SC1-PM-17-2017**		
Dissemination Level		
PU	Public, fully open	
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

Revision History, Status, Abstract, Keywords, Statement of Originality

Revision History

Revision	Date	Author	Organisation	Description
xxx	02/09	CMC	UM	Initial draft
	14/09	KC	TU	data access experience in PRECISE4Q
	16/09	CMC	UM	draft v1
	16/09	JA	ETH	ethical and legal considerations
	19/09	steschu	MUG	draft v2
	22/09	CMC	UM	draft v3
	26/09	CMC	UM	edit
	26/9	steschu	MUG	revision

Date of delivery	Contractual:	30.09.2022	Actual:	30.09.2022
Status	final ☒ /draft □			

Abstract (for dissemination)	This document describes some of the aspects we consider important when harmonizing and integrating data from heterogeneous sources, based on the experiences during the project and state of the art work. We briefly comment on the ethical and legal challenges regarding data access and sharing and we focus on the technical ones related to data harmonization and integration. Finally, we summarize our experiences as a list of recommendations that can serve as a basis for future projects and inform policy makers, industry and key stakeholders for future developments and research directions.
Keywords	Semantic data harmonization, ontology, terminology, SNOMED CT, Semantic Web, FAIR principles, Data privacy

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Content

List of Figures

Executive Summary

Based on the experiences during the project and state of the art work, we describe some of the aspects we consider important when harmonizing and integrating data from heterogeneous sources. We briefly comment on the ethical and legal challenges regarding data access and sharing and we focus on the technical ones related to data harmonization and integration.

Precise4Q has faced the challenge to integrate clinical data that were highly diverse regarding the following source characteristics: (i) different clinical disciplines, (ii) different institutions across Europe, (iii) different degrees of structure (from free text to structured data), (iV) different languages and (V) different data acquisition contexts and purposes.

Precise4Q aimed at improving stroke management through data-driven predictive models, implemented to offer personalized solutions to patients in all stroke phases from prevention to reintegration into the society, passing through acute treatment and rehabilitation.

In addition to all technological challenges derived from data heterogeneity, data sharing agreements did not allow all data to be centralised. Instead, data sharing had to face the challenge to include data that could not leave their source repositories.

Precise4Q opted for a semantic-driven approach and the use of federated learning and remote data access.

Finally, we summarize our experiences as a list of recommendations that can serve as a basis for future projects and inform policy makers, industry and key stakeholders for future developments and research directions.

The described work will be further revised and submitted to a medical informatics journal.

1 Introduction

The benefits of facilitating data sharing in health have been demonstrated during the COVID-19 crisis. The proposal of the European Commission for a 'European health data space' (EHDS) [1] shows the huge interest in improving access to health data across Europe for secondary use.

While data analytics methods are increasingly offering new ways to help optimise operations and drive decision-making, regulatory and public pressure on privacy issues make the creation of the EHDS a complex process [2].

Besides legal aspects, technological issues underlying the sharing and integration of health data are also being discussed within the medical informatics community. Despite many initiatives and projects for decades, data interoperability is still an unsolved task. Multiple health data interoperability standards such as openEHR [3], HL7 FHIR [4], ISO 13606 [5], OHDSI OMOP [6], etc. still co-exist, supporting sharing of data from electronic health records (EHRs) on the one hand, and research data on the other hand, but many implementations for them are not necessarily compatible. In fact, many hospital information systems and dedicated research databases do not yet comply with any of the mentioned standards although efforts are being made in that direction.

In the last years, there has also been a growing interest and effort from the different SDOs in aligning their models. Past European projects like NoE SemanticHealthNet [7] worked on the idea that there is a need to propose methods for ontology-based interoperability. In the U.S., the CTSA ontology group has issued a White Paper on interoperability desiderata [8].

OHDSI, a multi-stakeholder, interdisciplinary collaborative health data sharing initiative [9], has established an international network of researchers and observational health databases with a central coordinating center housed at Columbia University. The recent experience of the OHDSI initiative has demonstrated how fast a new data model is being successfully adopted by a large community of users thanks to being open source and providing an increasing tools ecosystem.

Reflecting on the past and current data interoperability landscape we can think of a future scenario in which new data models, interoperability approaches and tools will emerge. This challenges the community to build a common pan-European infrastructure that facilitates semantic data interoperability independently of specific standards, specifications or implementations.

The authors share the opinion of many renowned experts in the field that semantic resources and technologies, particularly health terminologies, ontologies and information models play a central role here. The practice of Applied Ontology, i.e. the provision of formal descriptions of entities of the domain, will help create bridges across all existing representations based on data semantics and not only on informal structures or envelopes. This is also mentioned in a recent work in the context of the building of the Swiss Personalized Health Network (SPHN) [10], in which the authors state that current trends in data interoperability have moved from a data model technocentric approach to sustainable semantics, formal descriptive languages, and processes [11].

Precise4Q, an European project on stroke management has faced the challenge to integrate clinical data that were highly diverse regarding the following source characteristics: (i) different clinical disciplines, (ii) different institutions across Europe, (iii) different degrees of structure (from free text to structured data), (iv) different languages, and (V) different data acquisition contexts and purposes.

Precise4Q aimed at improving stroke management through data-driven predictive models, implemented to offer personalized solutions to patients in all stroke phases from prevention to reintegration into the society, passing through acute treatment and rehabilitation.

In addition to all technological challenges derived from data heterogeneity, data sharing agreements did not allow all data to be centralised. Instead, data sharing had to face the challenge to include data that could not leave their source repositories. Furthermore, clinical data is very sensitive and therefore highly controlled under GDPR regulations to ensure patient data privacy.

Precise4Q opted for a semantic-driven approach and the use of federated learning and remote data access.

In the following, we will address the technological challenges experienced during the project and briefly comment on ethical and legal issues we had to deal with.

Finally, we provide a list of recommendations that can serve as a basis for future projects and inform policy makers, industry and key stakeholders for future developments and research directions.

2 Main aspects when accessing, harmonizing, and integrating healthcare data for in-silico modelling

Below we describe some of the aspects that we consider relevant when harmonizing and integrating data from heterogeneous sources, from the ethical and legal challenges related to data access and sharing to the technical challenges related to data harmonization.

2.1 Data access

Despite the promise to deliver better healthcare, data sharing is a big bottleneck in healthcare and biomedical research. Existing datasets are controlled by a few researchers at specific institutions or companies, and access for everyone else is laborious, costly, time-consuming, or just impossible despite the fact that the creation of nearly all health data is publicly funded [12]. The current understanding is that clinical data are exclusively collected to support individual care and decision-making, so-called primary use of clinical data. This explains why most clinical information systems are built to support this use case, and electronic health records are rather substitutes for the traditional paper record than data hubs that also support secondary use cases. The restriction to primary use scenarios is explained by personal data protection as a fundamental legal interest. Especially sensitive health data must therefore be handled in an ethical and responsible manner that protects the rights and interests of data subjects [13].

However, the collection, processing, storage, and sharing of personal data is key to advancing scientific research and healthcare delivery. Some authors even argue that data sharing - beyond healthcare delivery - is an ethical and scientific imperative [14] and there are continuing efforts undertaken to make data more accessible to health research, in the European Union, for instance, through the creation of a European Health Data Space [15].

According to the Swiss Personalised Healthcare Network's (SPHN) Ethical Framework for Responsible Data Processing in Personalised Health Research [16], there are four core ethical principles that should guide the processing of personal data and human biological material:

1) **Respect for Persons**, i.e., the rights and dignity of individuals, families, and communities contributing personal data and/or human biological material in the context of research and clinical care, as well as any other type of data that can be useful for biomedical research must be respected, protected, and promoted.
2) **Privacy**, i.e., the privacy of research participants and the confidentiality of their personal information must be safeguarded.
3) **Data FAIRness**, i.e., data that can be used for research purposes and research results should be made available for further research use to advance the common good of scientific knowledge.
4) **Accountability**, i.e., Accountability mechanisms should ensure fair, lawful, and transparent processing of personal data and handling of human biological material.

In addition to the ethical aspects, there are also a number of legal and regulatory considerations when it comes to data access, use, and governance. Particularly multisite research, as it is the case for PRECISE4Q, requires personal data to be shared between institutions and across jurisdictions. This means that within the same project, different datasets are subject to different national data protection, privacy, and research ethics laws. These regulations aim to protect patient privacy, autonomy, and safety, yet at the same time they may impede multisite research, particularly when several jurisdictions are involved. For researchers it is not easy to navigate these different systems where definitions, standards, rules, and requirements may vary significantly.

A comparative study which investigated the concept of data accessibility in data protection and research ethics laws across seven jurisdictions (Switzerland, Italy, Spain, the United Kingdom, the United States, Canada, and Australia) identified the requirements for consent, the standards for anonymization or pseudonymization, and adequacy of protection between jurisdictions as key barriers for data sharing [17]. The study also found differences between the European Union (EU) and other jurisdictions to pose a particular challenge for data accessibility. Possible solutions proposed by the authors include both regulatory and technical solutions.

Over the course of the PRECISE4Q project, there were numerous challenges concerning data accessibility, including the need for federated learning and remote data access, due to ethical and legal issues, and the need to comply with data privacy and compliance with General Data Protection Regulation (GDPR) regulations.

It was therefore a major effort to reconcile these principles with the goals of PRECISE4Q. Data sharing agreements took months and the conditions depended on each dataset and institution, with some institutions being more flexible than others. In particular, some institutions did not allow their patient data being stored at any place away from their servers.

Federated learning, or the training of models and machine learning algorithms across various decentralized data sources, was critical for the completion of PRECISE4Q, especially for those parts of the project where prevention models were developed. The data sources containing the information necessary to train the models were not stored in a single repository or server, but rather located on different servers located in different countries. Therefore, models had to travel to these sites, ensuring interaction with the data exclusively at the respective sites. Only model parameters (i.e., training results) left the institutions, so that data privacy issues, data security concerns and regulatory restrictions were not violated [18].

It is known that data anonymization alone does not guarantee data privacy in many cases. For instance, imaging data cannot be fully anonymized, as it is possible to reverse engineer an algorithm to reconstruct a face from computed tomography (CT) or magnetic resonance imaging (MRI) data [19]. Therefore, the nature and scope of the data processing (e.g., how data was used, stored and deleted, whether ethnic or genetic data were collected, who were the individuals and from what geographic area), as well as the context of the data processing (did the data subjects provide a written consent for the research and did they know how to withdraw their data from the research) and finally potential risks and measures to mitigate them had to be outlined in a Data Protection Impact Assessment (DPIA) [20].

In [21], the authors propose a data-flow protocol that describes all the steps to create an anonymous (non-personal) dataset, incorporating best practice and providing adaptable steps for handling data in accordance with UK and EU ethical and legal framework.

In [12], the authors state that although protecting patient privacy is the most relevant barrier to data sharing, many technical solutions to this problem exist, from sophisticated de-identification methods to highly secure cloud environments.

Focusing on developing tools for the automatic de-identification of texts, in Precise4Q, a method for the de-identification of free texts in Catalan and Spanish has been implemented [22]. A severe shortcoming in clinical text de-identification is the lack of annotated corpora for training. The implemented method with a small number of annotated samples was able to successfully de-identify most of the evaluated texts.

2.2 Data harmonization

2.2.1 Data description

After ethical and legal data sharing agreements are in place, the next step is preparing data to be shared.

Most existing healthcare data consists of more or less structured narratives. But also structured data usually includes a large amount of data elements like variables and values whose meaning is in the best case elucidated by some data dictionary, whereas in the worst case the user has no choice but to guess the meaning based on the variable names and the context of use (e.g. within a data acquisition form). This is explained by the fact that such data repositories were devised for data entry and data display by and for humans, not to machines. Consequently, tasks involving the combination of data from multiple data sets that are described using data dictionaries are not easily automated.

Even where data elucidations are provided by some kind of dictionary, crucial details of data descriptions are often missing. Examples are the use of foreign languages, specific terminologies, local dialects and jargons, missing data, mismatch of data elements, missing data types or descriptions, use of terminologies without specifying the version, lack of availability of the full list of permissible values, etc.

This makes reuse of the shared information a difficult and time consuming task, requiring repeated interactions with data owners, which is often not easily achievable.

Previous studies have provided recommendations to guide optimal data sharing and reuse. One is [23], whose CONSIDER statement among others make the following recommendations:

1. Provide data dictionary documentation separate from de-identified individual patient data. Since it does not contain patient data, it does not require ethical approval. It should be shared as soon as possible and be provided as a machine-readable file.
2. For each data element provide its data type (e.g. numeric, data, string, categorical). For categorical data elements provide a list of permissible values and distinguish them from numeric or string values.
3. Provide a complete data dictionary (all elements in the data are listed in the dictionary). Group data elements if necessary. Use a description field in addition to the title to fully describe the data element.

Dictionaries are useful for many data management tasks, such as aiding users in data conversion processes, test data generation, data validation and storing data usage criteria [24]. Information from dictionaries is usually known as metadata and stored in metadata registries or repositories following metadata standards like ISO 11179 [25].

Examples are the open-source Samply Metadata Repository [25], created to support the formulation of inquiries to networked biobanks based on their respective data elements, or Semantic Metadata Registry (Semantic MDR), an implementation of the ISO/IEC 11179 standard using Semantic Web technologies [27].

As mentioned in [28], metadata can be a powerful resource for identifying, describing, and processing information, but its meaningful creation is costly. In addition, the sheer number of relevant metadata standards has led to oversaturation and rejection [28].

The FAIR principles proposed by Wilkison et al [29] have attracted much attention as they define how to make data (and in general digital objects [30]) more Findable, Accessible, Interoperable and Reusable. These principles make a clear distinction between data and metadata.

The FAIR Data Point (FDP) is an approach to exposing semantically-rich metadata for a wide range of data (also known as digital objects) in a FAIR manner [31]. Its main goal is to establish a common method for metadata provisioning and accessing that is compliant with the FAIR principles. It can be used to expose metadata of datasets, but also from other digital resources (e.g. ontologies, ML workflows, etc.). It makes use of widely used vocabularies like DCAT, FOAF and Dublin Core Terms. Moreover, metadata records have references to ontological annotations using external ontologies (e.g. within the medical domain, using the SNOMED CT concept for certain disease), which ideally convey computable (logic-based) axioms and definitions. A reference implementation is provided in [32]. In the FDP, metadata is represented in RDF [33], being the storage usually an RDF store. Thus, representing metadata following the FDP approach will benefit from Semantic Web technologies, facilitating the automatic and meaningful combination of data from multiple data sets.

2.2.2 Semantic data harmonization and integration

Once data has been shared, independently of its data format or syntax it requires to be semantically harmonized in order to be used for comparison. Data might be provided at different levels of detail, using different protocols, (e.g. fever measured in different body sites) different units of measurements (e.g. body temperature measured in degree Fahrenheit vs. Celsius), etc.

Semantic harmonization is the process of combining multiple sources and representations of data into a form where items of data share meaning [34]. Harmonized data allows single given questions to be asked and answered across the data as a whole, without the need to modify or adapt queries for a given data source, invaluable as a tool for researchers [35].

Semantic harmonization is a time-consuming task, which requires consultation and agreement across a wide range of stakeholders, especially when data serves multiple purposes. It is therefore essential to properly describe all data sources with the exact meaning of every data element included in the dataset. Metadata registries such as based in FDP will facilitate data harmonization. Metadata can support data discovery and inter-comparison across variables or data elements from several sources. Guided by the metadata, ETL (extraction, transform, load) processes can then be implemented to semantically harmonize clinical data. For clinical data representation multiple options can be considered. Most of them are based on the use of a common data model like ISO/CEN 13606 [5], openEHR [3], OMOP CDM [6], being the two first ones devised for exchanging data for patient care and the latter one for research. A specification based on Semantic Web principles [36] is HL7 FHIR [4]. It is based on a content model in the form of modular components named ‘resources’ and a specification for their exchange using REST services.

Existing standards differ in their purpose, data model and modelling approach. However, it does not mean that some of them can not be conceptually and technically compatible [37]. In addition, their adoption has shown that much effort is still needed to achieve interoperability at the semantic level [38].

Nevertheless, there is recent evidence that aligning data models from existing standards is possible. Examples are the Vulcan project [39] for transforming FHIR data into OMOP, and the Common Data Model Harmonization (CDMH) project [40] for mapping and translating observational data represented using research oriented data models like OMOP and i2b2 [41] into FHIR and CDISC SDTM [42] specifications [43].

In line with the need to facilitate interoperability among existing data modelling standards, already in 2013, the Yosemite project recognised the need to provide translations between data models and vocabularies and proposed RDF as ‘universal healthcare exchange language’. The Yosemite manifesto [44] states that existing standard healthcare vocabularies, data models, and exchange languages should be leveraged by defining standard mappings to RDF, and any new standards should have RDF representations. In fact, HL7 FHIR provides RDF as one standard data format approaching the Semantic Web vision [36].

Thus, given that existing standard data models will coexist with new, non-standard ones, developed by many organizations, there is the need to facilitate data interoperability also beyond of the particular standards, specifications or data representations.

Semantic data harmonization and integration crucially depend on technologies and resources. In particular, semantic web technologies such as ontologies and semantic languages like OWL and RDF have contributed much to achieve semantic interoperability in health information systems [45]. Formal Ontologies, which we understand as "precise mathematical formulations" of the entities of a domain and the way how they are related, are well-established to support knowledge-intensive tasks related to EHR systems [46]. Formal ontologies given their integration capabilities and in concert with domain terminologies (often known as controlled Vocabularies (CVs)), play a central role in bridging across all existing representations based on the semantics of data [48], regardless of the structure or 'envelope' used [47].

Up until now, numerous project-specific ontologies have been built without any interoperability or standardization interest. They are maintained for the duration of a certain project and are then abandoned. They do not refer to foundational ontologies, nor do they re-use content from other domain ontologies. The fact that the foundational ontology BFO 2020 has become an ISO standard shows the interest in building standardized ontologies in the field of science and engineering [49].

Standardization has also been an issue in biomedical terminologies, particular in the case of SNOMED CT, which set off as an international terminology for EHRs, but which then increasingly adopted principles of formal ontology and logic, so that it can now be seen as clinical ontology of high coverage and granularity [50].

This was also one of the main outcomes of the FP7 SemanticHealthNet NoE [7], that highlighted the importance of formally representing the meaning of clinical information and differentiated between their structural and semantic representations [51] with the ultimate goal of achieving semantic interoperability independently of the standard or data representation used.

This has also been one of the main results of Precise4Q, which has specified and partially implemented a semi-automatic method for data harmonization, guided by semantics independent of the target representation model. It is based on the following principles:

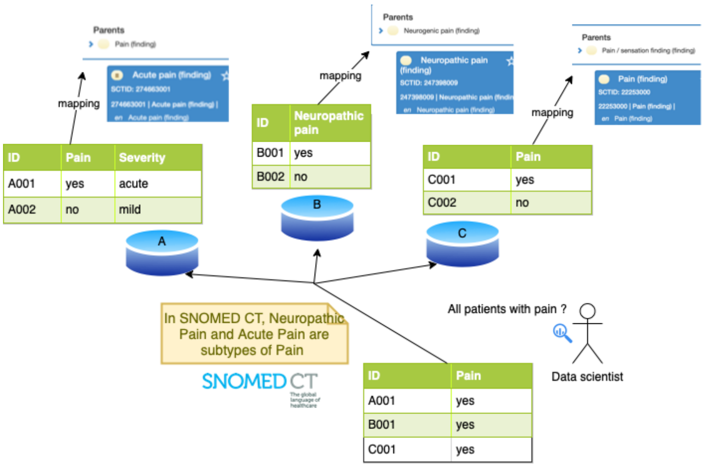
1. A standardized and simple data specification allows harmonization for research purposes
2. This has to be done semi-automatically, provided that formal ontology guides the creation of semantic data models that are independent of a specific clinical information representation standard.

The core of the method is an ontology-based data model, which uses a top-level ontology in order to standardize data modelling and to ensure interoperability among different ontologies. SNOMED CT plays the role of a reference ontology to represent the medical domain knowledge.

Having all data elements sufficiently described using a specific structured representation, data is automatically transformed into the desired target data model. PRECISE4Q used OMOP CDM as target representation, but other models such as HL7 FHIR would be equally eligible. Through predefined transformation rules, data are automatically generated in RDF according to the ontological model of OMOP [52]. Through SPARQL queries, OMOP RDF data can be also translated into relational database format.

The ontology-based data model allows standardizing the representation of data meaning, independently of the particular structure or syntax from the original data. Thus, the model acts as a bridge between data representation standards. Based on the ontology model, a set of predefined transformation rules are in charge of transforming data into RDF according to the specific target data model.

Having a precise definition of the data will support the integration of similar data from heterogeneous sources. Depending on the use case, a certain representation of the data will require doing some data transformations (e.g. use of certain units, harmonization of value ranges, etc.). Examples of well-known efforts for building agreed representations of data are the Patient Summary and e-Prescription within the epSOS project [53]. However, not always it is possible or even necessary to agree on certain data representation, but different data granularities are supported by ontologies (e.g. allowing different data granularities when representing pain as depicted by Figure 1).


Parents
Pain (finding)
Acute pain (finding)
SCTID: 274663001
274663001 | Acute pain (finding) |
en Acute pain (finding)
mapping
Parents
Neurogenic pain (finding)
Neuropathic pain (finding)
SCTID: 247398009
247398009 | Neuropathic pain (finding)
en Neuropathic pain (finding)
mapping
Parents
Pain / sensation finding (finding)
Pain (finding)
SCTID: 22253000
22253000 | Pain (finding) |
en Pain (finding)
mapping
A
B
C
All patients with pain ?
Data scientist
SNOMED CT
The global language of healthcare

ID	Pain	Severity
A001	yes	acute
A002	no	mild

ID	Neuropathic pain
B001	yes
B002	no

ID	Pain
C001	yes
C002	no

In SNOMED CT, Neuropathic Pain and Acute Pain are subtypes of Pain

ID	Pain
A001	yes
B001	yes
C001	yes

Figure 1 – Semantic data harmonization example. Three heterogeneous representations of pain but all of them subtypes of the SNOMED CT concept for pain

In [11], the authors recognize the coexistence of multiple data models. They highlight that there is no one-size-fits-all data model and that current trends in data interoperability have moved from a technocentric data model approach to sustainable semantics, formal descriptive languages, and processes.

As part of the building of the Swiss Personalized Health Network (SPHN) [10] they are implementing a semantic-driven data model-independent framework for semantic interoperability. This framework departs from concepts (variables or data elements) encoded using international standards such as ICD-10, SNOMED CT, LOINC, or ad hoc knowledge representation using RDF for representing data. Then, RDF data is converted to any target data model needed.

From the above described approaches and experiences we can take some important messages:

1. The need for a semantic definition of the original data (e.g. using existing terminologies and ontologies)
2. Flexible descriptions of data, agnostic regarding any data model that facilitates data conversion into other representations
3. Need to implement specific converters for any target data specification and format (e.g. OMOP, FHIR, but also XML, JSON, etc.)

The first and the second items are the most difficult to address. The need for a semantic definition of the original data highlights the need for rich metadata repositories and good data terminology/ontology data mapping tools to support the semantic standardization of the metadata.

As previously mentioned, metadata should include not just the name of the data elements but also their descriptions. Both should support the mapping of the data elements to existing terminologies and ontologies and thus facilitate semantic data harmonization.

Manual mapping is time-consuming and error-prone. Text mining tools and terminology servers are used to support the automatic mapping of data to ontologies or vocabularies. On the one hand, text mining systems performance depends on the available vocabularies and corpora, as well as language models derived from the latter. Collections of technical terms that primarily represent the language used in the clinic - so-called user interface terminologies - , together with open annotated corpora are essential to improve performance on text mining tools [8].
State of the art tools for mapping free text to SNOMED CT are mostly based on rules but a combination of machine learning approaches with rule-based ones could be a way to improve performance [54].
On the other hand, terminology servers like Snowstorm for SNOMED CT [55], can be used to map structured data to an ontology or vocabulary. In [56], a mapping approach where structured text and a terminology server are used together is described.

In addition, although SNOMED CT is considered a reference terminology for medicine, and an important resource to achieve semantic interoperability in healthcare, it is not sufficient for encoding all medical domain knowledge, because its coverage for specific medical domains is limited [57]. In fact, when data is encoded, often different ontologies and CVs or terminologies are used. Besides, across different data sources, multiple vocabularies and ontologies might be used for encoding the same concept (e.g. using SNOMED CT and ICD-10 for encoding certain diseases).

Thus, it is necessary to make available resources like ATHENA, part of the OMO approach [58] that acts as a common repository of biomedical vocabularies and provides equivalences among concepts from several of them. Among others, ATHENA uses UMLS [59], which integrates more than 214 vocabularies and ontologies. Other works like [60] provide mappings from OMOP vocabularies to OBO Foundry ontologies [61]. Finding correspondences between concepts from vocabularies and ontologies is not trivial and usually requires manual validation. Contributing to their coordinated building by following certain criteria or principles like ontologies from the OBO Foundry [61], and ideally within an overarching ontological framework, will facilitate their integration. In this line, the Clinical and Translational Science Ontology Group (CTSA) proposes an increased investment in the research and development of ontologies to address the limitations in their use with EHRs [8].

The second message refers to the description of the data in a flexible way, which is agnostic regarding any data mode and that facilitates data conversion into other representations. Here, it is necessary to link previously identified concepts with informational aspects, contextualizing them (e.g. for the identified concept 'neuropathic pain', we need to know when (e.g. 12/01/2007), how (e.g. 'conversation with the patient') and by whom (e.g. 'the patient') this was stated). Using high-level categories and relations as provided by a top-level ontology can support us to link the different pieces in a standardized way [48,62]. In addition, using RDF as a representation language provides a flexible data model and serves as the 'lingua franca' for exchanging machine-processable information. RDF does not depend on a specific semantic standard, but allows using different ontologies and vocabularies. Ontologies provide the formal definition that allows both machines and human beings to understand the intent of the information [63]. In addition, RDF can be used together with other formalisms when needed for other types of information and purposes (e.g., Guidelines Interchange Format for guidelines and Java Business Process Model for workflows) [11].

Additionally, specific converters of the data into other formats such as OMOP, HL7 FHIR or any other data model or syntax can be implemented. There are already works for transforming RDF data into relational data models or standard common data models like OMOP [63] or HL7 FHIR. The use of a top-level ontology for representing data, as well as vocabularies and ontologies for encoding medical domain knowledge, will facilitate the implementation of the specific converters.

Finally, knowledge graphs represented in RDF enable the use of graph algorithms and machine learning techniques to find hidden patterns in the data and infer new knowledge [64]. They are more computationally efficient when data is very interconnected and scale to very large sizes [65]. Recent studies advocate their use as a tool for explainable machine learning [66].

In PRECISE4Q, data represented in RDF is stored in a graph database and can be queried by using a REST API or a specific query system for non-experts in semantic technologies. In addition, we have begun to exploit the structure of the data graph to support machine learning, by applying algorithms that facilitate data exploration and analysis like page rank and community detection algorithms.

3 Summary of recommendations

Based on the Precise4Q data sharing experiences and current state of the art research and implementations in data sharing, we summarize what we consider the most relevant recommendations from the issues discussed above:

1. *Data should be described by **rich metadata** (data about the collected data) and the latter should be shared as soon as possible since it does not contain patient data and does not require ethical approval.*

   Metadata is a powerful resource for identifying, describing and processing data. It is crucial for semantic harmonization and data integration. Data providers should provide metadata such as name and description of the data elements, their data types and list of permissible values. Based on this metadata, automatic methods can be implemented for their semantic enrichment such as annotating them with identifiers from CVs and ontologies.

2. *Metadata should follow the **FAIR principles** and be based on **semantic metadata standards**.*

   Using Fair Data Points (FDPs) for representing and sharing metadata has been proposed by the authors of the FAIR recommendations. It is a promising solution to follow **a common approach** to publish semantically-rich and machine-processable metadata according to FAIR. Representing metadata using FDPs will benefit from Semantic Web technologies, facilitating the automatic and meaningful combination and integration of data from multiple and possibly heterogeneous data sets. It is important to highlight that FDPs address interoperability of the metadata but not of the data.

3. **Metadata and data should be described using ontologies or terminologies as far as possible.**

   Data and metadata should be mapped to CVs and ontologies (e.g. SNOMED CT, LOINC, etc.). Automatic methods based on text mining can be implemented to support the mapping process, e.g. via terminology server implementations.

4. **The construction of ontologies and terminologies should be better coordinated.**

   An overarching ontological framework, e.g. a foundational ontology, can facilitate their integration.
   Since data and metadata will be encoded using multiple CVs and ontologies, correspondences between their terms and concepts must be established. Contributing to the coordinated building of CVs and ontologies by following certain criteria or principles like ontologies from the OBO Foundry, and ideally within an overarching ontological framework, will facilitate their integration. Thus, increased investment in ontology research and development should aim at best practices and engineering standards.

5. **Data should be described in a flexible way** to facilitate data conversion into multiple representations and follow FAIR principles.

   RDF allows the representation of data using a flexible data model and facilitates accomplishing FAIR data principles. In addition, RDF can be used together with other formalisms when needed for other types of information and purposes

6. **Bridging methods for transforming data** into other standardized or not representations should be based on ontologies.

   Since there is no universal agreed standard or representation for clinical data, specific converters need to be implemented. They should be guided by the clinical content, encoded using ontologies developed under an overarching ontological framework.

In addition to the above recommendations, none of them will succeed without appropriate open-source tools. The existence of reference implementations like in the case of the FDPs can facilitate the adoption of the corresponding technology.

4 Conclusions

This deliverable summarizes project experiences regarding data sharing and puts them in the context of state of the art work. It proposes a list of recommendations for data sharing that can serve as a basis for future projects and inform policy makers, industry and key stakeholders for future developments and research directions.

The described work will be further revised and submitted to a medical informatics journal.

5 References

1. European Health Data Space. European Commission site. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en (last accessed Sept. 22)
2. Horgan D, Hajduch M, Vrana M, Soderberg J, Hughes N, Omar MI, Lal JA, Kozaric M, Cascini F, Thaler V, Solà-Morales O, Romão M, Destrebecq F, Sky Gross E. European Health Data Space—An Opportunity Now to Grasp the Future of Data-Driven Healthcare. Healthcare. 2022; 10(9):1629. https://doi.org/10.3390/healthcare10091629
3. Kalra, D., Beale, T., & Heard, S. (2005). The openEHR foundation. *Studies in health technology and informatics*, *115*, 153-17
4. HL7 FHIR. https://hl7.org/fhir/ (last accessed Sept. 22)
5. ISO 13606 Standard, Part 1: Reference model. https://www.iso.org/obp/ui/#iso:std:iso:13606:-1:ed-2:v1:en (last accessed Sept. 22)
6. Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., ... & Ryan, P. B. (2015). Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. In *MEDINFO 2015: eHealth-enabled Health* (pp. 574-578). IOS Press.
7. SemanticHealthNet project. https://cordis.europa.eu/project/id/288408/es (last accessed Sept. 22)
8. CTSA Whitepaper on ontology desiderata. ArXiV 4495999
9. Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., Suchard, M. A., Park, R. W., Wong, I. C., Rijnbeek, P. R., van der Lei, J., Pratt, N., Norén, G. N., Li, Y. C., Stang, P. E., Madigan, D., & Ryan, P. B. (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Studies in health technology and informatics, 216, 574–578
10. Swiss Personalized Health Network (SPHN). https://sphn.ch. (last accessed Sept. 22)
11. Gaudet-Blavignac C, Raisaro JL, Touré V,Österle S, Crameri K, Lovis C. A National, Semantic-Driven, Three-Pillar Strategy to Enable Health Data Secondary Usage Interoperability for Research Within the Swiss Personalized Health Network: Methodological Study. JMIR Med Inform 2021;9(6):e27591 doi: 10.2196/27591
12. Mullainathan, S., & Obermeyer, Z. (2022). Solving medicine's data bottleneck: Nightingale Open Science. Nature Medicine, 28(5), 897-899.
13. Vayena, E., Blasimme, A. Biomedical Big Data: New Models of Control Over Access, Use and Governance. Bioethical Inquiry 14, 501–513 (2017). https://doi.org/10.1007/s11673-017-9809-6
14. Bauchner H, Golub RM, Fontanarosa PB. Data Sharing: An Ethical and Scientific Imperative. JAMA.2016;315(12):1238–1240. doi:10.1001/jama.2016.2420
15. Vayena, E. (2021). Value from health data: European opportunity to catalyse progress in digital health. The Lancet, 397(10275), 652-653.
16. Ethical Framework for Responsible Data Processing in Personalized Health Research. https://swissethics.ch/assets/pos_papiere_leitfaden/ethical_framework_20180507_sphn.pdf (Last accessed: Sept'22)
17. James Scheibner, Marcello Ienca, Sotiria Kechagia, Juan Ramon Troncoso-Pastoriza, Jean Louis Raisaro, Jean-Pierre Hubaux, Jacques Fellay, Effy Vayena, Data protection and ethics requirements for multisite research with health data: a comparative examination of legislative governance frameworks and the role of data protection technologies, Journal of Law and the Biosciences, Volume 7, Issue 1, January-June 2020, lsaa010, https://doi.org/10.1093/jlb/lsaa010
18. Rieke, N., Hancox, J., Li, W. et al. The future of digital health with federated learning. npj Digit. Med. 3, 119 (2020). https://doi.org/10.1038/s41746-020-00323-1

19. Schwarz, C. G. et al. Identification of anonymous mri research participants with face-recognition software. N. Engl. J. Med. 381, 1684–1686 (2019)
20. Yuan B, Li J. The Policy Effect of the General Data Protection Regulation (GDPR) on the Digital Public Health Sector in the European Union: An Empirical Investigation. Int J Environ Res Public Health. 2019 Mar 25;16(6):1070. doi: 10.3390/ijerph16061070. PMID: 30934648; PMCID: PMC6466053.
21. Crossfield, S. S., Zucker, K., Baxter, P., Wright, P., Fistein, J., Markham, A. F., ... & Hall, G. (2022). A data flow process for confidential data and its application in a health research project. PloS one, 17(1), e0262609.
22. Saadullah Amin, Noon Pokaratsiri Goldstein, Morgan Wixted, Alejandro Garcia-Rudolph, Catalina Martínez-Costa, and Guenter Neumann. 2022. Few-Shot Cross-lingual Transfer for Coarse-grained De-identification of Code-Mixed Clinical Texts. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 200–211, Dublin, Ireland. Association for Computational Linguistics
23. Mayer, C. S., Williams, N., & Huser, V. (2020). Analysis of data dictionary formats of HIV clinical trials. Plos one, 15(10), e0240047.
24. Sabbir M. Rashid, James P. McCusker, Paulo Pinheiro, Marcello P. Bax, Henrique O. Santos, Jeanette A. Stingone, Amar K. Das, Deborah L. McGuinness; The Semantic Data Dictionary – An Approach for Describing and Annotating Data. Data Intelligence 2020; 2 (4): 443–486. doi: https://doi.org/10.1162/dint_a_00058
25. ISO/IEC 11179 Metadata registries (MDR). http://metadata-standards.org/11179/ (last accessed Sept. 22)
26. Kadioglu, D., Breil, B., Knell, C., Lablans, M., Mate, S., Schlue, D., ... & Prokosch, H. U. (2018, January). Samply. MDR-A Metadata Repository and Its Application in Various Research Networks. In GMDS (pp. 50-54).
27. Sinaci, A. A., & Erturkmen, G. B. L. (2013). A federated semantic metadata registry framework for enabling interoperability across clinical research and care domains. Journal of biomedical informatics, 46(5), 784-794.
28. Ulrich H, Kock-Schoppenhauer A,Deppenwiese N, Gött R, Kern J, Lablans M, Majeed RW, Stöhr MR, Stausberg J,Varghese J, Dugas M, Ingenerf J. Understanding the Nature of Metadata: Systematic Review J Med Internet Res 2022;24(1):e25440 doi: 10.2196/25440
29. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.
30. De Smedt, K., Koureas, D., & Wittenburg, P. (2020). FAIR digital objects for science: From data pieces to actionable knowledge units. Publications, 8(2), 21.
31. Luiz Olavo Bonino da Silva Santos, Kees Burger, Rajaram Kaliyaperumal, Mark D. Wilkinson; FAIR Data Point: A FAIR-oriented approach for metadata publication. Data Intelligence 2022; doi: https://doi.org/10.1162/dint_a_00160
32. Fair Data Point Reference Implementation. https://github.com/FAIRDataTeam/FAIRDataPoint (Last accessed: Sept'22)
33. Resource Description Framework (RDF). https://www.w3.org/RDF/ (last accessed Sept. 22)
34. Holmes, J. H., & Naylor, M. D. (2013). Conducting research using the electronic health record across multi-hospital systems: semantic harmonization implications for administrators. The Journal of nursing administration, 43(6), 355–360. https://doi.org/10.1097/NNA.0b013e3182942c3c
35. Cunningham, J. A., Van Speybroeck, M., Kalra, D., & Verbeeck, R. (2017). Nine Principles of Semantic Harmonization. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2016, 451–459.
36. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 34-43.

37. Pedrera-Jiménez, M., Spanish Expert Group on EHR standards, ., Kalra, D., Beale, T., Muñoz-Carrero, A., & Serrano-Balazote, P. (2022). Can OpenEHR, ISO 13606 and HL7 FHIR work together? An agnostic perspective for the selection and application of EHR standards from Spain (Version 1). TechRxiv. https://doi.org/10.36227/techrxiv.19746484.v1 ([])
38. de Mello, B.H., Rigo, S.J., da Costa, C.A. et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol. 12, 255–272 (2022). https://doi.org/10.1007/s12553-022-00639-w
39. The Vulcan project http://www.hl7.org/vulcan/ (Last accessed: Sept '22)
40. Common Data Model Harmonization project (CMDH). https://www.healthit.gov/sites/default/files/page/2020-07/CDMH-Project-Summary.pdf (Last accessed Sep'22)
41. I2b2 Common Data Model Documentation. https://community.i2b2.org/wiki/display/BUN/i2b2+Common+Data+Model+Documentation (Last accessed: Sept'22)
42. CDISC SDTM standard. https://www.cdisc.org/standards/foundational/sdtm (Last Accessed: Sept'22)
43. Common Data Model Harmonization (CDMH) and open standards for evidence generation. https://aspe.hhs.gov/sites/default/files/private/pdf/259016/CDMH-Final-Report-14August2020.pdf (Last Accessed: Sept'22)
44. Yosemite manifesto. http://yosemitemanifesto.org (last accessed: Sept'22)
45. Burse, R., Bertolotto, M., O'Sullivan, D., & McArdle, G. (2021). Semantic interoperability: the future of healthcare. In Web Semantics (pp. 31-53). Academic Press.
46. Rector, A., Schulz, S., Rodrigues, J. M., Chute, C. G., & Solbrig, H. (2019). On beyond Gruber: "Ontologies" in today's biomedical information systems and the limits of OWL. Journal of biomedical informatics, 100S, 100002. https://doi.org/10.1016/j.yjbinx.2019.100002
47. Schulz, S., Martínez-Costa, C., Karlsson, D., Cornet, R., Brochhausen, M., & Rector, A. L. (2014, September). An ontological analysis of reference in health record statements. In FOIS (pp. 289-302).
48. Martínez-Costa, C., Cornet, R., Karlsson, D., Schulz, S., & Kalra, D. (2015). Semantic enrichment of clinical models towards semantic interoperability. The heart failure summary use case. Journal of the American Medical Informatics Association, 22(3), 565-576.
49. Schulz S, Stegwee R, Chronaki C. Standards in Healthcare Data. In: Kubben P, Dumontier M, Dekker A, editors. Fundam. Clin. Data Sci., Cham (CH): Springer; 2019.
50. Schulz, S., & Martínez-Costa, C. (2015). Harmonizing SNOMED CT with BioTopLite: an exercise in principled ontology alignment. In MEDINFO 2015: eHealth-enabled Health (pp. 832-836). IOS Press.
51. Martínez-Costa, C; Karlsson, D; Schulz, S SEMANTIC INTEROPERABILITY BY ONTOLOGY BASED REPRESENTATION OF CLINICAL INFORMATION In: Ammenwerth, E; Hörbst, A; Hayn, D; Schreier, G; editors(s). Health Informatics meets eHealth - von der Wissenschaft zur Anwendung und zurück. Big Data - eHealth von der Datenanalyse bis zum Wissensmanagement. 293: 1080 Wien, Piaristengasse 19: Druckerei Riegelnik; p. 65-71. 2013(ISBN: 978-3-85403-29)
52. Lamy, J. B., Mouazer, A., Sedki, K., & Tsopra, R. (2021). Translating the Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) electronic health records to an OWL ontology. In MEDINFO.
53. epSOS project. http://www.epsos.eu/home.html (last accessed: Sept'22)
54. Gaudet-Blavignac C, Foufi V, Bjeloglrlic M, Lovis C. Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for Processing Free Text in Health Care: Systematic Scoping Review J Med Internet Res 2021;23(1):e24594URL: https://www.jmir.org/2021/1/e24594
55. Snowstorm Github. https://github.com/IHTSDO/snowstorm (last accessed: Sept'22)

56. Rajput, A. M., Triep, K., & Endrich, O. (2022). Semi-Automated Approach to Map Clinical Concepts to SNOMED CT Terms by Using Terminology Server. Studies in health technology and informatics, 293, 67-72.
57. Elkin, P. L., Brown, S. H., Husser, C. S., Bauer, B. A., Wahner-Roedler, D., Rosenbloom, S. T., & Speroff, T. (2006, June). Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. In Mayo Clinic Proceedings (Vol. 81, No. 6, pp. 741-748). Elsevier.
58. OMOP ATHENA. https://athena.ohdsi.org/search-terms/start (Last Accessed: Sept'22)
59. Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. Methods of information in medicine, 32(4), 281–291. https://doi.org/10.1055/s-0038-1634945
60. Callahan, T. J., Stefanski, A. L., Wyrwa, J. M., Zeng, C., Ostropolets, A., Banda, J. M., ... & Kahn, M. G. (2022). Ontologizing Health Systems Data at Scale: Making Translational Discovery a Reality. arXiv preprint arXiv:2209.04732.
61. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., ... & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology, 25(11), 1251-1255.
62. Kaliyaperumal, R., Wilkinson, M. D., Moreno, P. A., Benis, N., Cornet, R., dos Santos Vieira, B., ... & Roos, M. (2022). Semantic modelling of common data elements for rare disease registries, and a prototype workflow for their deployment over registry data. Journal of biomedical semantics, 13(1), 1-16.
63. Xiao, G., Pfaff, E., Prud'hommeaux, E., Booth, D., Sharma, D. K., Huo, N., ... & Jiang, G. (2022). FHIR-Ontop-OMOP: Building Clinical Knowledge Graphs in FHIR RDF with the OMOP Common Data Model. Journal of Biomedical Informatics, 104201.
64. Santos, A., Colaço, A. R., Nielsen, A. B., Niu, L., Strauss, M., Geyer, P. E., ... & Mann, M. (2022). A knowledge graph to interpret clinical proteomics data. Nature biotechnology, 40(5), 692-702.
65. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., ... & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys (CSUR)*, *54*(4), 1-37.
66. Tiddi, I., & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. Artificial Intelligence, 302, 103627.