InteropEHRate
EHR in people's hands across Europe

Towards Interoperable Health Data: the Contribution of the InteropEHRate Project

White Paper

May 2022

InteropEHRate project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826106.

This document has been produced in the context of the InteropEHRate Project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826106. All information provided in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose.

InteropEHRate is a research and innovation project funded by the Horizon 2020 programme that focuses on developing open specifications to allow citizen-centred data sharing. It is due to complete its work by September 2022.

Authors and Contributors

Gábor Bella, University of Trento, Italy

Stefano Dalmiani, Fondazione Toscana Gabriele Monasterio, Italy

Patrick Duflot, Centre Hospitalier Universitaire de Liège, Belgium

Vincent Keunen, Andaman7, Belgium

Luc Nicolas, EHTEL, Belgium

Diane Whitehouse, EHTEL, Belgium

Final version May 2022

Cover photo: proto-cuneiform tablet, courtesy of the Louvre Museum, Paris, France.

Table of Contents

Abstract

This White Paper, written by InteropEHRate project partners, addresses **challenges** related to European health data exchange across institutions, regions, and countries. It also draws its arguments from a series of meetings held with experts in semantics and other fields.

The paper targets **a heterogeneous readership** that includes technologists and healthcare practitioners. More specifically, it is oriented towards hospital chief information officers (CIOs) and data managers, decision-makers and application developers from the healthcare information technology (IT) industry, and clinicians and medical directors concerned about the reuse of health data produced in clinical settings.

First, the paper provides **an overview** of the problems of data heterogeneity and interoperability, which are especially complex due to the breadth and the ever-evolving spread of the health data domain. Second, it presents **a solution** based on an in-depth and exhaustive analysis of data, made possible through innovative semi-automated methodologies and tools developed in the framework of the InteropEHRate project. Third, it shows how these **tools and methods** were applied in complex clinical and research scenarios requiring cross-border data interoperability. Fourth, the paper outlines some conditions that are needed in order to be able to **exploit health data**. Last but not least, **conclusions** are reached.

Introduction

Health data interoperability has been a challenge for decades and, despite major advances, as of today it remains an open problem. Rather than being the solution, the use of national and international standards as a *lingua franca* for data exchange has itself become a never fully resolved challenge. The complexity and the ever-shifting nature both of standards and of the healthcare landscape have so far prevented—and in the authors' view will always prevent—attempts at all-encompassing standardisation. Despite the promise of "miracle" technological solutions, such as ontologies in the 2000s and neural networks in the 2010s, it is our position that no serious attempt at health data interoperability can be based on any single "easy win" approach.

Instead, innovation in data interoperability means finding the optimum compromise between possibly contradictory needs. The first need is for **a meticulous transformation of data**—initially destined for local primary use, thus based on implicit local conventions, and often consisting of informal natural-language text—into a fully specified, correct, and unambiguous representation. The second need is to be able to perform such transformations efficiently and regularly on **ever-growing quantities of health data**.

InteropEHRate innovates in terms of **methodologies and tools**, by allowing data transformations that are agile and yet manage to delve deep into every single data value and into natural-language text expressed in multiple European and global languages. By design, the underlying semi-automated InteropEHRate methodology always involves **human experts**—knowledge and data managers—who maintain automated processes and ensure the quality of the output.

The InteropEHRate project has piloted three challenging **cross-border clinical and research scenarios**. The methodology and tools developed in the project and presented in this White Paper have been put to the test on a realistically broad set of health data involving **hospitals from four European countries**. This experience results in insights for **future exploitation and use** of the InteropEHRate tools and methods.

1. Why is Interoperability over Health Data Hard to Achieve? Problems and Existing Solutions

Words do not travel well across borders. Internationally-agreed and maintained codes do—whether this is across organisations, countries, languages, or domains. Health data, however, are "living organisms" which do not accept to be captured easily and many conditions need to be met to achieve satisfactory results.

Health data coding has become an activity of considerable importance. In many countries, it has proved **difficult to prioritise health data interoperability** given the efforts which need to be developed by all actors to achieve satisfactory results. Practice evaluation and decision support were usually considered as the main incentives for engaged clinicians to code data, provided that they were supported by the necessary tools (Middleton et al., 2016). In many countries, the adoption of pay-for-performance schemes has obliged healthcare organisations to code data, but in many cases this coding remains disconnected from clinical practice. Even if most clinicians appreciate the fact that medical information can be summarised and sorted in a meaningful way, they do not see the need to invest their activities further and become more involved (since most healthcare encounters take place in a limited geographical scope).

Several new developments have made health data coding even more of a priority. The recent focus in 2020-2022 on Machine Learning (ML), Artificial Intelligence (AI) and European Health Data Space (EHDS) is bringing new impetus to this strategic element of interoperability. The series of waves underpinning the COVID-19 pandemic in 2021-2022 has also contributed to move semantic interoperability higher on the eHealth agenda. Data needed to flow quickly from the point of care to epidemiologists and to be analysed on the flow. Public safety and the capacity to adapt quickly to the evolution of the pandemic's variants was everywhere a priority. Value-based medicine also requires the patient to be put at the centre. Here again, semantic interoperability can play a key role by making medical data both accessible and understandable by non-specialists.

Today, data structuring and coding is thus becoming essential. Many countries and organisations are, however, still **ill equipped** to act on this. They experience a dire lack of national info-structure and tools, and a majority of products continue to rely on proprietary codes.

1.1 The Depth and Breadth of Heterogeneity in Health Data

Interoperability is a solution to the problem of data heterogeneity which, in turn, is a consequence of the complexity of the healthcare domain.

This complexity is multi-faceted: it involves, among others, the breadth and depth of domain knowledge; the technological constraints that shape the way data are represented; and the heterogeneity present across institutions, regions (especially in decentralised health systems in countries such as Germany, Italy, and Spain), and countries. From a data perspective, heterogeneity is present in the language in which data is written; the terminologies and codings used; and the way data is structured. Heterogeneity is sometimes clearly in sight, such as when natural-language text is used within data. At other times, the heterogeneity is hidden in implicit interpretations and everyday practices. For example, a "quantity prescribed" data attribute in a prescription may contain the value "5" without any indication as to whether it refers to the number of tablets or packs. Interoperability means that such implicit assumptions all need to be understood, made explicit, and formalised so that data can be exchanged and correctly interpreted on the receiving side.

The effort of **undertaking an in-depth understanding of data**—to which the term “semantic” in “semantic interoperability” refers—thus cannot be avoided.

1.2 Adoption of Standards is Easy on Paper but Hard in Practice

International standards have been used in the healthcare domain for **more than a century** (for example, ICD has existed since 1893). Thus, healthcare practitioners and administrators are well aware of both the benefits and limitations of standardisation. The adoption of international healthcare data standards and data schemas is, however, **complex and challenging**, as is the cooperation that needs to take place among standardisation bodies.

Any non-trivial standard is **a balancing act** between universality (i.e. applicability in a wide range of contexts through time) and ease of understanding and implementation. For instance, language and terminology tend to be relatively stable and evolve slowly. The same observation is valid, although to a lesser degree, for codings and classifications. In consequence, standards on these levels have been applied with a certain success, although not without considerable effort (e.g. The International Classification of Diseases (ICD) is in its 11th version). Even these international standards often prove to be insufficient, and are frequently extended on national, regional, or institutional levels (e.g. experts deem the national extensions of ICD not to be fine-grained enough).

Data schemas are even more **complex, diverse, and rapidly changing**, and related standards have gone through many evolutionary cycles. Health Level Seven (HL7) has existed since 1989, and *Clinical Document Architecture* (CDA) was introduced as part of HL7 v3 to standardise electronic health records (EHRs) in 2000. Soon, due to diverging industry implementations, *Consolidated CDA* was introduced. Still, HL7 was eventually considered to have become too complex. Attempts to simplify HL7 have led to the development of new standards, such as Fast *Health Interoperability Resources* (FHIR) for healthcare or *Observational Medical Outcome Partnership Clinical Data Model* (OMOP CDM) for use in research. This dialectic of standardisation has not stopped since it began, nor is it expected to reach long-term stability any time soon.

For standards of such complexity, the existence of mutually incompatible—and thus, by definition non-standard—implementations is not an exception but the **unavoidable reality of the healthcare world**. Attempts at interoperability that rely solely on standards only work so far as the partners involved are genuinely motivated in successful data interchange and are ready to pay the huge price of understanding and implementing them to the letter. The long-term experience of using and misusing standards in healthcare has two sides to it. On the one hand, standards are an absolute necessity for successful communication beyond single organisations; on the other hand, however, they are far from being sufficient in themselves.

Today, there are **increasing pressures to shift towards interoperability** among standards. Standardisation bodies which once used to function independently have started to converge on their efforts, as shown, for example, by the JIC initiative[1] since 2020. Major standardisation actors have initiated collaborations with an increasing number of mappings between them. Although considerable progress has been made, a lot remains to be done before the goal of a fully integrated standardisation ecosystem can be reached.

[1] Joint Initiative Council, http://www.jointinitiativecouncil.org/

1.3 Even Advanced Technologies are Not Enough

The formal, electronic representation of standards has been shown to be useful in automating data exchange, through term bases such as SNOMED CT, and codings such as LOINC. On the level of data schemas, ontologies have been proposed and widely publicised as the ultimate solution to interoperability (especially in the first decade of this century). It is easy to see, however, that **the complexity inherent in the use of standards**, such as the problem of appropriateness to local needs or the high cost of implementation, is not addressed by technology alone. Still, ontologies have found their place as one of the many bricks among the building blocks that constitute modern data integration systems.

If AI technology in the 1990s and 2000s was knowledge-centric, the decade of the 2010s saw an impressive surge in statistical and neural learning-based approaches. These developments were applied very successfully in certain well-defined segments of healthcare, such as prediction in medical imagery or natural language processing of medical text (Wang et al., 2018). Their large-scale adoption, however, has not happened, despite huge public investments into ambitious industrial solutions such as Watson Health (Ross and Swetlitz, 2017).

More specifically in the area of data interoperability, there are two major reasons for the **slow adoption of AI-driven techniques**: (1) the lack of explainability and traceability, and (2) the problem of context-dependent results.

Explainability means that a system offers the means to the human to understand how it reached a certain result, while **traceability** means that the provenance of every piece of data is made explicit (Holzinger et al., 2019). In healthcare, these are major requirements motivated by the need to understand and manage risks, above all for use in care, but also for certain secondary uses such as research. In the context of data interoperability, mappings, conversions, and translations need to be traceable and explainable, e.g. by referring to the crosswalks used (i.e. how the conversion has been performed). Learning-based AI technology, however, behaves as a black box that, once deployed, is hard to control and offers no explanation about the predictions or transformations it produces. As the responsibility for taking decisions with respect to patients lies ultimately on the shoulders of physicians, a lack of understanding and control of this black box poses a major problem for the clinicians.

Context-dependence refers to the fact that any learning-based AI system can lose quality if it is used in a context different from where it was originally trained. This is due to assumptions behind the data (e.g. units of measure, scope, interpretation of words) that are left implicit in the input, and that may “silently” change when the solution is transferred elsewhere. Retraining a large AI system from scratch is extremely onerous, and the very idea of machine learning is to train once and reuse as many times as possible. While such cases of reuse are realistic in certain well-circumscribed scenarios (such as the recognition of tumours in X-ray images, where it is assumed that the human morphology does not drastically change across hospitals), context-dependence is much more of a problem in types of healthcare data, such as health records in general or free text in particular, that incorporate many implicit human assumptions. While partial solutions for the AI-based parsing of specific types of data do exist—and InteropEHRate relies on some of them, as exposed in Section 3—a robust machine understanding, conversion, and translation of health data in its generality are still unsolved problems, despite remarkable progress in AI research (Laï et al., 2020).

1.4 The Time Dimensions of Evolution and Adaptation

Health data is evolving and adapting over time. The European health data landscape keeps evolving, as it also does anywhere else in the world.

Several **changes** are taking place in tandem. The EU is pushing for ever-growing continent-wide integration; knowledge of the domain is increasing (with data on new diseases and new procedures); the market is expanding (through new drugs, new devices, and IT applications); and health data is used and reused in increasingly varied manners (e.g. research, precision medicine, and wearable devices).

As a result, due to this parallel evolution, on the one hand of local practices within healthcare institutions and, on the other hand, of international standards, **data integration and interoperability cannot be solved by one-shot implementations**.

No single European or international initiative, however complete, can resolve the interoperability challenges once and for all. This is one of the reasons why neither standardisation (tackled in sub-section 1.2) nor technology (handled in sub-section 1.3) can be sufficient on their own: the **agility of adaptation of any solution** is key to maintaining interoperability in healthcare data in the long term.

1.5 Current Practice in Most Healthcare Organisations

A wide range of different practices occur in healthcare organisations, whether in the status of the actual technologies (information and communication technology, ICT) or in the vocabularies, communication models, and languages used to handle health data.

Current **ICT status** in healthcare provider organisations covers a wide range of approaches. The differences include the level of investments in ICT infrastructures and tools, regional and national constraints and regulations, and the fact that the care processes themselves are implemented in a variety of ways. In addition, the situation in terms of infrastructures is very different among (European Union) EU countries depending on the level of investment made by public authorities.

Despite this huge variance, some common points can be identified: first, every IT system needs to have a set of **vocabularies** to represent coded information; second, every IT system is implementing an **information model** (eventually at database level); and third, every IT system communicates with others according to a given protocol. The adoption of standardised models, such as HL7, can facilitate knowledge management in response to the first difficulty, and simplify the handling of the second and the third difficulties.

The evolution of local or national/global vocabularies requires continuous mapping and adaptation, most of the time performed manually on the different copies of the vocabularies. (This is also the case in relation to events, as in the case of the coronavirus (COVID) pandemic that required a prompt extension of vocabularies with COVID-19 codes and related procedures.) Where local codes are preferred, for the many real-world problems not addressed by international vocabularies, a mapping towards the higher-level vocabularies is required, which therefore increases the level of effort to be made by healthcare organisations.

The adoption of a standardised model for communication (e.g. HL7 or Digital Imaging and Communications in Medicine (DICOM)) also requires a continuous adaptation of business logic and internal information models of Hospital Information Systems to support, in a certifiable way, interactions, trigger events, and content of the communication, designed to match with the chosen standard. In the long run, this adoption should decrease the level of investments to be made by the healthcare organisation, even if—at a first glance—it is more expensive (because there is a lack of availability of off-the-shelf solutions supporting state-of-the-art standards, e.g. HL7 FHIR).

In a cross-border healthcare use case, understanding **the foreign language** used by the patient and in any documentation (e.g. records, prescriptions) is a fundamental feature. Most of the time there is no official support for this challenge in a healthcare facility itself. Many clinicians prefer, as a rapid solution, the use of English as a standard communication natural language. However, this solution only works for patients who are skilled in using English vocabulary and terms as applied to the clinical field. Most people are in fact unable to describe precise body parts or symptoms in clinical terms, even in their own language.

2. The InteropEHRate Solution

Any long-term solution to (health) data interoperability has to balance a deep and precise understanding of data with the need for scalable and explainable solutions. This must happen within the context of the complex and always evolving European healthcare landscape.

The InteropEHRate project firmly believes that a truly effective, real-world solution to health data interoperability needs to combine technological and methodological innovations: the use of novel **automated tools** within a **human-centric methodology** that clearly defines the actors and the steps through which the tools are activated.

The tools and methodology proposed within the InteropEHRate project have been implemented as the InteropEHRate Health Services, a data integration system designed to be installed in the premises of healthcare institutions. The role of the system is to **transform local health data into interoperable representations** that can benefit both local data reuse and data exchange with bodies outside the institution.

By 2022, these tools and methodology have twice been acknowledged by the European Commission's *Innovation Radar*[2] for their **outstanding innovation**.[3]

2.1 A Human-Centric Methodology for Automated Interoperability

Healthcare institutions, clinicians, and researchers expect that information systems that deal with health data respect a very strong set of constraints:

- **precision**: transformations must not degrade the data as it could put the lives of patients in risk, either directly or indirectly;
- **responsibility**: a healthcare practitioner or institution are responsible of their decisions and actions with respect to their patients, which also covers their reliance on the output of automated systems; in other words, even if certain processes are automated, the **ultimate responsibility lies on the human experts** (similarly to the case of self-driving cars where, despite the autonomy of the vehicle, the presence and attention of a responsible human is required by law);
- **legal framework**: according to new EU medical device regulations,[4] information systems that perform automated reasoning or prediction for healthcare must be considered as **medical devices**, with all the legal and technological implications that this implies;

[2] https://digital-strategy.ec.europa.eu/en/activities/innovation-radar

[3] https://www.innoradar.eu/innovation/37051

[4] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=uriserv:OJ.L_.2017.117.01.0001.01.ENG

- **explainability and traceability**: automated operations need to be transparent enough so that human experts are able to follow the underlying reasoning and to decide to what extent the results are trustable.

InteropEHRate addresses these requirements by making its methodology human-centred: while data processing is designed to run in a fully automated way, it is overseen by one or more **data managers** who maintain the automated processes, curate their results, and fine-tune them if necessary. Such data scientists and managers are already employed by healthcare institutions as their familiarity with both technology and local data-related practices is now understood to be increasingly essential (Subrahmanya et al., 2021).

The system in charge of data interoperability adopted by the InteropEHRate project is expected to be deployed in the premises of healthcare organisations and research centres that are in control of health data. The **goal of the system** is twofold: to transform local datasets in order to enable their exploitation outside the context of the institution in medical research or cross-border care; and to improve local care through a better integration of local data.

As with any other system with a similar goal, the InteropEHRate system takes local datasets as input and produces representations that fulfil FAIR (findable, accessible, interoperable, reusable) principles. The process consists of the **in-depth understanding** of the data, followed by the **in-depth adaptation** of the data to the required formats (e.g. FHIR), terminologies and codings (e.g. ATC[5], EDQM, ICD-10, LOINC, SNOMED CT), and target language.

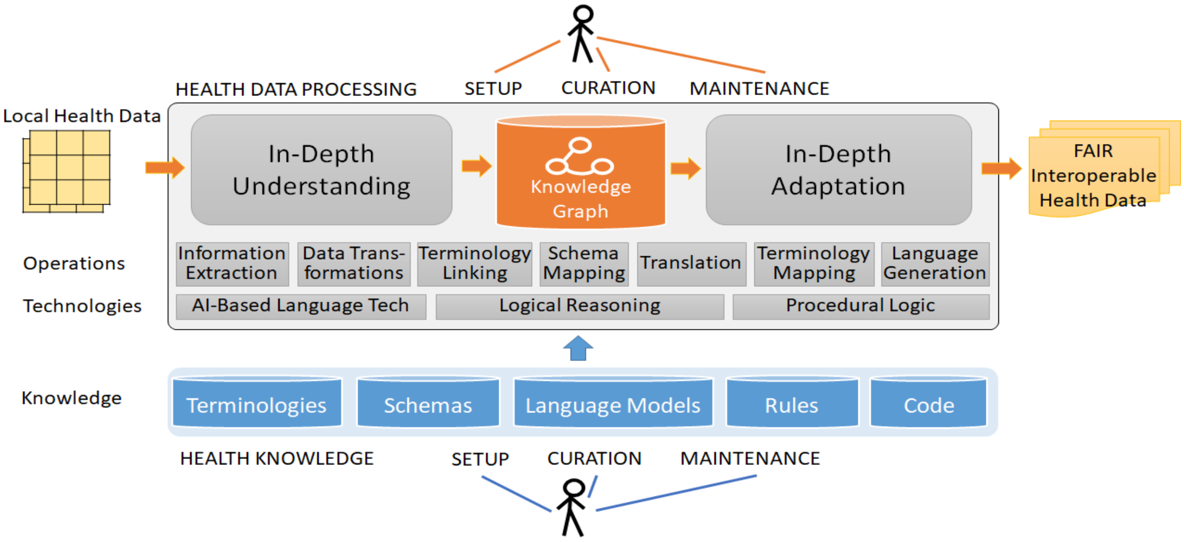


Figure 1: High-level architecture of the InteropEHRate Health Services and the way they are overseen by a human data manager.

Figure 1 above provides a high-level overview of the system. The initial understanding step is crucial: only if the data is formally and unambiguously represented, down to the lowest level of individual data values, is it possible to apply automated transformations without the risk of corrupting health data. In order to ensure the quality of understanding, it is overseen by a data manager who (1) **sets up** (configures) the system before the process, (2) **curates** the results (or samples of the data) during the process, and (3) **maintains** the system setup to follow the evolution of the requirements. The proportion of automated and human intervention can be set according to the level of criticality of mistakes (these may differ e.g. in terms of care vs research vs accounting).

[5] The full names of many of these standards are detailed in Table 1 of this White Paper.

The goal of setup and maintenance is to adapt the system to local needs like data formats, terminologies, practices, and to follow their evolution. To make this process agile, almost all adaptations are carried out through pluggable **knowledge** (depicted in blue at the bottom of Figure 1) rather than undertaking lengthy software development cycles.

In the system, knowledge is understood in a broad sense. It includes ontologies, terminologies, classification hierarchies, codings, machine learning models, rules, and small snippets of code.

2.2 Innovative Solutions

Elements of technologies are already present in existing systems. Examples include formal domain knowledge (e.g. SNOMED CT) and terminology servers that help automate mappings towards standards; information extraction that helps formalise unstructured data; and graphical extract, transform, and load (ETL) tools that provide a more agile method for defining transformations.

InteropEHRate innovates by incorporating all these components into an **end-to-end data integration methodology**, supported by a suite of graphical tools, while simultaneously pushing the boundaries of precision and automation (Bella et al., 2020a). The methodology and the tools support three main outputs that are described below.

In-depth analysis. For a more precise result, data analysis needs to go deep down to the level of individual data values and, within them, individual words if necessary. Unstructured and semi-structured text (i.e. short phrases often found in otherwise structured data) are analysed through multilingual and high-precision (>95% precision) AI-based natural language processing. Extracted terms and codes are converted into a fully formal and language-independent representation, and by doing so the entire dataset contents become a single interconnected **knowledge graph**, as shown (in orange) in the centre of Figure 1. From this deep and precise representation, the subsequent adaptation to the target interoperability profile can be automated via simple and reliable techniques.

Efficient human input. Human supervision of otherwise automated processing is crucial for health data in both primary as well as many secondary use cases (e.g. medical research). In the InteropEHRate Health Services, the data manager plays a central role in the data integration methodology. Firstly, the data manager gives step-by-step instructions to the InteropEHRate Health Services: **the system records human input** that it can subsequently replay in fully automated mode on different but similarly structured datasets. This operation is example-driven: the data manager works on a significant data sample (e.g. a thousand records) that is characteristic of the heterogeneity typically encountered in real life. Secondly, the data manager can **curate** automated results, with single inputs propagated through entire datasets. Thirdly, thanks to ongoing AI research, the automated system will be able to **learn** from the human curation undertaken about any errors made, e.g. in information extraction from unstructured text, and gradually improve its predictions. This learning ability, together with the automated replay of pre-recorded human input described above, guarantee the scalability of human intervention over large datasets.

Explainability and traceability. Through the possible curation of every micro-step, the data manager has the possibility to oversee and understand the entire data transformation process. The curation of micro-steps is optional: the same series of operations can be executed in fully automated mode or interactively, step by step. The human or algorithmic components that intervene in each step are logged by the system. These logs can be reused to annotate results by their producer, ensuring their traceability.

Agile adaptation. The InteropEHRate system is almost entirely **adapted to local needs** through the configuration of knowledge: terminologies and schemas with crosswalks (code mappings), rules, language models, and code snippets. The main benefit is the agility with which the system can be adapted to the evolution of the healthcare landscape. Various examples of potential changes can be identified. The emergence of a new disease, a new version of ICD, or a new data variable can be immediately addressed by the data manager through extension of the knowledge using graphical tools. Changes are made incrementally. Such incremental evolution steps require minutes to days at most, as opposed to lengthy software development cycles which are typically measured in months.

3. Validation of the InteropEHRate Solution

The system (in this case, the focus is on semantic tools) has been used in the InteropEHRate pilot projects to ensure the cross-border interoperability of a wide range of electronic health record datasets. It has been applied in the following three scenarios[6].

1. **Scenario 1:** Patients upload their health records located in a hospital data centre onto their smartphones. The records are converted into an EU-wide cross-border interoperable representation and, if necessary, are translated into the patients' native tongue.
2. **Scenario 2:** Through a smartphone, a patient provides his/her health record to a hospital in a foreign country that he/she is visiting, where the health data are automatically translated into the local language before being displayed to the healthcare practitioner who is treating the patient.
3. **Scenario 3**: For use in a cross-border medical research study which is being organised, a patient cohort is established automatically by involving citizens from different EU countries. Anonymous research data is extracted automatically from the health records provided.

These three cross-border scenarios require the full power of in-depth analysis and adaptation of health records. For this reason, InteropEHRate has involved hospitals from four countries and has supported five separate languages: English, French, Greek, Italian, and Romanian. The health records used originate from two hospitals: Fondazione Toscana Gabriele Monasterio (FTGM), Pisa, Italy, and the Centre Hospitalier Universitaire (CHU) de Liège, Belgium.

3.1 Experience from Multilingual Data Mapping

The table below sums up the mapping challenges encountered at the Italian and Belgian sites. They all are a consequence of the heterogeneity of data on multiple levels: linguistic (due to moving the data across borders), structural (hospitals are using different data schemas to communicate with the outside world), as well as terminological (different local, national, and international coding system and terminologies).

	FTGM Pisa	**CHU Liège**	**Standards-based representation**
Language	Italian	French	English, Italian, French, Greek, Romanian
Data structures	HL7 CDA, DICOM, PDF	Belgian SumEHR, hospital-internal	FHIR 4.0 + extensions
Coding systems	HL7 codes, ICD-9, LOINC, ATC	Local codes, CNK, ICD-10, ATC	ICD-10, LOINC, ATC

[6] For more detail, see the InteropEHRate White Paper entitled "Unleashing personal health data for care and research: The InteropEHRate approach", https://www.interopehrate.eu/wp-content/uploads/2021/08/InteropEHRate-White-Paper.pdf

Terminologies	regional and national	local and national terminologies	SNOMED CT, UCUM
Dataset types mapped	380 attributes are mapped as part of the following FHIR resources: Patient, Condition, Observation, Diagnostic Report (Laboratory Result, Image Report, Biosignal, Image Study), Encounter, Practitioner, Allergy intolerance, Medication Statement, Medication, Medication Request, Media, Imaging, Care plan, Procedure		

Notes related to Table 1: ATC (Anatomical Therapeutic Chemical), CNK (a unique product code used in Belgium), DICOM (Digital Imaging and Communications in Medicine), FHIR (Fast Healthcare Interoperability Resources), HL7 CDA (Health Level Seven Clinical Document Architecture), ICD-10 (International Classification of Disease), LOINC (Logical observation identifier names and codes), PDF (Portable Document Format), SNOMED CT (SNOMED Clinical Terms—a brand name), UCUM (Unified Code for Units of Measure).

Table 1: Local versus standard representations in the IEHR Italian and French clinical sites

In order to be able to provide all the needed expertise, three distinct sub-roles had to be specified. They related to the local data, the standards, and the data mapping:

- **a local data manager** (one from FTGM, one from CHU) with a deep understanding of local hospital data structure;
- **a standards expert** with a good understanding of the target interoperability profiles (in this case, FHIR-based);
- **a data mapping expert** with the necessary technical experience to use the data mapping tools and adapt them to local needs (e.g. French and Italian language models, and terminologies).

In the case of InteropEHRate, four **people were involved in the process** in total (one Belgian and one Italian local data manager, one FHIR standards expert, and one data mapping expert). Depending on the expertise available at healthcare organisations, these roles can be fulfilled by fewer persons. The people involved, however, have to work in close collaboration due to the complementarity of their expertise.

The **cooperation** needs to extend over all the datasets to be mapped. Achieving a satisfying end result may require intervention in the production of local data but, in some cases, also on the international standard itself. Here, the example used is FHIR, which may not cover certain attributes and/or local needs and may need to be updated through its extension mechanism. The parties involved need to discuss and make decisions together on each problematic case. It was observed that a single FHIR resource type (as in the table above) required, on average, 1-2 combined, full-time working days to define and formalise the mappings of the resource type's 20-30 attributes.

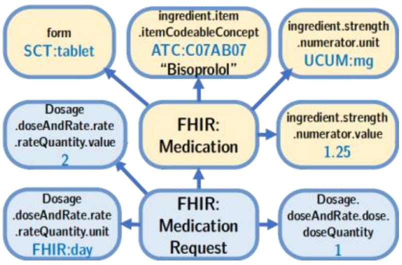
A key experience from the extensive data mapping and transformation activities summarised in Table 1, was that a single person cannot carry out the entire activity. The limitations related to work to be done by a single person were not due to the amount of work but, rather, to the wide-ranging forms of expertise required to do a proper job.

The experience also confirmed the absolute necessity of having access to sufficient **example source data** during the mapping process. Having such example data greatly reduces the effort needed to understand mapping requirements, and is also crucial for defining transformations in a way that is robust enough to handle heterogeneity in the source data.

3.2 Multilingual Information Extraction in Use

As part of the system's in-depth understanding of health data, one of the biggest challenges is to deal with the natural-language text present in both structured and unstructured datasets. In the scope of the InteropEHRate project, the researchers have implemented **robust and high-precision**

information extraction for textual prescriptions, based on the principles of parsing short textual labels within structured data, known as the *Language of Data* (Bella et al., 2020b). The starting point is a text string such as the one in this Italian prescription shown immediately below in three parts:



Prescription
Bisoprololo (Bisoprololo san) 1,25 mg 1 cpr x 2/die

Information extraction generates the structured equivalent automatically, and then converts relevant names and keywords inside the cells into formal terminology equally automatically (from e.g. ATC, EDQM, SNOMED CT, and UCUM). Thus, it achieves a fully language-independent and formal representation of the prescription as a knowledge graph, shown here both in graph-based and tabular forms:

FHIR:Medication					**FHIR:MedicationRequest**		
ingredient	Identifier	strength value	strength unit	form	dose Quantity	rate Quantity	rateQuantity.unit
Bisoprololo ATC:123456	Bisoprololo san	1,25	**mg UCUM:mg**	**cpr SCT:tablet**	**1**	2	**die SCT:daily**

Thanks to the deep, language-independent, formal representation, in turn, the result can be translated automatically in the subsequent in-depth adaptation phase into e.g. French, by using standard high-quality terminology translation:[7]

Principe actif	Produit	Dosage	Unité de mesure	Quantité	Forme pharmaceutique	Fréquence	Période
Bisoprolol	Bisoprololo san	1,25	**mg**	1	**comprimé**	2	**jour**

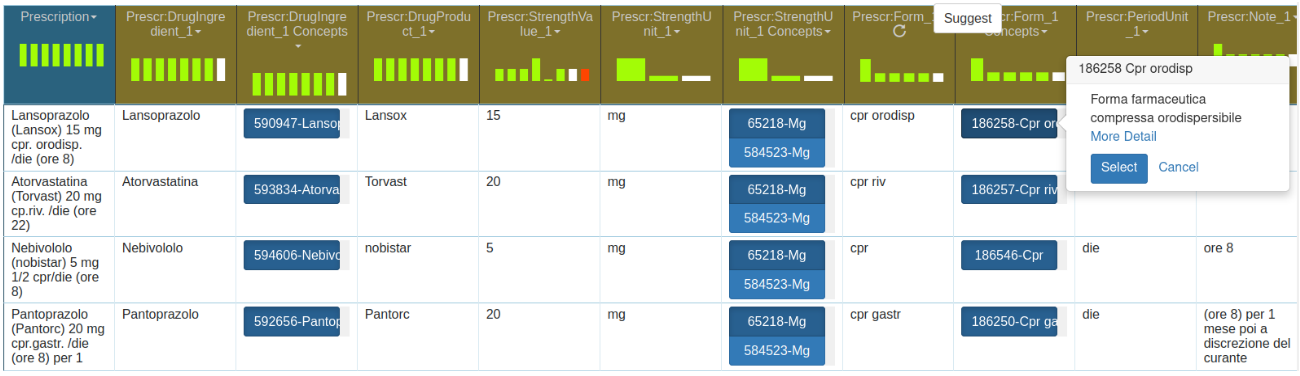
In InteropEHRate, information extraction from prescriptions was implemented for the French and Italian languages, and was tested and shown to have both precision and recall over 90% in both languages.[8] With this high level of quality, it would be a realistic proposal to apply the solution in production, also outside of the scope of the InteropEHRate project. Original data are, nevertheless, always to be kept alongside results obtained through these automated data processing operations. It is the responsibility of the data manager to set up the extraction process accordingly.

For applications where data quality is critical (e.g. in care), human validation and curation of the results are also made possible through a tight integration of information extraction into the data mapper tool and its graphical user interface, as shown in Figure 2 below. The data manager pilots the information extraction process from the data mapper tool, observes the results, and fixes any mistakes introduced

[7] While prescriptions happen to be fairly easy to understand even when written in a foreign language, the principle remains the same for other kinds of unstructured data.

[8] For evaluation results, see the InteropEHRate project deliverable (InteropEHRate, D5.12). For the definition of precision and recall, see https://en.wikipedia.org/wiki/Precision_and_recall.

by automated extraction.



Prescription	Prescr:DrugIngredient_1	Prescr:DrugIngredient_1 Concepts	Prescr:DrugProduct_1	Prescr:StrengthValue_1	Prescr:StrengthUnit_1	Prescr:StrengthUnit_1 Concepts	Prescr:Form_	cr:Form_1 Concepts	Prescr:PeriodUnit_1	Prescr:Note_1
Lansoprazolo (Lansox) 15 mg cpr. orodisp. /die (ore 8)	Lansoprazolo	590947-Lanso	Lansox	15	mg	65218-Mg 584523-Mg	cpr orodisp	186258-Cpr or		
Atorvastatina (Torvast) 20 mg cp.riv. /die (ore 22)	Atorvastatina	593834-Atorva	Torvast	20	mg	65218-Mg 584523-Mg	cpr riv	186257-Cpr riv		
Nebivololo (nobistar) 5 mg 1/2 cpr/die (ore 8)	Nebivololo	594606-Nebivo	nobistar	5	mg	65218-Mg 584523-Mg	cpr	186546-Cpr	die	ore 8
Pantoprazolo (Pantorc) 20 mg cpr.gastr. /die (ore 8) per 1	Pantoprazolo	592656-Panto	Pantorc	20	mg	65218-Mg 584523-Mg	cpr gastr	186250-Cpr ga	die	(ore 8) per 1 mese poi a discrezione del curante

186258 Cpr orodisp

Forma farmaceutica compressa orodispersibile

Figure 2: Screenshot from the Data Mapper Tool of the InteropEHRate Health Services, showing prescriptions in their initial textual and final fully formalised forms

4. Prerequisites for Exploitation and Use

As shown in Figure 1, the InteropEHRate Health Services consist of multiple components that are combined together via a data integration methodology and carried out by a (human) data manager.

This section of the White Paper considers the requirements of using either the entire InteropEHRate system or its **components** by **two types of players** in the healthcare ecosystem. They are considered to be prerequisites in order to be able to exploit/use the InteropEHRate services.

The components for use include information extraction, schema mapping, or terminology mapping. The players involve healthcare organisations (and healthcare professionals—see section 4.1), and companies in the private sector developing healthcare applications (e.g. apps developers—see section 4.2).

4.1 For Healthcare Organisations and Healthcare Professionals

Healthcare organisations, such as hospitals, often already have more or less extensive policies and governance regarding data management for primary and secondary uses, supported by a team of data managers and IT experts. The adoption of the solution proposed by InteropEHRate implies the involvement of these teams for carrying out the underlying methodology.

As shown in Section 2, in order to map, transform, or translate health data, the system needs to acquire an **in-depth understanding** of the data. For this to happen, the system needs to be fed with knowledge about local data and practices by a human data manager. The data manager (1) **sets up the formal knowledge** relevant to the interoperability problem to be solved; (2) **defines the transformations and mappings** on terms, codes, data attributes and values; and (3) **curates and maintains** these transformations and mappings over time.

The data manager typically uses and needs material help in the form of (a) example data; (b) specifications of the interoperable target data representations; and (c) existing crosswalks. Firstly, a sufficient amount of **example data** is crucial for the system to be able to acquire a deep understanding of the heterogeneity and complexity of the data that needs to be addressed. The proposed data integration methodology has a very strong, bottom-up, example-driven element, that is reflected in the operation of the graphical tools. Secondly, the data manager needs access to **detailed specifications and, ideally, also expert advice on target data representations**, e.g. of target data schemas, such as FHIR, and international code systems, such as SNOMED CT, ICD, or LOINC. Experience gathered from real-world data mapping efforts in InteropEHRate shows that achieving a human understanding of the correspondences between source and target representations is much more time-consuming than actually codifying the computational transformations themselves. Thirdly, in order to reduce the effort required to understand and formalise the mappings, the data manager should **reuse existing mappings and crosswalks** between codes, terms, and data attributes as much as possible. Central terminology servers, national or international language systems (an example of which is UMLS[9]), are extremely useful in reducing the efforts involved in finding and obtaining such knowledge.

4.2 For Application Developers

While the InteropEHRate Health Services and Tools, as described above, were designed for use within large healthcare organisations employing professionals paid for knowledge and data management, individual tools and components developed and piloted in the project can also be integrated into third-party applications to add useful functionality to their applications and thus be in a position to support critical use cases such as the ones supported by the IEHR protocols. Apps need to support a minimum level of semantic interoperability, such as field names, unit systems, data and time, to name only the very basic terms. Going **beyond libraries**, the integration of InteropEHRate technologies into applications extends the semantic capabilities thanks to InteropEHRate's cross-mapping tool and its reliance on major internationally used terminologies. The app is able not only to store and share the data in a structured way, but also to translate them into any of the languages supported.

This task on the part of application developers still requires, however, that the information should be validated *ex ante* by the (clinical) data producers. As described in section 3, the system needs **initial investments** to be used, in terms of preloaded knowledge (i.e. crosswalks, ontologies, and translations) and logic (information extraction language models).

One cannot expect that the expertise needed will thus be available in small and medium-sized enterprises (SMEs). It is thus assumed that all **data transformations are executed "upstream"**, i.e. inside the healthcare organisation or by the healthcare professional him or herself, with the support of adequate tools. However, the current reality is that a majority of health data information is delivered in free text or with only a very basic level of structuration.

In the case in which data are delivered in free text by health care professionals, the IEHR tools embedded in the app might also be used in a fully automated way to support specific use cases that are related mainly to secondary use of data and thus provide a critical contribution to the creation of meaningful health data spaces.

Improvements and streamlining of energy-demanding processes have significantly increased the capabilities of **applications running on mobile phones**. Transparent disclosure of the fully automated process and the intended use of the data will, in this case, always accompany the data sharing process.

[9] The Unified Medical Language System, https://www.nlm.nih.gov/research/umls/index.html

5. Conclusions

With a fast-growing number of use cases and an increased focus on data, **health data interoperability** in a global and open context is now placed higher on the agenda of all the actors in the eHealth value chain than it was before.

In the healthcare domain, health data interoperability remains a complex issue. Several challenges and problems need to be handled. One needs to deal with the breadth and depth of domain knowledge, various technological constraints, and the wide heterogeneity that operates across systems, domains and countries. The process at stake consists of two main actions: the **in-depth understanding** of the data, followed by an **in-depth adaptation** to the requirements of downstream use: language, terminologies, and schemas.

Much progress has been achieved during the last decade. The basis for **an integrated semantic ecosystem** is now emerging (see especially sections 1.2 and 4).

AI has brought new perspectives but, overall, the adoption of AI-driven techniques has been slow due mainly to the lack of explainability, traceability, and the difficulty of adaptation to local contexts. Furthermore, data integration and interoperability cannot be solved by one-shot implementations.

The report authors advocate that **a human-centric methodology** needs to be applied for automated interoperability. Specific human skills, coordinated by a data manager, need to be present in the healthcare organisation in order to establish a meaningful and efficient semantic process.

The InteropEHRate approach is thus to combine **state-of-the-art technology** with **a human-centric**, yet scalable, **methodology**. The InteropEHRate Health Services clearly define the actors and the steps through which those tools are activated. While human supervision and efficient processing of large amounts of data are *a priori* contradictory requirements, InteropEHRate aims to bridge this gap through using technological and methodological innovations that improve the efficiency of human input and let the IT system learn from it.

The current reality is that most health information is delivered in free text or with a very basic level of structuration by healthcare organisations. While this phenomenon will evolve over time, in the meantime, the **possibility of fully automated semantic transformation** has been raised. The use of the **health tools developed by InteropEHRate** might then still be considered for limited use cases related mainly to secondary use of data. This situation would place the **citizen/patient in the position of the data controller** (provided that adequate information on the process and the intended use always travel with the data).

References

- Bella, G., Elliot, L., Das, S., Pavis, S., Turra, E., Robertson, D., and Giunchiglia, F. (2020a) *Cross-Border Medical Research using Multi-Layered and Distributed Knowledge*. Prestigious Applications of Intelligent Systems, ECAI 2020, Santiago de Compostela, Spain.

- Bella, G., Gremes, L., and Giunchiglia, F. (2020b) *Exploring the Language of Data*. Proceedings of the 28th International Conference on Computational Linguistics (COLING).

- Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). *Causability and explainability of artificial intelligence in medicine*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, *9*(4), e1312.

- Laï, M.C., Brian, M. and Mamzer, M.F. (2020) *Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France.* Journal of translational medicine, 18(1), pp.1-13.

- InteropEHRate project deliverable D5.12: Design of information extractor and natural language translator—v2.

- Middleton B, Sittig DF, and Wright A. (2016) *Clinical Decision Support: a 25 Year Retrospective and a 25 Year Vision.* Yearbook of Medical Informatics. 2016;Suppl 1(Suppl 1):S103-S116. Published 2016 Aug 2. doi:10.15265/IYS-2016-s034

- Ross, C. and Swetlitz, I. (2017) *IBM pitched its watson supercomputer as a revolution in cancer care. It's nowhere close*, STAT Investigation.

- Subrahmanya SVG, Shetty DK, Patil V, Hameed BMZ, Paul R, Smriti K, Naik N, Somani BK. (2021) *The role of data science in healthcare advancements: applications, benefits, and future prospects.* Irish Journal of Medical Science. DOI: 10.1007/s11845-021-02730-z.

- Wang, Y. et al. (2018) *Clinical information extraction applications: a literature review*. Journal of biomedical informatics, 77, pp.34-49.

The tools and methodology proposed by the InteropEHRate project have been cited twice for outstanding innovation by the European Commission’s *Innovation Radar*.

https://www.innoradar.eu/innovation/37051

InteropEHRate
EHR in people’s hands across Europe

Web: www.interopEHRate.eu

Email: info@interopehrate.eu