TITLE:

New implementation of data standards for AI research in precision oncology. Experience from EuCanImage.

AUTHORS:

Teresa García-Lezana[1], Maciej Bobowicz[2], Santiago Frid[3], Michael Rutherford[4], Mikel Recuero[5], Katrine Riklund[6], Aldar Cabrelles[1], Marlena Rygusik[2], Lauren Fromont[1], Roberto Francischello[7], Emanuele Neri[7], Salvador Capella[8], Fred Prior[4], Jonathan Bona[4], Pilar Nicolas[5], Martijn P. A. Starmans[9], Karim Lekadir[10,11], Jordi Rambla[1,12] and EuCanImage Consortium.

AFFILIATIONS:

1. Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr.Aiguader 88, Barcelona 08003, Spain
2. 2nd Department of Radiology, Medical University of Gdansk, Mariana Smoluchowskiego 17, 80-214, Gdansk, Poland
3. Clinical Informatics Service, Hospital Clínic de Barcelona, Villarroel 170, 08036 Barcelona, Spain
4. Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, United States
5. Social and Legal Sciences Applied to the New Technosciences Research Group, University of the Basque Country (UPV/EHU), Bilbao, Spain
6. Department of Diagnostics and Intervention, Diagnostic Radiology, Umeå university, Umeå, Sweden
7. Academic Radiology, Department of Translational Research, University of Pisa, Via Roma 67, 56126 Pisa, Italy
8. Barcelona Supercomputing Center (BSC), Jordi Girona 29, 08034, Barcelona, Spain
9. Department of Radiology and Nuclear Medicine and Department of Pathology, Erasmus MC Medical Center, Rotterdam, the Netherlands
10. Artificial Intelligence in Medicine Lab (BCN-AIM), Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain
11. Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, Spain.
12. Universitat Pompeu Fabra (UPF), Barcelona, Spain

FUNDING: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952103.

KEYWORDS: interoperability, data model, precision oncology, artificial intelligence, FHIR.

ABSTRACT:

An unprecedented amount of personal health data, with the potential to revolutionise precision medicine, is generated at healthcare institutions worldwide. The exploitation of such data using artificial intelligence relies on the ability to combine heterogeneous, multicentric, multimodal and multiparametric data, as well as thoughtful representation of knowledge and data availability. Despite these possibilities, significant methodological challenges and ethico-legal constraints still impede the real-world implementation of data models. The EuCanImage is an international consortium aimed at developing AI algorithms for precision medicine in oncology and enabling secondary use of the data based on necessary ethical approvals. The use of well-defined clinical data standards to allow interoperability was a central element within the initiative. The consortium is focused on three different cancer types and addresses seven unmet clinical needs. This article synthesises our experience and procedures for healthcare data interoperability and standardisation.

BACKGROUND

Artificial intelligence (AI) for precision oncology is an exponentially growing field[1] built over large amounts of patient-related data. The volume and depth of personal health data necessary to support precision medicine is unprecedented, and the integration and analysis of such heterogeneous data types require a thoughtfully structured representation of knowledge[2,3]. Besides, for a broad implementation, data needs to be shared across diverse institutions and even across multiple nations, increasing the complexity to data integration and flows.

Personal health data includes a wide range of different data types, among others, demographic characteristics, patient's symptoms, diagnoses, laboratory results, medications, imaging data and genomics. Generating and collecting the mentioned data types involve a multitude of different technological platforms and their storage in a wide range of data formats and information systems, contributing to increased data diversity[4]. In fact, healthcare data has been shown to be more heterogeneous than other types of research data[5]. This high data complexity, together with the cross-border legal constraints and massive data volumes required, makes interoperability of healthcare information a significant hurdle in the development of AI models[6,7]. The barrier is even more prominent in oncological observational research since cancer diagnoses require a set of attributes usually registered separately (imaging, histology, topology, stage, grade, and biomarkers), and the complex patient trajectory involves personalised treatment regimens[7].

Data needs to be harmonised at three levels to address the interoperability challenge: technical, syntactic and semantic. On one hand, requirements in technical interoperability facilitate basic data exchange conventions (file formats), and on the other hand, syntactic and semantic interoperability define data structure and the use of ontologies for unambiguous representation of medical concepts, respectivelly[8]. Across the different healthcare ecosystems, equivalent information can be represented in diverse ways. The use of standards that can be universally interpreted, both human and machine-readable, facilitates harmonisation efforts. The structured exchange of health-related data is supported by international standards, such as the Health Level Seven (HL7®) or fast healthcare interoperability resources (FHIR®) specification. FHIR defines the structure of medical data in modular components called "Resources"[9]. It is envisioned that the FHIR framework can become critical for implementing AI technologies in the health sector, just as Digital Imaging and Communication in Medicine (DICOM) or Picture Archiving and Communication System (PACS) for imaging data[4].

The EuCanImage consortium comprises multidisciplinary teams with the overall aim of building AI models integrating imaging, clinical and phenotypic data from different EU countries to improve cancer patients' outcomes. Here, we synthesise our experiences as an

overview of the methods and challenges we identified while working towards the standardisation and interoperability of healthcare data and implementing a data model for AI in oncological research.

DATA DESCRIPTION

Purpose for data collection and data description

EuCanImage is a complex project centred around addressing seven key unmet clinical needs in cancer imaging. A multidisciplinary team of different speciality physicians (radiologists, oncologists, radiotherapists, surgical oncologists, pathologists), sociologists, psychologists, AI developers, Small and Medium Enterprises (SMEs), imaging and oncology research associations, as well as patients' organisations, collaborated to identify and refine core research questions. The consortium pinpointed the most urgent topics in liver, colon and breast cancer and designed seven clinical use cases to respond to each specific clinical need. More specifically, we concentrated our efforts on one use case on hepatocellular carcinoma, three on colorectal cancer (one on colorectal liver metastasis and two on rectal cancer) and three on breast cancer. Our ambition is to integrate clinical, pathological and genetic data (non-imaging data) and radiological images (imaging data) to build algorithms going beyond standard practice, allowing personalised approaches informed by the best quality data. The initial effort to obtain such high-quality data was dedicated to defining clinical consensus and requirements for the use cases with specifications of clinical data variables.

The mentioned factors necessitate a comprehensive data model incorporating multifactorial inputs from multiple data sources. In EuCanImage, data is submitted from six university hospitals in Italy, Lithuania, Poland, Spain, and Sweden, national registries and two research institutions from the Netherlands. Each of the centres uses its own imaging infrastructure and PACS as well as electronic or paper health records that include demographic, clinical, pathological and phenotyping information recorded in Health Information Systems.

Regarding the clinical data defined for each use, some common variables exist for all use cases: patient ID, biological sex, age at diagnosis, diagnosis, and pathology (ICD-O-3 codes). On the other hand, there are use case-specific variables such as the hormone receptor status, HER2 mutational status, or Ki67 status for breast cancer, or information on specific chemotherapy agents with dosing regimens. The dialogue between physicians and AI developers on clinically relevant variables that can be meaningfully incorporated in AI algorithms, with the GDPR-compliant data minimisation principle, led to the final set of defined variables (Figure 1).

The final number of clinical variables used as ground truth or additional input parameters varies between 8 and 39 parameters per use case. Three levels of data provision were defined: minimal, mandatory and recommended. The minimal set contains essential information from the pathology assessment of specimens, e.g., cancer vs. other findings and the presence of complete pathological response vs. partial or no response. It allows the assembly of standard-level algorithms primarily using imaging information as input with

pathology information as ground truth. The mandatory set contains important enriching information. These are all variables that should be included as input together with cancer images for more advanced and complex algorithms. Finally, the recommended set addresses additional clinical data points (e.g., risk factors for breast cancer) and phenotyping information (PAM50 results) available only from selected centres but with adequate numbers of patients for AI research. This recommended set would create a very interesting and promising asset in the project repository for future research that is otherwise not readily available from other data repositories.

Next, the value sets for each clinical variable were discussed in detail and defined to ensure understanding and a manageable range of values. It allows good description of cohorts and, at the same time, prevents very fine-grained stratification of data with limited instances and unbalanced distribution in some of the cohorts. This consensus approach represents a compromise between the need for a precise representation of the clinical range of disease presentations and the goals of data clarity and homogeneity.

DATA CURATION

Analysis of semantic interoperability and health standards

Semantic interoperability represents a remarkable challenge for medical research. Data captured through health information systems are usually stored in locally-modelled clinical repositories, mostly in non-structured ways, thus hindering cross-national data source integration and translational research. Health information standards play a crucial role in defining the structure and meaning of clinical information for it to be unequivocally interpreted by different systems. However, there is no single standard that solves every need in the biomedical field, but rather different standards that either complement each other or compete with one another. This includes standardised vocabularies and classifications and also health information standards. Examples of vocabularies include the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)[10] or International Classification of Diseases (ICD)[11], Logical Observation Identifiers Names and Codes (LOINC)[12], OHDSI Standardised Vocabularies, and International Cancer Genome Consortium[13] - Accelerating Research in Genomic Oncology (ICGC ARGO[14]). Examples of health information standards include HL7 FHIR[15] and open Electronic Health Records (openEHR)[16].

Vocabularies and classifications represent concepts that pertain to the biomedical domain in a standard fashion[17], although they require a common structure that provides the syntactic interoperability required to achieve semantic interoperability. Common Data Models (CDMs) serve as representations of collected data aimed at facilitating the exchange, pooling, sharing, or storing data from multiple sources, and can provide this common structure[18].

Health information standards also provide a syntactic base to allow the formal representation of the structure of clinical information and its meaning.

FHIR was introduced in 2011 by the standard-developing organisation HL7[15]. The information within FHIR is organised in basic building blocks named Resources. Those blocks define the structure of the contained information. Although it is widely used in health informatics, its uptake in research environments is less prevalent[19]. Most studies using FHIR in health research focus on clinical research (including clinical trials), and just about 12% are oncology-related[20]. In these studies, FHIR has been mainly used for standardisation and data capture and to a lesser extent for data analysis[20].

Observational Medical Outcomes Partnership (OMOP) enables the systematic analysis of disparate observational databases through a common data model and a closed dictionary of terminologies, vocabularies, and coding schemes. Several authors consider it an adequate data model for sharing data in electronic health record (EHR)-based longitudinal studies[18,21].

ICGC ARGO[14] is an initiative that provides a fixed schema for creating 15 clinical tables oriented to genomic oncology research, thus oriented for addressing cancer-specific issues in the representation of clinical data. Unlike other CDMs and health information standards, its use of standardised vocabularies and classifications is limited.

Data Model Design

Data standardisation within the project is required to a) support content organisation and subsequent development of AI algorithms, and b) facilitate interoperability and the secondary use of the data. After thoroughly evaluating various CDM alternatives, we decided to use FHIR due to its wide adoption, flexibility and suitability for real-world data exchange.

As previously outlined, clinical elements necessary for each hypothesis were established by domain experts and interdisciplinary teams including clinicians and AI developers, considering different key data aspects. For the purpose of the project, data to cover the seven use cases was arranged following five different data schemas. As a general overview, the highest level components of the FHIR model are the Resources, which contain hierarchical sub-layers of descriptive elements for more detailed data classification. The content and format of a Resource has controlled properties, meaning that the different data elements and data types allowed need to follow specific requirements. To design the data architecture needed for each EuCanImage use case, the following FHIR resources were identified as relevant: Patient, Condition, Observation, Procedure, Medication Administration and Diagnostic report (Figure 2A). When choosing the most suitable resource for each selected variable, each resource's constraints were considered. Once the variables were assigned to a Resource, they were mapped to the appropriate FHIR element. In cases where the clinical variables could be assigned to more than one suitable profile (ex:

Histological type), simplicity criteria were applied to minimise the number of Resources used. Many data elements within the FHIR Resources require coded values. Some are fixed values defined by the FHIR specification, but others require external ontologies. As a general rule, HL7/FHIR terminology was used in a few established fields, basically status profiles. SNOMED was the preferred terminology for general clinical concepts, ICD-O3 for histology, LOINC for some test observations and RxNORM for medication. We used NCIT when the concept did not exist in previous ontologies (Figure 2B). The summary of the different stages we followed to conceptualise the data model is described in Figure 3.

It is essential to point out that the project presented some particular needs not fully represented by FHIR (and standards ontologies), requiring alternative paths solutions to overcome limitations. The gaps identified relate to the fact that FHIR was designed to support interoperability and data exchange in healthcare rather than specifically focusing on research needs. The primary limitations we faced were a) the need to represent concepts without available standard terminology, b) variables not structured as in healthcare practice, c) the representation of dates to comply with the de-identification of personal data, d) the representation of not provided (missing) information, e) the implementation of the model without a FHIR server.

For each clinical variable, we defined the limited set of permissible values (value set) that this variable can adopt. Some of these value sets need to include ambiguous concepts for simplicity and data harmonisation reasons. An example is the term ‘other’, which is required to group less frequent or more irrelevant values. Those terms, isolated from additional context, suppose a challenge for interoperability. We used SNOMED post-coordinated expressions to build more specific clinical ideas by combining relevant terms with compositional grammar. Another challenge posed were concepts that are not used in healthcare but are essential to contextualise the specific use cases for research purposes and which are not captured by standard terminologies. Some examples are the variable ‘breast cancer subtype-by proxy’ to group patients according to hormone receptor expression levels or ‘time interval between the end of the neoadjuvant treatment and surgery’. Additional difficulties include tumour grading systems such as modified Ryan Scheme for Tumor Regression Grade, Miller and Payne's Tumor Regression Grade, Residual Cancer Burden class or grading of DCIS. Our approach was using the specific grading scales in NCIT, if available, or using generic grading scales in SNOMED (ex: grade 1 on a scale of 1 to 5), despite the fact it could affect interoperability.

There were some variables characterised under the Medication Administration Resource that presented significant difficulties to be represented as needed for the purpose of the project. To detail the chemotherapy dosage, we required the total number of chemotherapy cycles, dose (amount of medication per dose) and the accumulation dose within the same

'Medication administration' entry. However, that Resource is designed to collect that information differently (single entry).

In compliance with GDPR, personal data pseudonymization entails the removal of indirect data identifiers, such as dates. The collection of exact dates was replaced by the collection time intervals (months, weeks, etc.). Most FHIR Resources allow time periods as valid data types, however the Resource: Medication Administration, only allows dates (*ddmmyy*). To fulfil this FHIR restriction we recodified the time periods into arbitrary dates starting on January 1st of 1970 to mimic the Epoch Unix system, with the end date calculated based on the collected time interval and starting date in mind.

TECHNICAL IMPLEMENTATION OF DATA STANDARDS

Transforming and loading 'raw' data from various hospital data systems into the newly developed data schema proved to be challenging and labour-intensive. The FHIR implementation format has a hierarchical architecture, that, while having many advantages to encode the relationships between the variables and facilitating data storage, supposes an additional barrier for data providers given that most of the required information was not structured data inside their health records. To minimise the need to re-encode and simplify the data capture process, we created electronic case-report forms (eCRF) with REDCap. REDCap is a secure web application that supports data capture primarily for research studies[22,23]. This software allows the custom design of data entry forms and data collection workflows. It features a user-friendly interface to design the forms, field validation, custom logic patterns, calculated fields, data import/export options, data quality control and role-based user access. Additionally, it offers a set of APIs for integration with other platforms[24]. REDCap was deployed at the European Genome-phenome Archive (EGA) servers to design and manage data entry forms for clinical data collection within the consortium. Different data entry forms were conceived to support each of the five different data schemas.

Data from hospitals can be imported to the EuCanImage REDCap database following two paths: 1) by directly filling the online forms or 2) by entering data into CSV files complying with the specific REDCap format requirements, and then uploading the files into REDCap. Patient IDs were previously pseudo-anonymised at the hospitals, and only hashed patient IDs (EuCanImage ID) were introduced in the platform. Consequently, all related clinical data from the different institutions merged into a single harmonised database for each use case. Once harmonised data was at the database quality control checks were performed. Then, all data was exported from REDCap as a CSV file for conversion into FHIR-compliant files.

To implement the FHIR model, we created each Resource using individual persistent identifiers with Uniform Resource Names (URN), more specifically with Universally Unique Identifiers (UUID). These identifiers were generated for each patient, resource and bundle.

In FHIR, a bundle is a way to envelope all the Resources belonging to a single patient. In our case, a bundle is generated from a single row of the exported CSV files.

For the subsequent data standardisation stage, we built Extract Transform Load (ETL) pipelines to transform the output CSV files into JSON files compliant with the FHIR schema. This process was automated per use case using different Python scripts (available on Github: https://github.com/EGA-archive/EuCanImage-FHIR/), with the additional use of external validators, such as FHIR Validator GUI (https://validator.fhir.org/) and Simplifier (https://simplifier.net/). While building the Python scripts, the mapping of the dictionaries was hardcoded using FHIR-compliant ontologies. The results of the ETL process are JSON files containing the patients' information standardised to the CDM, one file per patient. Those files will be, on one side, the data source on which AI algorithm development relies, and on the other, standardised data available for the scientific community after proper data access request (Figure 4).

Data quality & consistency

Quality data can be defined as data that is fit for purpose, e.g., the data are sufficient for the specified purpose for which it is intended[25,26]. In most cases, data quality for the purpose of machine learning cannot be limited to a single focus but must cater to the needs of multiple audiences. Data quality issues can be introduced at any point in the data management and collection lifecycle. Whether during data acquisition, storage, analysis, or publication, diminished quality can inadvertently affect downstream tasks such as AI training[25–27].

We employ both quality assurance and quality control techniques over the course of the data life cycle including strict conformance to requirements during input and the assessment methods for repeatable feedback and improvement. The overall objective of our quality analysis is to rate the individual records based on multiple dimensions of quality and use this as a filter for downstream tasks. To achieve the measure of data quality needed for superior AI training and results, we defined quality control rules and procedures based on standard dimensions of quality and built tools to integrate into our pipelines for data collection and storage.

Since we use REDCap as the intermediary data store where all collection methods funnel their data, we found REDCap's data quality module[22,23] valuable for organising our data quality rules. This module allows for the execution of quality checks for all data entered into the system, whether by direct entry or by CSV imports. This also enabled the capability to export these rules for use in customised tools for data collection.

Data quality can be evaluated over many different dimensions[26] and we have focused our evaluation on three critical dimensions: completeness, conformance and plausibility. For completeness, we focused on value requirements. For conformance, we analysed the

various data types and permissible values to ensure adherence. Plausibility applied to ranges, such as age. The types and dimensions are outlined in Table 1.

Much of the needed data quality assessment functionality was already built into the REDCap quality module including pre-established rules handling blank values, data type errors, outliers and invalid permissible values. We also included custom rules covering multiple levels of requirements: minimal, mandatory and recommended. These rules aligned with the required fields outlined for each use case and agreed by representatives from each clinical centre.

After assessment of the data based on these organising quality criteria, we generate a score for each quality check based on the number of successes and failures. Here we can also apply weights if we deem an assessment more important than others. Table 2 shows an example of a scoring report.

LEGAL INTEROPERABILITY

EU legal and regulatory landscape: personal and non-personal data, secondary uses and cross-border data sharing for AI research in precision oncology

Huge data processing and sharing needs for the advancement and support of precision oncology have not been ignored by the European legislative and regulatory effort. On the contrary, the last decade has witnessed the flourishing of a plethora of regulations and legal initiatives relevant to AI and data in this domain. Besides the well-known General Data Protection Regulation (GDPR), which is only applicable to the processing of personal data, noteworthy are the legislative developments taking place under the European Data Strategy. Notably, this includes the recently passed Data Governance Act and the proposals for Regulations on the European Health Data Space (EHDS) and on the Artificial Intelligence Act. The latter two are expected to reshape the European regulatory environment and would be applicable to both personal and non-personal data processing.

Personal data vs non-personal data: pseudonymisation, anonymisation and de-identification

Early discussions within research projects and consortia often concern the personal or non-personal nature of the data to be processed. This is mainly due to the fact that processing operations involving personal data, namely information relating to an identified or identifiable natural person, fall under the scope of the GDPR. Therefore, non-personal data such as anonymised data (i.e., personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable) are not bound by the Regulation. Pseudonymised data, which could be attributed to a natural person by the use of additional information, would qualify as personal data (see Article 4(5) and Recital 26 of the GDPR[28]). Generally, de-identification, which is understood as the process of removing or substituting all personal

information and identifiers, is not, per se, sufficient to achieve the anonymisation threshold required in the EU. As a result, if raw data is retained at source or any key or additional information is reasonably likely to be used to reverse the process and re-identify the data subject, information shall be considered as pseudonymised data and thus subject to the GDPR

Within EuCanImage, the removal of direct and indirect personal identifiers has been carried out. In addition, the data is double-hashed both by the data providers and by the platform. Additionally, processing operations carried out within the project are actually aligned with the GDPR and supported by a contractual and governance structure that further enables data sharing.

Secondary use of data for AI research in precision oncology

Further processing of personal data for scientific research purposes, which comprises AI research in precision medicine as pursued by EuCanImage, is compatible with the GDPR (art. 5(1)(b)). Moreover, under the forthcoming EHDS Regulation, secondary use of pseudonymised electronic health data is expressly permitted for "training, testing and evaluating of algorithms, including in medical devices, AI systems and digital health applications, contributing to the public health or social security, or ensuring high levels of quality and safety of health care, of medicinal products or of medical devices" (art. 34(1)(g)). It should be noted, nonetheless, that these two terms ("further processing" and "secondary use") are not legally analogous, but the latter will be preferred here for the sake of clarity.

Cross-border data sharing for AI research in precision oncology

Despite the advent of the GDPR and the harmonisation effort, processing operations involving organisations and researchers from several EU Member states still face slight differences between national, regional or sectoral regulatory frameworks. Hence, although data sharing within Member states of the European Economic Area (EEA) is not hindered by any additional requirements, EuCanImage's partners and their legal teams still must cope with a complex and fragmented scenario. Not only from a legal interoperability perspective, but also concerning divergent ethical oversight layers and internal procedures of each centre, hospital, institution or country.

Transfers of personal data to third countries or international organisations (i.e., data sharing with researchers or organisations located outside the EEA) remain a controversial issue[29,30]. Notwithstanding the fact that EuCanImage members do not plan to store or process data outside the EEA, controversies have arisen in relation to the transfer of data to international organisations and to UK-based institutions after Brexit. Potential routes for transfers remain

limited, particularly, in light of the strict requirements and threshold set by the Court of Justice of the European Union[31,32].

DISCUSSION

In all scientific disciplines, but especially in health research, working with large scale datasets and engaging in cross-border data sharing is becoming increasingly vital for the adoption and development of AI technologies. EuCanImage focuses on leveraging existing healthcare data to address various scientific research questions using AI models. Our experience uncovers obstacles to data interoperability and reuse, as well as realistic solutions. In this report, we outline our procedures to achieve data interoperability, including a thorough description of the data model design, the standards used, harmonisation efforts, and methodological aspects concerning the practical implementation, along with legal interoperability considerations. The successful development and deployment of our data models and related standards represents a significant milestone, laying the groundwork for future AI applications in cancer healthcare. Furthermore, our work highlights the need for improvements in data collection, annotation, and cross-border dissemination.

We anticipate that our approach and methods will not only benefit individual institutions but also serve as a guide for future large-scale consortia requiring harmonisation and interoperability of cancer-related clinical data for AI and machine learning advancements. Similar efforts have been developed by other consortia[33], including projects within the AI4HI initiative such as Chaimilion, ProCancer-I, Incisive or Primarge[34]. These collective endeavours have all involved meticulous, collaborative, expert-driven analysis, spanning model design, data curation, standards usage, and infrastructure development. The knowledge and experience gained from our combined efforts are crucial in laying the groundwork for future healthcare data standardisation initiatives for AI research across Europe.

Achieving interoperability in healthcare data raises complex issues that need to be addressed. The development of supervised AI models trained for prediction or classification tasks relies on data labelled with ‘ground truth’ classifications. Reaching a consensus on data labelling requires common standard definitions for diagnosis and agreements on the level of data granularity; these are critical factors that affect the reproducibility and quality of the results[35,36]. Cancer diagnosis involves the integrating complex criteria based on a variety of disparate data components, such as pathology reports, laboratory results, radiology findings, and advanced molecular and genetic tests. Close collaboration among different medical specialists has enabled the establishment of key principles for data harmonisation: 1) the selection and definition of essential clinical variables to address the medical needs, 2) the identification of common data available across all centres and 3) striking a balance between the volume and granularity of the data that can be provided by various hospitals and the optimal information required for AI models (Figure 3).

Within the healthcare-research ecosystem, data sharing remains a barrier. Yet, it is a crucial mechanism for ensuring that high quality data, obtained through exhaustive and expensive processes like defining data labels, harmonisation tasks, and the use of common standards, can be reused by other researchers and thus, maximise the impact. The FAIR principles– Findability, Accessibility, Interoperability and Reuse– provide the framework for such data re-use[37]. Despite progress in adopting interoperability standards, data from different sources still contain discrepancies. To make data fully reusable and reproducible, methods for data cleaning, harmonisation, and standardisation must be transparent[38].

While the presented work demonstrates the feasibility of using HL7 FHIR to achieve interoperability, it also has limitations. FHIR resources were employed for structural interoperability, while SNOMED, LOINC, NCIT, and RxNorm were mainly used for semantic interoperability. By leveraging the comprehensive information model in FHIR, clinical data can be organised hierarchically in a manner that captures its context and remains unambiguous[39]. However, utilising FHIR for building a model for research oncology presents specific constraints and unique requirements for maintaining data interoperability (described previously - data model section). To maximise the potential of FHIR and encourage broader adoption in the specialised scientific context of AI for precision medicine, alternative FHIR configurations or detailed methodological explanations should be considered to ensure reproducibility. At this moment, the main limitation is the fact that the suitability of the models to develop the AI algorithms have not been validated yet.

In summary, we demonstrated that large-scale, real-world, multicentre clinical data harmonisation and curation for AI research is feasible through the use or adaptation of common standards. The standardised datasets that we will make available, which include data from over 20,000 cancer patients, will provide an invaluable resource for investigators. Expanding the understanding of these complex diseases and opening the door for cutting edge translational research beyond the scope of EuCanImage.

METHODS

Methods used are thoroughly detailed within the results section. Briefly:

Data model

The clinical data necessary to address each use case was established by interdisciplinary teams including clinicians and AI experts considering different key data aspects. Data was arranged following five different data schemas compliant with the FHIR (FHIR Release 4B) architecture. The FHIR Resources used were: Patient, Condition, Observation, Procedure, Medication Administration and Diagnostic report.

Ontologies

HL7/FHIR terminology was used in status profiles required by FHIR. SNOMED (SNOMED version International 2022-12-31) was the preferred terminology for general clinical concepts, ICD-O3 (ICD-O3 version 20220429) for histology, LOINC (LOINC version 2.73) for some test observations and RxNorm (RxNorm version 03-Jan-2023) for medication. We used NCIT (NCIT version 23.8d) when the concept did not exist in previous ontologies.

Data capture, standardisation and quality control

Patient IDs are pseudo-anonymized at the hospitals, and only hashed patient IDs (EuCanImage ID) are introduced in the platform. Data from hospitals is captured in REDCap [REDCap version 13.10.0; PHP 8.1.3 (Linux/Unix OS); MySQL 8.0.30] by filling the online forms or uploading CSV files complying with the specific format requirements. As a result, all clinical data merges into a single harmonized database for each data schema. At this stage we perform quality control checks. We focus our evaluation on three critical dimensions: completeness, conformance and plausibility and we generate a score for each quality check based on the number of successes and failures. We built ETL pipelines in python to transform the harmonized output data into JSON files compliant with FHIR. We use FHIR Validator GUI (https://validator.fhir.org/) and (https://simplifier.net/) as external validators for quality control.

European Genome-Phenome Archive

The European Genome-phenome Archive (EGA) is a service for permanent archiving and sharing of personally identifiable genetic, phenotypic, and clinical data. The standardised clinical data, one JSON file (FHIR compliant) per patient, obtained after the previously described process will be stored encrypted at the EGA repository.

CODE AVAILABILITY

The python scripts to transform harmonised data to the FHIR compatible schemas are available at https://github.com/EGA-archive/EuCanImage-FHIR/

ACKNOWLEDGEMENTS

We acknowledge support of the Spanish Ministry of Science and Innovation through the Centro de Excelencia Severo Ochoa (CEX2020-001049-S, MCIN/AEI /10.13039/501100011033), and the Generalitat de Catalunya through the CERCA programme. We are grateful to the CRG Core Technologies Programme for their support and assistance in this work.

We acknowledge the Social and Legal Sciences Applied to the New Technosciences Research Group, University of the Basque Country (UPV/EHU). Grant from the Department of Education of the Basque Government to support the activities of Research Groups from the Basque University System (Reference IT 1541-22)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952103.

CORRESPONDING AUTHOR

Correspondence to Teresa García-Lezana (teresa.garcia@crg.eu)

AUTHOR CONTRIBUTIONS

J.R., L.F. and T.G.L. developed the concept. F.P., K.L., M.S., and S.C. substantially contributed to the development of the concept. M.B., K.R., S.F., E.N. provided clinical advice. J.R., L.F., T.G.L., M.B., K.R., S.F. and R.F. conceptualised the data model. A.C. and M.Ry. Implemented the model and tests. A.C. and M.Ru. wrote the code. M.Ru., F.P., J.B. and A.C. conceptualised and implemented quality control. M.Re. and P.L. provide legal guidelines. T.G.L., M.B., S.F., M.Ru., M.Re. and A.C. conceptualised and wrote the manuscript. T.G.L. and M.B. created figures. All authors contributed to the critical revision of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests

FIGURE LEGENDS

Figure 1. Data minimisation of clinical (non-imaging) parameters

Figure 2. A) Representation of the proportion of FHIR resources needed for each use-case
B) Ontologies used in each resource

Figure 3. Description of the different steps followed to conceptualise the data model

Figure 4. Clinical data processing workflow from clinical institutions to data analysis/sharing

TABLES

Table 1. Clinical data quality assessment Dimensions

Type	Dimension	Description
Minimal_req	Completeness	Meets minimum requirements
Mandatory_req	Completeness	Meets mandatory requirements
Length	Conformance	Conforms to length restrictions
Datatype	Conformance	Conforms to data type restrictions
Permissible	Conformance	Conforms to list of permissible values
Range	Plausibility	Meets known range limits

Table 2. Data quality scoring report

	Dimension	Type	Fail	Pass	Total	Weight	Score
0	Completeness	minimal_req	2	16	18	50	44.44%
1	Completeness	mandatory_req	3	32	35	10	9.14%
2	Conformance	length	1	10	11	10	9.09%
3	Conformance	datatype	0	40	40	10	10.00%
4	Conformance	permissible	3	52	55	10	9.45%
5	Plausibility	range	1	3	4	10	7.50%
Total:			**10**	**153**	**163**	**100**	**89.63%**

REFERENCES

1. Serra-Burriel, M., Locher, L. & Vokinger, K. N. Development Pipeline and Geographic Representation of Trials for Artificial Intelligence/Machine Learning-Enabled Medical Devices (2010 to 2023). *NEJM AI* (2023) doi:10.1056/AIpc2300038.
2. Haendel, M. A., Chute, C. G. & Robinson, P. N. Classification, Ontology, and Precision Medicine. *N. Engl. J. Med.* **379**, 1452–1462 (2018).
3. Dinov, I. D. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *GigaScience* **5**, s13742-016-0117–6 (2016).
4. He, J. *et al.* The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).
5. Kruse, C. S., Goswamy, R., Raval, Y. J. & Marawi, S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. *JMIR Med. Inform.* **4**, e5359 (2016).
6. Sweeney, S. M. *et al.* Challenges to Using Big Data in Cancer. *Cancer Res.* **83**, 1175–1182 (2023).
7. Frid, S. *et al.* Evaluation of OMOP CDM, i2b2 and ICGC ARGO for supporting data harmonization in a breast cancer use case of a multicentric European AI project. *J. Biomed. Inform.* **147**, 104505 (2023).
8. Näher, A.-F. *et al.* Secondary data for global health digitalisation. *Lancet Digit. Health* **5**, e93–e101 (2023).
9. Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R. & Stiawan, D. The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. *JMIR Med. Inform.* **9**, e21929 (2021).
10. SNOMED International. https://www.snomed.org.
11. International Classification of Diseases (ICD). https://www.who.int/standards/classifications/classification-of-diseases.
12. LOINC. *LOINC* https://loinc.org/.
13. Reich, C. *et al.* OHDSI Standardized Vocabularies-a large-scale centralized reference

ontology for international data harmonization. *J. Am. Med. Inform. Assoc. JAMIA* ocad247 (2024) doi:10.1093/jamia/ocad247.

14. ICGC ARGO. https://www.icgc-argo.org/.
15. FHIR. https://www.hl7.org/fhir/.
16. openEHR International. https://www.openehr.org/about/what_is_openehr.
17. Plazzotta, F., Luna, D. & González Bernaldo de Quirós, F. [Health information systems: integrating clinical data in different scenarios and users]. *Rev. Peru. Med. Exp. Salud Publica* **32**, 343–351 (2015).
18. Garza, M., Del Fiol, G., Tenenbaum, J., Walden, A. & Zozus, M. N. Evaluating common data models for use with a longitudinal community registry. *J. Biomed. Inform.* **64**, 333–341 (2016).
19. Duda, S. N. *et al.* HL7 FHIR-based tools and initiatives to support clinical research: a scoping review. *J. Am. Med. Inform. Assoc. JAMIA* **29**, 1642–1653 (2022).
20. Vorisek, C. N. *et al.* Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. *JMIR Med. Inform.* **10**, e35724 (2022).
21. OMOP Common Data Model. https://ohdsi.github.io/CommonDataModel/index.html.
22. Harris, P. A. *et al.* The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inform.* **95**, 103208 (2019).
23. Harris, P. A. *et al.* Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* **42**, 377–381 (2009).
24. Cheng, A. C. *et al.* REDCap on FHIR: Clinical Data Interoperability Services. *J. Biomed. Inform.* **121**, 103871 (2021).
25. Talburt, J. R. *Entity Resolution and Information Quality*. (Elsevier, 2011).
26. Bian, J. *et al.* Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *J. Am. Med. Inform. Assoc. JAMIA* **27**, 1999–2010 (2020).

27. Priestley, M., O'donnell, F. & Simperl, E. A Survey of Data Quality Requirements That Matter in ML Development Pipelines. *J. Data Inf. Qual.* **15**, 1–39 (2023).

28. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/ EC (General Data Protection Regulation)Quality*. vol. 119(1) (2016).

29. Bernier, A. *et al.* Reconciling the biomedical data commons and the GDPR: three lessons from the EUCAN ELSI collaboratory. *Eur. J. Hum. Genet.* **32**, 69–76 (2024).

30. Molnár-Gábor, F. *et al.* Bridging the European Data Sharing Divide in Genomic Science. *J. Med. Internet Res.* **24**, e37236 (2022).

31. *Court of Justice of the European Union, Judgement of 16 July 2020, Case Schrems II (C-311/18).*

32. *European Data Protection Board (EDPB), Recommendations 01/2020 on Measures That Supplement Transfer Tools to Ensure Compliance with the EU Level of Protection of Personal Data, Adopted on 18 June 2021.*

33. Sweeney, S. M. *et al.* Case Studies for Overcoming Challenges in Using Big Data in Cancer. *Cancer Res.* **83**, 1183–1190 (2023).

34. Kondylakis, H. *et al.* Position of the AI for Health Imaging (AI4HI) network on metadata models for imaging biobanks. *Eur. Radiol. Exp.* **6**, 29 (2022).

35. Wang, S. Y., Pershing, S., Lee, A. Y. & Committee, on behalf of the A. T. on A. and A. M. I. T. Big data requirements for artificial intelligence. *Curr. Opin. Ophthalmol.* **31**, 318 (2020).

36. Cirillo, D., Núñez-Carpintero, I. & Valencia, A. Artificial intelligence in cancer research: learning at different levels of data granularity. *Mol. Oncol.* **15**, 817–829 (2021).

37. Scheffler, M. *et al.* FAIR data enabling new horizons for materials research. *Nature* **604**, 635–642 (2022).

38. Heacock, M. L. *et al.* Enhancing Data Integration, Interoperability, and Reuse to Address Complex and Emerging Environmental Health Problems. *Environ. Sci. Technol.* **56**, 7544–

7552 (2022).

39. Leroux, H., Metke-Jimenez, A. & Lawley, M. J. Towards achieving semantic interoperability of clinical study data with FHIR. *J. Biomed. Semant.* **8**, 41 (2017).

FIGURE 1

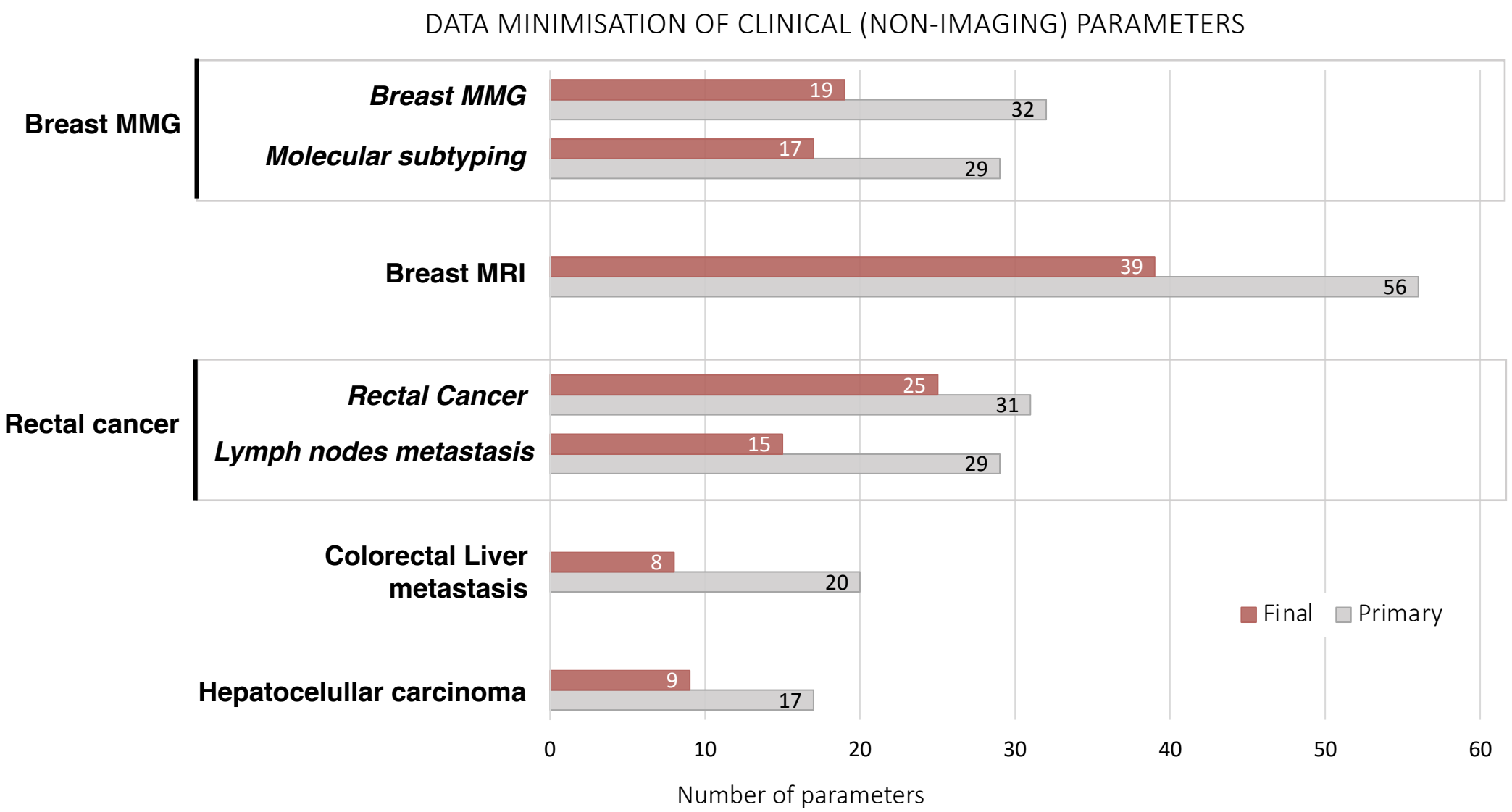
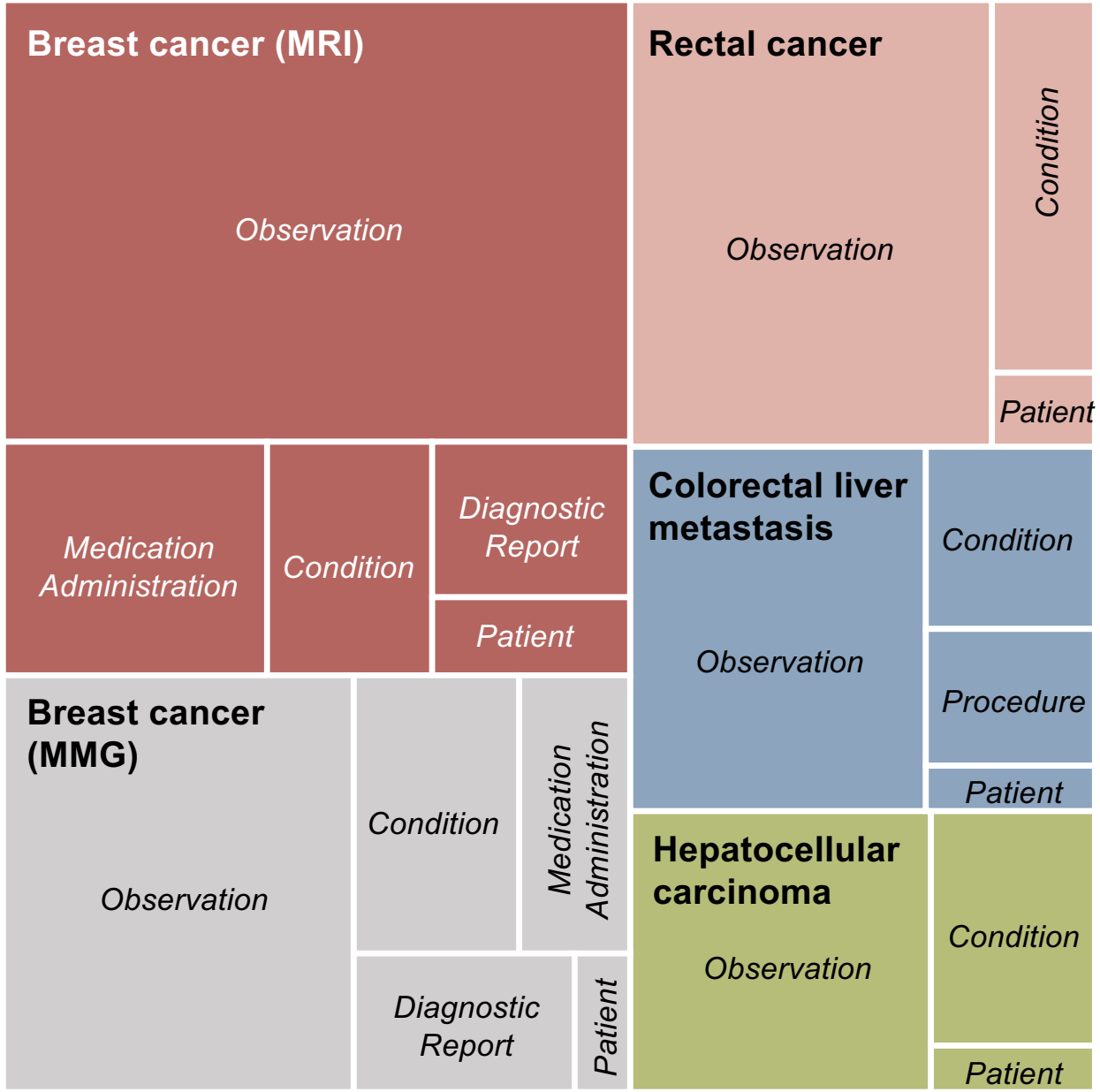

DATA MINIMISATION OF CLINICAL (NON-IMAGING) PARAMETERS
Breast MMG
Breast MMG
Molecular subtyping
19
32
17
29
Breast MRI
39
56
Rectal cancer
Rectal Cancer
Lymph nodes metastasis
25
31
15
29
Colorectal Liver metastasis
8
20
Hepatocelullar carcinoma
9
17
Final
Primary
0
10
20
30
40
50
60
Number of parameters

FIGURE 2

a



b

TERMONOLOGY	
RESOURCE: Patient	
Patient ID	
RESOURCE: Condition	
Condition – code *For histopathology*	SNOMED ICD-O3 (preferred) / SNOMED
Condition - clinical status	HL7/FHIR terminology
Condition – body site	SNOMED
RESOURCE: Observation	
Observation - code	SNOMED (preferred) / LOINC / NCIt
Observation - value	SNOMED (preferred) / LOINC / NCIt
RESOURCE: Procedure	
Procedure - code	SNOMED
Procedure - status	HL7/FHIR terminology
RESOURCE: Medication Administration	
Medication administrator - status	HL7/FHIR terminology
Medication administrator - medication	RxNorm (preferred) / SNOMED
Medication administration - dose	For units: UCUM
RESOURCE: Diagnostic report	
Diagnostic report - code	LOINC (preferred) / SNOMED
Diagnostic report - conclusion code	SNOMED

FIGURE 3

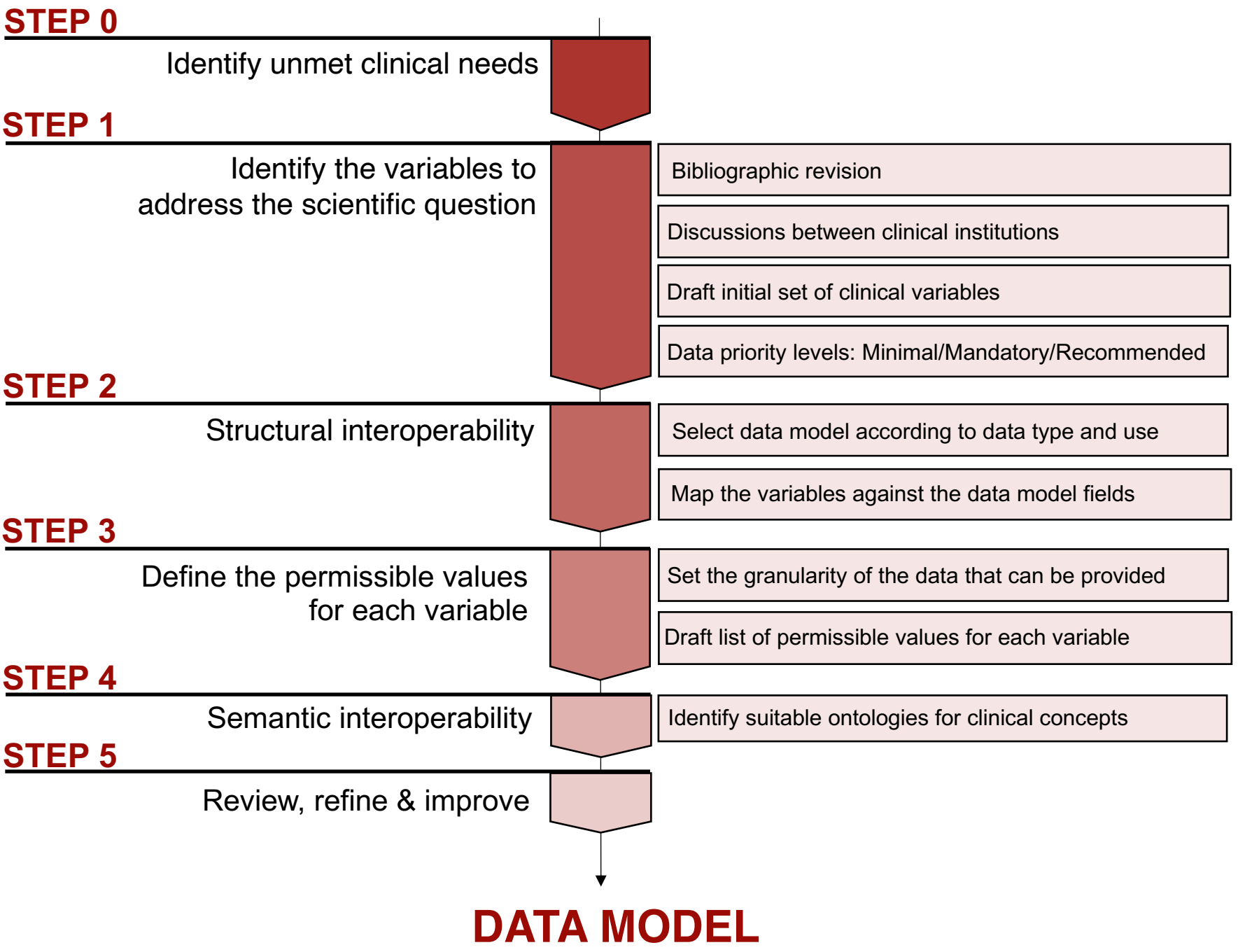
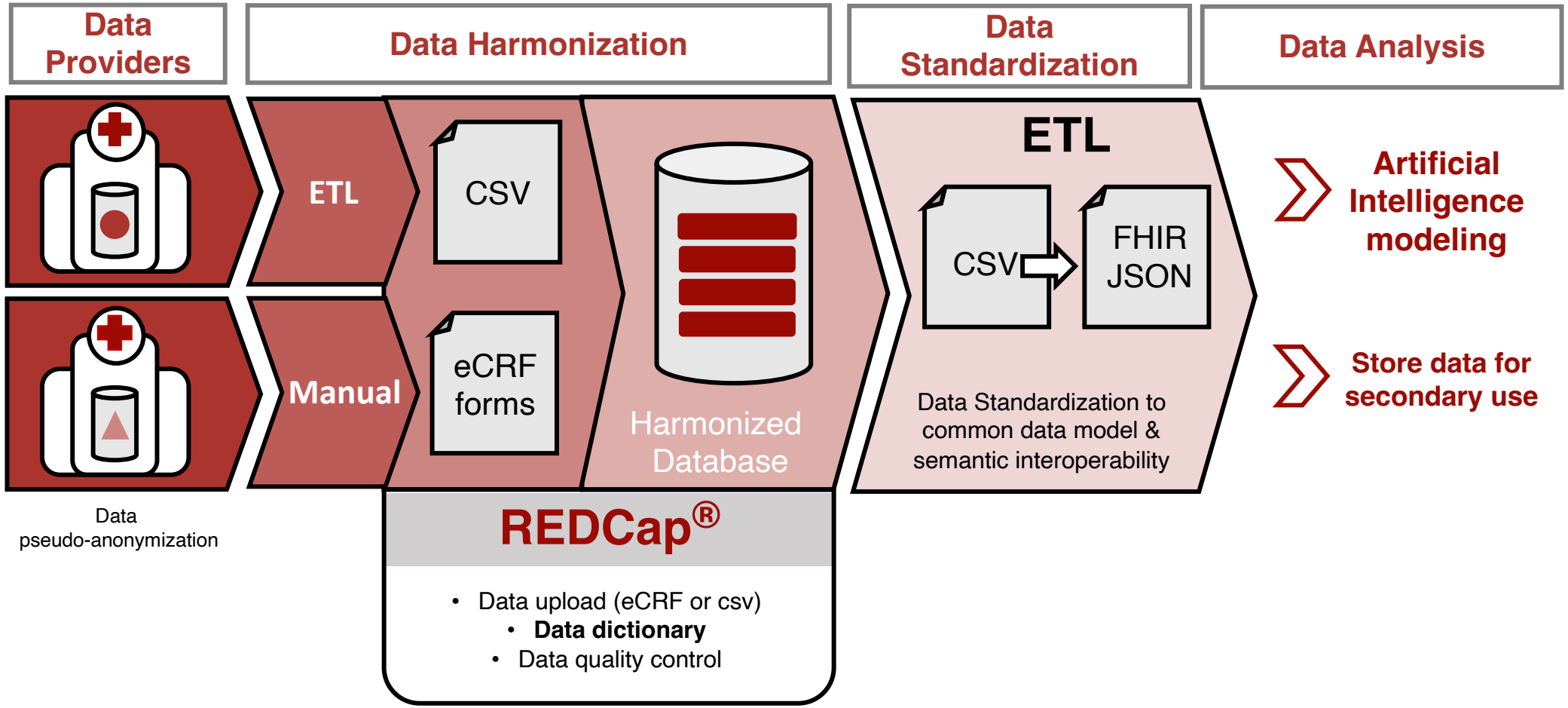

STEP 0
Identify unmet clinical needs
STEP 1
Identify the variables to address the scientific question
Bibliographic revision
Discussions between clinical institutions
Draft initial set of clinical variables
Data priority levels: Minimal/Mandatory/Recommended
STEP 2
Structural interoperability
Select data model according to data type and use
Map the variables against the data model fields
STEP 3
Define the permissible values for each variable
Set the granularity of the data that can be provided
Draft list of permissible values for each variable
STEP 4
Semantic interoperability
Identify suitable ontologies for clinical concepts
STEP 5
Review, refine & improve
DATA MODEL

FIGURE 4


Data Providers
Data Harmonization
Data Standardization
Data Analysis
ETL
Manual
CSV
eCRF forms
Harmonized Database
ETL
CSV
FHIR JSON
Data Standardization to common data model & semantic interoperability
Artificial Intelligence modeling
Store data for secondary use
Data pseudo-anonymization
REDCap®
• Data upload (eCRF or csv)
• Data dictionary
• Data quality control