DIGITAL-2023-CLOUD-AI-04-ICU-DATA

INDICATE

Deliverable D2.1

Minimal dataset description for all clinical use cases

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Cover Page

PROJECT INFORMATION	
Project number:	101167778
Project acronym:	INDICATE
Project name:	A federated INfrastructure for Data of Intensive CAre uniTs in Europe
Project starting date:	01-12-2024
Project duration:	42 months

DOCUMENT INFORMATION	
Deliverable:	D2.1 Minimal dataset description for all clinical use cases
Work Package:	WP2 – Common Data Model
Lead Contributor	UNIVREN
Primary Author(s)	Boris Delange (Université de Rennes)
Dissemination Level[1]	PU
Deliverable Type[2]	DMP
Contractual delivery date	31-07-2025
Actual delivery date	29-07-2025
Document status	Final

TEAM	
Authors:	Boris Delange (Université de Rennes)
Contributors:	Celia Alvarez (SAS) Jan van den Brand (Erasmus MC)
Acknowledgements:	Maria Gonzalez (SAS)
Reviewers:	Falk von Dinklage, Gregor Lichtner (UMG) Janno Schouten (EMC) Christel Daniel (APHP) Elias Grünewald (Charité) Filip Haegdorens (INDICATE EEAB member)

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

HISTORY OF CHANGES			
Version	**Date (DD-MM-YYYY)**	**Author/reviewer**	**Description**
0.1	21-02-2025	Boris Delange Celia Alvarez-Romero	Table of Contents and Initial draft of the deliverable
0.2	29-04-2025	Jan van den Brand	Review for consistency with Data Space Architecture
0.3	07-05-2025	Boris Delange	Description of minimal dataset
0.4	03-07-2025	Boris Delange	Final version ready for review
0.5	23-07-2025	Boris Delange Celia Alvarez-Romero	Reviewers' comments addressed and improvements included after the review process. New version ready for the official submission

[1]PU = Public; SEN = Sensitive, limited under the conditions of the Grant Agreement; CO = Confidential, only for members of the Consortium.
[2]R= Document/Report; DEC = Website; DEM = demonstrator; DATA = federated datasets; DMP – Data Management Plan

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

Table of Content

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

1. Introduction

The INDICATE project (A federated INfrastructure for Data of Intensive CAre uniTs in Europe) aims to establish a pan-European federated data infrastructure for secure cross-border access to Intensive Care Unit (ICU) datasets. This infrastructure must navigate a complex regulatory landscape encompassing data protection, privacy, security, and healthcare-specific regulations.

The INDICATE project is co-funded by the European Commission in the “Digital Europe” programme, in the 2023-2024 call under the topic “Federated European Infrastructure for Intensive Care Units' (ICU) data” and under Grant Agreement no. 101167778.

INDICATE aims to establish a secure federated infrastructure for standardised Intensive Care data. *Federated* means that software used for data or to train artificial intelligence (AI) models is brought to the data instead of transferring the data to the researchers. This way, patient data never leaves the hospital, and unauthorized users never see the actual health data, only the aggregated, non-identifiable results from their analyses. The federated approach is secure and promotes patient privacy by design. Importantly, INDICATE partners are spread across Europe to ensure that the available data reflects the diversity of patients in Europe in the short term, and also in the long term as the network grows.

In this way, INDICATE will form the foundation for training and operationalizing AI models in intensive care units to support applications in personalised medicine, benchmarking between ICUs, and disaster preparedness. For example, INDICATE will facilitate the development of an AI model to predict bacterial bloodstream infection in newborn infants, helping to improve early recognition of this disease.

The Data Management Plan (DMP) of the project, presented in this document, describes the data management life cycle, including specific standards for the databases in terms of formats, metadata, sharing, archiving and preservation. The DMP will also clarify how and which data will be accessible, and outline the procedure to gain such access.

The INDICATE DMP will continue to evolve throughout the project and will be periodically updated as required. This document represents only an initial draft of the description of the data management life cycle for all datasets to be collected, processed or generated by INDICATE.

In INDICATE, data will be generated, collected and managed through the activities conducted within the following work packages:

- WP1 - Project Management and Coordination
- WP2 - Common Data Models and Vocabularies
- WP3 - Governance, Business Models, and Legal and Ethical Framework
- WP4 - Design and Implementation of the Technical Infrastructure
- WP5 - Dissemination, Exploitation and Communication
- WP6 - Clinical Demonstrations

The INDICATE project receives funding from the European Union’s Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

2. Data Summary

Purpose of the data collection/generation: relation to the objectives of the project

The primary goal of INDICATE is to establish an interoperable infrastructure that facilitates secure and widespread access to standardised ICU data from across different European countries. This will allow clinicians and researchers to access large sets of health data without physical data transfers. This approach will promote the exchange of best practices with an expedited response to emerging needs. INDICATE aims to aggregate high-quality data from across Europe, ensuring that its population diversity is accurately represented in terms of nationality, age, sex, ethnicity and other characteristics.

The INDICATE consortium will establish and deploy a pan-European federated infrastructure for ICU datasets. To fulfil this aim, we have the following objectives:

- **Objective 1:** Creation and implementation of data standards via common data models (Work Package 2). We will create a foundation for the standardisation of ICU datasets using common data models (CDMs) and according to Findable, Accessible, Interoperable and Reusable (FAIR) principles. This will enable the querying of and access to high-quality, multi-modal data located on federated nodes and will therefore support research and innovation. This ICU data standardisation protocol will allow secure cross-border federated access and analysis of datasets.
- **Objective 2:** Development of a governance framework for the INDICATE infrastructure to ensure long-term viability (Work Package 3). It addresses critical aspects of governance, including legal and ethical considerations, facilitating data exchange within the infrastructure, and the formulation of a sustainable business model. Its primary aim is to ensure the long-term viability of the infrastructure beyond the conclusion of the project, while adhering to data security and privacy regulations.
- **Objective 3:** Technical implementation of the federated infrastructure for ICU data (Work Package 4). It addresses the technical implementation, optimisation and maintenance of the federated infrastructure to ensure its functionality for all data providers and users.
- **Objective 4:** Dissemination, Exploitation, and Communication for infrastructure establishment and sustainability (Work Package 5). It addresses dissemination, exploitation and communication and will as a priority support the establishment and sustainability of the INDICATE infrastructure. Such activities will include the data user on-boarding process and the provision of digital skills training, educational resources, and documentation. Dissemination and communication will appropriately target audiences and stakeholders with technical, non-technical, clinical, and lay backgrounds.
- **Objective 5:** Completion of real-world testing via six clinical use cases to validate platform functionality and demonstrate value in ICU healthcare (Work Package 6). Six clinical use cases will be used to validate the feasibility, functionality and added value of the INDICATE infrastructure in real-world ICU settings. Use case 1 will focus on building a European Medical Information Mart for Intensive Care (MIMIC-EU) open access database that will serve as an 'atlas' of anonymous acute care cases for research and development of Machine Learning (ML) and AI models. Use case 2 will focus on the early detection of organ failure via ML model development and prediction, while use case 3 will focus on development of a virtual digital twin. Use case 4 will focus on the prediction of neonatal and paediatric sepsis with ML. Use case 5 will focus on quality benchmarking dashboards.

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

Finally, use case 6 will focus on the development of a Grand Rounds workspace: this use case aims to support the discussion of clinical cases through a federated approach.

Types and formats of data generated or collected

INDICATE aims to identify, prioritize, and define a minimal dataset to support clinical use cases in intensive care units (ICUs), particularly in alignment with WP6.

Types of data

The project will involve the generation or reuse of two main types of data sources:

- Electronic Health Records (EHR): routine care data captured during patient management. These are the preferred source, as they enable real-time availability and future scalability of AI-based models in data-rich settings.
- Research databases: may be used as an alternative when EHR data are not available. These databases typically contain data collected under specific study protocols.

The following categories of data will be included, either from EHR systems or research databases (depending on availability and the setup of each data provider):

- Administrative and demographic data – such as patient ID (pseudonymised), age, sex, admission/discharge dates.
- Vital signs and physiological monitoring data – high-frequency time-series data (e.g., heart rate, oxygen saturation).
- Clinical scores, summarizing vital signs and the overall status of a patient
- Laboratory test results – numerical and categorical lab values.
- Imaging metadata – associated with diagnostic imaging (e.g., DICOM headers, not the image files themselves).
- Therapeutic interventions – medication (RxNorm-coded), surgeries, and other clinical procedures.
- Staffing and skill-mix data – including number and types of healthcare professionals per shift, ICU-specific qualifications, and care hours per patient day.
- Outcome variables – such as ICU mortality, length of stay, or readmissions.

Data models

We will use the following data models and code systems to standardize the data:

- OMOP-CDM (Observational Medical Outcomes Partnership – Common Data Model): This is the target format for harmonisation across all sites. All local ICU data will be mapped to this model during the Extract, Transform and Load (ETL) processes.
- HL7 FHIR (Fast Healthcare Interoperability Resources) - Used in two ways:
    - As input format for ETL in centres that already expose data in FHIR.
    - As an interoperability layer (output format) to enable access to OMOP-transformed data by AI tools, dashboards, and clinical applications.
- Additionally, FHIR will serve as a foundation to guide the deployment of AI-based decision support tools in hospitals, through seamless integration with Hospital Information Systems via FHIR-based APIs.

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

Standard terminologies

Standard terminologies will be used to semantically harmonise data:

- SNOMED CT (clinical concepts)
- LOINC (laboratory tests)
- RxNorm (medications)
- ICD-10 (diagnoses)
- DICOM (for imaging metadata)
- HealthDCAT-AP (for health metadata)

All data will remain stored locally at each data provider site (INDICATE federated infrastructure). For federated analyses, only aggregated data will be shared. For federated learning, algorithms will be trained locally at each data provider site, and only model parameters (not data) will be exchanged – following a federated design.

Sources of the data for INDICATE project

No new data will be collected specifically for INDICATE, instead existing data generated during routine clinical care at ICUs of participating Data Providers will be integrated (secondary use). Data Providers are varying in size of their ICU wards, ranging from approximately 10 to 100 ICU beds. However, the data volumes generated through automated monitoring systems at ICUs are large. We expect - even with sampling and reduction techniques - that data providers will be able to provide access to between 10 gigabytes to several terabytes of health data.

Minimal dataset

To support data harmonisation and facilitate the implementation of the use cases, a minimal dataset has been defined. This dataset was developed by collecting and synthesising the data needs expressed by the use case leaders, with the goal of providing a viable first implementation of the INDICATE use cases.

The minimal dataset consists of a comprehensive list of clinical concepts that need to be made available. These concepts are coded using international standard terminologies, including ICD-10, RxNorm, LOINC, SNOMED CT, and UCUM. The dataset is independent of the underlying data model and is designed to support both OMOP and FHIR Extract-Transform-Load (ETL) processes.

The minimal dataset represents a balance between including a limited number of concepts - so that as many data providers as possible can make their data available without overly complex ETL processes - and including a sufficiently rich set of concepts to ensure the scientific relevance and feasibility of the use cases.

However, we anticipate that the minimal dataset may need to be adjusted during the practical implementation of the ETL processes at each Data Provider's site. Despite efforts to select concepts that are likely to be available across all sites, data availability and local specificities may require modifications to the dataset. All modifications and successive versions of the minimal dataset will be tracked and maintained in a version-controlled Git repository (e.g., GitHub, GitLab) to ensure transparency and traceability. The governance and decision-making process for

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

managing changes to the minimal dataset is yet to be fully defined and will be established as part of the project's data governance framework.

Category	Number of concepts	Description / Examples
Demographics and Encounters	14	Demographic information and hospital stay details (e.g., age, gender, birth date, hospital and ICU admission/discharge dates, height, weight).
Conditions	230	Diagnoses and medical conditions (e.g., acute respiratory distress syndrome, sepsis).
Ventilation	52	Mechanical ventilation settings and modes (e.g., FiO_2, PEEP, tidal volume).
Laboratory Measurements	199	Blood and urine test results (e.g., lactate, creatinine, CRP, sodium, leukocytes).
Vital Signs	27	Basic physiological measurements (e.g., heart rate, blood pressure, temperature).
Drugs	11,313	Medications and active ingredients (e.g., norepinephrine, midazolam, antibiotics).
Microbiology	50	Culture results and identified pathogens (e.g., Pseudomonas aeruginosa, blood culture).
Clinical Observations	21	Non-numeric clinical assessments (e.g., Glasgow Coma Scale, RASS score).
Procedures	18	Medical or surgical procedures (e.g., intubation, dialysis, ECMO).

Table 1. Overview of the INDICATE minimal dataset categories.

This table presents the categories of clinical concepts included in the minimal dataset, along with the number of concepts per category and illustrative examples.

For each concept, the dataset provides:

- A direct link to the OHDSI ATHENA vocabulary browser, to support the OMOP ETL mapping process.
- A link to a FHIR terminology API, to assist with FHIR-based transformations.

The dataset is distributed in Excel format to ease human readability, and a CSV version to facilitate technical integration in the ETL process. A web-based application has been developed to help users explore and search through the list of concepts interactively (*Figure 1*).

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

Additionally, each concept is annotated with expert comments to guide selection. For instance, for ventilation-related concepts, comments created through collaboration between clinicians and data scientists highlight the most appropriate choices based on clinical relevance and data availability.

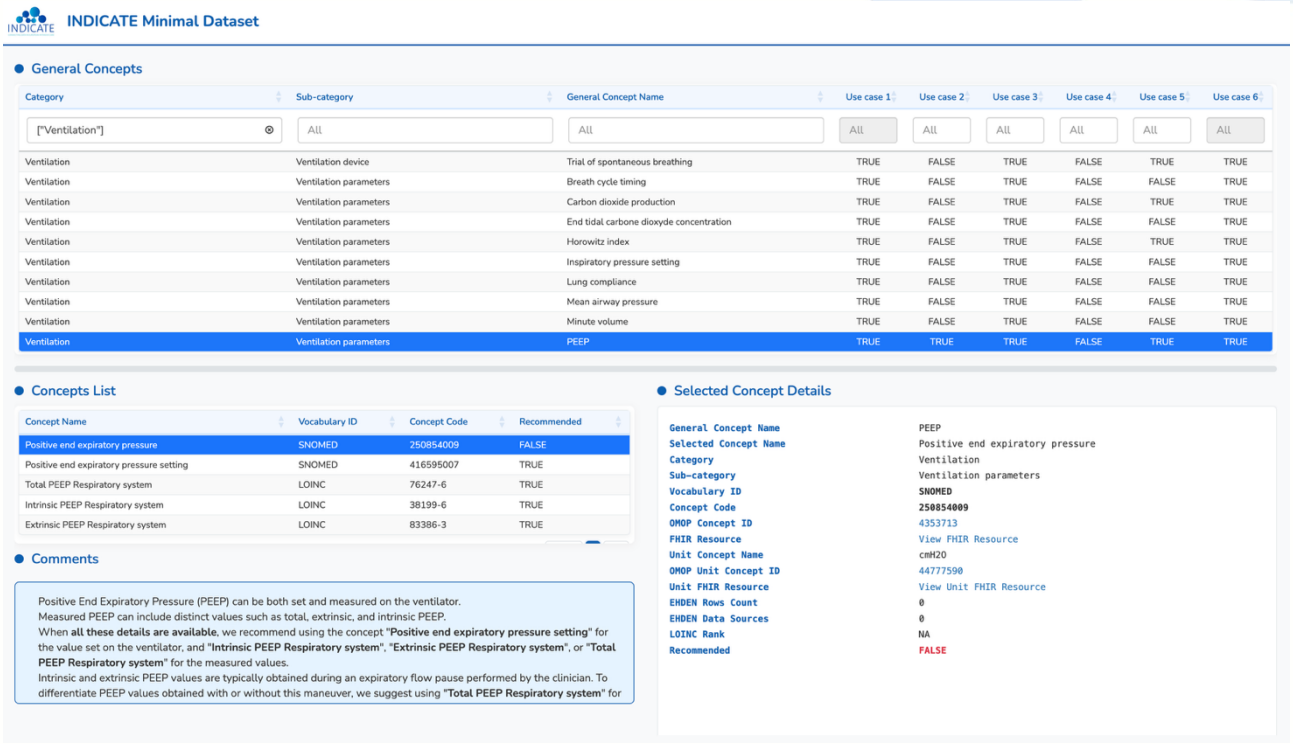

Figure 1. Web interface of the "INDICATE Minimal Dataset" application

This interactive application allows users to explore the clinical concepts identified as necessary for the implementation of INDICATE project use cases. In the top section ("General Concepts"), users can browse the full list of concepts and apply filters by category (e.g., ventilation), general concept name, or by their relevance to each of the six use cases. When a general concept is selected (e.g., PEEP – Positive End Expiratory Pressure), the "Concepts List" section displays all associated standardised terminology concepts from vocabularies such as LOINC and SNOMED, along with an indication of whether the concept is recommended for ETL. Upon selecting a specific concept, the "Selected Concept Details" section displays comprehensive metadata (e.g., terminology, concept code, units, OMOP/FHIR mappings), while the "Comments" section provides detailed guidance to support concept selection. These comments, created collaboratively by clinicians and data scientists, offer recommendations and clarification to help select the most appropriate concept during the ETL process to OMOP or FHIR.

The minimal dataset does not impose any restrictions on the level of data granularity to be provided. The approach is pragmatic: data providers are simply asked to make available all the data they have that correspond to the selected concepts.

In the specific case of high-frequency signals (e.g., electrocardiogram measured from bedside monitors at approximately 300 Hz), a separate task within Work Package 2 is dedicated to this issue. The minimal dataset itself lists the relevant concepts (such as heart rate, systolic blood

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

pressure, etc.) without making assumptions about the data granularity, storage constraints, or data model structure - these aspects are addressed in the scope of that dedicated task.

Data volume estimates will vary depending on the level of granularity of the data made available by each Data Provider. One of the main drivers of variation will be the in- or exclusion of high-frequency data from bedside monitors. These data (e.g., electrocardiogram signals recorded at approximately 300 Hz) can represent several terabytes per site, depending on the number of patients, the recording duration, and the retention strategy.

For example, use cases such as the Digital Twin will rely heavily on time-series ICU data, which may result in significant storage requirements at local nodes. In contrast, other use cases - such as benchmarking dashboards or sepsis detection - may rely on aggregated indicators or smaller patient cohorts, resulting in more modest data volumes.

It is important to note that:

- Medical images are not included, only their metadata (DICOM headers) are handled.
- All data stays local and only aggregated and anonymised outputs are shared.
- Given the federated nature of the INDICATE infrastructure, data will not be centralised but will remain at each Data Provider site.

To support the ongoing harmonisation of data across Data Providers, the project will also develop an Extended Dataset in the next phases. This dataset will build upon the foundations established by the Minimal Dataset and will include additional clinical concepts required for more advanced analyses or for use cases with broader data requirements. The same collaborative and pragmatic methodology - based on input from clinical experts and data scientists, alignment with international terminologies (e.g., LOINC, SNOMED CT, ICD-10), and dual compatibility with OMOP and FHIR - will be applied to ensure continuity and coherence. The Extended Dataset will enable Data Providers to progressively enrich their data contributions, while maintaining interoperability and facilitating the reuse of ETL tools and mappings.

Data availability

Note: The data provided in the table represents an initial estimate based on preliminary input from the INDICATE data providers regarding their data sources and the expected number of patients available per year. However, data providers will need to obtain approval from their respective ethics committees before they can proceed with their role.

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

Data provider	Origin of data	Data Availability Period	Number of Patients per Year
Vall d'Hebron Institute of Research (VHIR – Spain)	EHR	From 2017	Adult ICU: 3,600 patients/year NICU: 0
Ghent University Hospital (UGENT – Belgium)	Research database	From 2013	Adult ICU: 3,000 patients/year NICU : 0
Centre Hospitalier Universitaire de Rennes (CHU Rennes – France)	EHR	From 2021	Adult ICU: 2,600 patients/year NICU: 400 patients/year
Servicio Andaluz de Salud - Virgen del Rocío University Hospital (SAS – Spain)	EHR	From 2026 (prospective study)	Adult ICU: 300 patients/year NICU: 0
St. James's Hospital / Trinity College Dublin (TCD – Ireland)	Research database	From 2014	Adult ICU: 2,000 patients (number of patients per year: NR) NICU : 0
Spitalul Clinic Judeţean de Urgenţă "Pius Brînzeu" (SCJUT – Romania)	EHR	From 2023	Adult ICU: 1,200 patients/year NICU : NR
Region Stockholm - Karolinska University Hospital (RS – Sweden)	EHR	From 2005	Adult ICU: 1,500 patients/year NICU : NR
Heinrich-Heine-Univeristaet Duesseldorf (UDUS – Germany)	EHR	From 2025	Adult ICU: 2,000 patients/year NICU : 0
Charité - Universitaetsmedizin Berlin (Charité - Germany)	EHR	From 2015	Adult ICU: NR NICU: NR
Universitair Medisch Centrum Utrecht (UMCU – Netherlands)	EHR	From 2011	Adult ICU: 2,400 patients/year NICU: NR

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

Medizinische Universitaet Wien (MUW – Austria)	EHR	NR	496,500 patients Adult ICU: NR NICU: NR
Bērnu klīniskā universitātes slimnīca Valsts Sia - Children's Clinical University Hospital (BKUS – Latvia)	EHR	From 2021	Adult ICU: 0 NICU: 400 patients/year
Stichting Amsterdam UMC (AUMC – Netherlands)	EHR	From 2016 for EHR 2003 – 2023 for Amsterdam UMCdb	EHR: 3,300 patients / year A-UMCdb: 3,000 patients / year Adult ICU: NR NICU: NR
Erasmus Universitair Medical Center Rotterdam (Eramus MC – Netherlands)	EHR	From 2017	2,500 ICU patients / year Adult ICU: NR NICU: NR

Table 2. Overview of data availability from participating INDICATE Data Providers.

For each partner, the table summarizes the origin of the data (e.g., EHR or research database), the period of data availability, and the estimated number of ICU patients per year. Some data are still not reported (NR) and will be completed as part of ongoing data collection efforts. EHR: Electronic Health Record; NICU: Neonatal Intensive Care Unit.

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

The majority of INDICATE Data Providers have confirmed that they will provide access to the full scope of ICU data available at their institution, typically covering all ICU stays recorded during the available time period. The earliest datasets go back to 2003 (AUMC – Netherlands), while several partners provide data from 2015 or later. Data availability and patient volumes vary across sites. While some providers report annual estimates (ranging from approximately 1,500 to 3,300 ICU patients per year), others indicate total patient counts over multiple years.

Information regarding high-frequency bedside monitor signals (with up to 500 Hz) is currently available for Charité (Germany) and CHU Rennes (France). For other sites, this information is still pending. It will be collected in detail through a dedicated questionnaire as part of the specific task addressing high-frequency signal data in Work Package 2.

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

3. FAIR data

Making data findable, including provisions for metadata

Datasets that are selected in INDICATE will be assigned persistent identifiers (PIDs) as part of the Study Package concept. The Study Package is the central data artifact that encapsulates all information needed to execute a research study across the federated infrastructure. It serves as a container for study protocols, supporting documentation such as data sharing agreements and ethics approvals, configurations and scripts in both Python and R languages for analysis tasks.

The Study Package is:

- Created by Data Users in their own development environments
- Published to the Study Package repository, following a process similar to version control in software development (e.g., Git local and remote repositories)
- Reviewed and approved by Data Providers
- Executed in Secure Processing Environments
- Produces Aggregated Results

The Study Package contains the following components:

- Protocol: Formal description of the research question, methodology, and expected outcomes
- Ethics Approval: Documentation of research ethics committee approval
- Scripts: Executable code for data processing and analysis (Python, R)
- Environment Configuration: Required libraries, dependencies, and runtime settings
- Data Requirements: Specification of required data elements from the OMOP CDM
- Terms and Conditions: Legal framework for data usage.

Likewise, within the metadata structures of OMOP-CDM and HL7 FHIR, internal identifiers (such as concept IDs, resource IDs, and encounter identifiers) will ensure local data traceability and reusability.

INDICATE project will provide structured and rich metadata to support data discovery, interoperability, and reuse, in line with the FAIR principles.

Types of metadata to be created:

- INDICATE Data Dictionary: A comprehensive dictionary will be developed to describe each minimal dataset element required for clinical use cases. This data dictionary serves as a core metadata resource, as it documents the structure, semantics, and priority of each data element, thus enabling data discovery and semantic interoperability. It will include:
    - Field name
    - Description
    - Unit of measurement
    - Data type
    - Linked standard terminology (e.g., SNOMED CT, LOINC, RxNorm, ICD-10)
    - Priority level (mandatory, recommended, optional).

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

    - Semantic mappings: Each dataset item will be mapped to international standard vocabularies to ensure semantic alignment.
- ETL process metadata: Metadata on data provenance, transformation date, applied CDM version, and quality control metrics will be recorded.

Standards to be followed:

- HealthDCAT-AP: HealthDCAT-AP is a health-related extension of the DCAT application profile for sharing information about Catalogues containing Datasets and Data Services descriptions in Europe (DCAT-AP). DCAT-AP is maintained by the SEMIC action, Interoperable Europe.
- OMOP-CDM: Defines structured metadata fields such as concept_id, source_value, vocabulary_id, and domain_id to standardise health data representation.
- HL7 FHIR: Includes embedded metadata for clinical context, coding, status, timing, authorship, etc., especially for interoperability purposes.
- DICOM metadata: Only DICOM headers will be managed to capture technical identifiers and clinical metadata from imaging systems.
- Standard terminologies: SNOMED CT, LOINC, RxNorm and ICD-10 will be used for semantic consistency.

If metadata standards do not exist for specific data types, modelling decisions will be documented in the INDICATE Data Provider Handbook to ensure coherence and traceability across the consortium.

INDICATE pays special attention to the public metadata catalogue, which describes its datasets by following metadata standards, such as the DCAT-AP for Data Portals in Europe. INDICATE project will generate structured metadata aligned with international standard (HealthDCAT-AP), which in principle supports metadata harvesting and indexing.

Making (meta)data accessible

Repository

The INDICATE project will not deposit personal health data into a centralised trusted repository, due to the federated nature of the infrastructure and the sensitivity of ICU patient-level data. All personal data of patients will remain under the control of the data providers and will be stored locally at each institution. Data will be accessed and queried through federated tools and infrastructures, following GDPR-compliant anonymisation and governance procedures.

However, other types of research outputs — such as ETL scripts or metadata, may be considered for deposition in trusted repositories, depending on their type, purpose, and licensing model. Trusted repositories, in this context, are understood as platforms that provide clear terms of deposit, transparent governance, and well-defined preservation policies that enhance trust and long-term accessibility. To that end INDICATE will deploy a Metadata Catalog Service that enables discovery and understanding of available datasets without accessing the data itself.

The Metadata Catalog Service:

- Collects and standardises metadata from data providers
- Implements the HealthDCAT-AP standard
- Provides search and browsing capabilities

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

- Links to terms and conditions for data use

The Metadata Catalog Service provides the following capabilities:

- Metadata Publication: Tools for data providers to publish standardised metadata
- Search and Discovery: Advanced search capabilities for finding relevant datasets
- Quality Assessment: Validation of metadata completeness and conformance
- Metadata Federation: Integration with other metadata catalogs

The INDICATE central metadata catalog - whether as a catalog of catalogs or a central store - will also be accessible to other data spaces, provided these external spaces use a DCAT-AP interoperable metadata catalog, thus ensuring cross-domain metadata interoperability.

INDICATE recognises the importance of persistent identifiers in supporting data discovery and reuse, and this requirement will be considered when selecting trusted repositories for research outputs other than health data. Further guidance will be provided in subsequent deliverables and technical documentation associated with metadata and repository integration.

Data

Data generated or reused within the INDICATE project will not be made openly available. Health data from intensive care units (ICUs) are highly sensitive and fall under strict data protection regulations, including the General Data Protection Regulation (GDPR).

Legal and contractual reasons:

- Health data remain under the full control of each data provider and are hosted locally, in accordance with the federated architecture of the INDICATE infrastructure.
- Data sharing is subject to national and institutional data governance policies, as well as patient privacy and informed consent limitations.
- The Grant Agreement allows beneficiaries to restrict access when opening the data would violate applicable laws or contractual obligations.

Intentional restrictions:

- Raw health data are not exposed outside the institutions of origin. Access to anonymised or harmonised data is foreseen only through federated infrastructure and under strict governance rules defined by WP3.

Other research outputs (e.g. ETL tools, metadata schemas, documentation) may be shared openly when feasible and appropriate, in accordance with FAIR principles.

The Data Provider Handbook and subsequent deliverables will further specify the procedures for controlled data access, including technical, legal, and ethical safeguards.

No embargo periods have been explicitly defined in the project documentation. However, if specific research outputs are considered for intellectual property protection (e.g. patents) or prior publication, embargoes may be applied in accordance with the conditions of the Grant Agreement. In such cases, embargo periods will be kept as short as possible and justified on a case-by-case basis.

Health data in the INDICATE project will not be made available through a public access protocol

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

due to privacy constraints and the federated nature of the infrastructure. Access to data will be mediated by governed workflows and federated tools.

For other research outputs such as ETL tools or metadata schemas, standardised protocols (e.g. HTTPS, APIs based on HL7 FHIR) may be used to enable controlled and documented access. These outputs will be made accessible in compliance with FAIR principles, but only when no legal or ethical restrictions apply.

For the metadata publication, public repositories will be analysed and made accessible in accordance with FAIR principles.

Due to the federated architecture and sensitivity of ICU health data, access will be subject to strict legal and ethical restrictions both during and after the project.

During the project:

- Access to data will occur only through federated analysis workflows.
- No patient-level data will leave local data providers.
- All access will be subject to the governance and anonymisation procedures defined in INDICATE.

After the project ends:

- Access to data will continue to rely on the federated infrastructure.
- Long-term governance of access mechanisms will depend on sustainability plans defined in the project.
- No direct data downloads or central publication of sensitive health data is foreseen.

Access to other types of research outputs (e.g. tools, schemas, synthetic datasets) will follow the project's open science and FAIR data strategy, if no legal or contractual restrictions apply.

Details on governance, anonymisation, and access request procedures will be developed in the Data Provider Handbook and other INDICATE deliverables, as part of the ongoing implementation of access governance within the project.

The identity of individuals accessing the data will be managed through strict authentication and authorisation mechanisms integrated into the federated infrastructure. These mechanisms will be developed by WP3 as part of the implementation of the access governance framework.

Access to sensitive health data will not be possible, because INDICATE's approach is to establish a secure federated infrastructure for standardised Intensive Care data, and in compliance with governance rules established in collaboration with data providers, ensuring traceability and compliance with GDPR and institutional policies.

A Data Access Committee will be created to review applications for access from third parties. This committee will take into account all relevant circumstances and data protection regulations. INDICATE will design and implement a comprehensive governance framework, defining roles and procedures for federated analysis. This will include criteria for approval, audit trails, and alignment with GDPR and institutional policies.

Metadata

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

Metadata generated in INDICATE will follow FAIR principles and will be made available openly whenever possible.

Metadata related to the platform's technical and functional operations, such as format specifications, data provenance, ETL documentation, and access logs, may be openly shared and potentially released under public domain dedication (e.g. CC0), depending on the nature of the information and the absence of sensitive content, including trade-secrets.

Metadata describing health datasets will be openly licensable under CC0 because they will not contain indirect identifiers or sensitive contextual information, including trade-secrets.

Metadata will include references and documentation to enable discovery and access procedures, including identifiers, data source descriptions, and contact information or access request mechanisms where applicable.

The duration for which data will remain available and findable within INDICATE depends on the sustainability model to be developed under WP5 and the governance framework defined in the project.

Health data will remain hosted locally by data providers and accessible for the federated analysis as long as the nodes and governance framework are maintained. Long-term accessibility beyond the project duration is subject to future institutional and infrastructure commitments.

Metadata will be managed and stored in such a way as to ensure continued availability and findability, even if the corresponding datasets are no longer accessible. Metadata describing datasets, tools, and processes will be retained in FAIR-compliant registries or repositories in INDICATE infrastructure.

Where applicable, open-source software used for ETL processes, data validation, or federated queries will be made available, including source code, dependencies, and instructions for deployment. This will support transparency, reusability, and reproducibility of the processes applied to the data.

Software developed within the project will follow open science practices, subject to licensing and intellectual property considerations. When possible, it will be deposited in trusted repositories and referenced in the corresponding metadata.

These practices are inspired by The Book of OHDSI, which illustrates how open-source tools and public documentation foster community-driven collaboration, reproducibility, and transparency in health data standardisation and analytics.

This documentation will enable data users to understand the tools involved and reproduce the transformation pipeline applied to datasets.

Making data interoperable

To enable interoperability and data exchange within and across disciplines, INDICATE will adopt community-endorsed standards, vocabularies, and methodologies for both data and metadata.

The project will primarily use the OMOP Common Data Model (CDM) from the Observational Health Data Sciences and Informatics (OHDSI) community for harmonising health data across ICU sites. OMOP CDM enables consistent representation of health data and supports

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

standardised analytics.

In parallel, HL7 FHIR will be adopted for semantic interoperability and for enabling API-based access to health data elements. This dual approach supports both structured data analysis and integration with clinical decision support systems and electronic health records.

For data, the project will apply internationally recognised vocabularies such as SNOMED CT, LOINC, ICD-10, and RxNorm to ensure semantic consistency and support reuse.

For metadata, FAIR principles will guide the use of standardised vocabularies and ontologies. The project will apply internationally recognised standard such as DCAT-AP.

These practices reflect established best practices in health data standardisation and are aligned with methodologies described in The Book of OHDSI and the implementation strategy of WP2.

If uncommon or project-specific ontologies or vocabularies are required, the INDICATE project will provide mappings to established, widely used standards to maintain semantic and technical interoperability.

Any new ontologies or controlled vocabularies generated within the project will be properly documented and openly published to facilitate their reuse, refinement, or extension by other researchers or initiatives. When relevant, INDICATE may also propose the inclusion of missing concepts into existing standard vocabularies (such as SNOMED CT) by submitting them to the appropriate maintaining organisations.

INDICATE will incorporate qualified references to other datasets where applicable. These may include:

- References between datasets generated within different use cases of the project.
- Links to external data sources or research datasets used for validation, comparison, or enrichment.

These references will follow recognised metadata standards and FAIR principles to ensure they are machine-readable, persistent, and semantically clear. The inclusion of qualified references supports data discoverability, reuse, and contextual understanding.

Increase data re-use (through clarifying licences)

INDICATE will provide detailed documentation to support data validation, transparency of analytical processes, and data reuse.

This will include:

- README files for each dataset or data package.
- Methodological descriptions of data collection, transformation, and cleaning procedures.
- Codebooks defining variables, units of measurement, and value domains.
- Descriptions of analytical workflows and data pipelines.

Licences for data use and documentation will be publicly shared and versioned on the INDICATE project website to ensure long-term traceability and reproducibility of the research processes applied to health datasets.

Data generated or reused in INDICATE include health datasets, synthetic or derived data,

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

metadata, and software tools. Not all of these data types will be made freely available in the public domain. Health data, in particular, is subject to legal and ethical restrictions due to their sensitive nature.

Instead, access will be managed under a federated infrastructure model, with governance and compliance mechanisms in place. Data will remain under the control of each provider and accessible only under approved conditions.

When appropriate and legally permissible, non-sensitive derived outputs will be shared under standard open licences to promote reuse (e.g. EUPL, CC-BY or CC0).

Metadata will be made openly available under a CC0 licence when possible.

Where legally and ethically permitted, certain types of data produced in INDICATE—such as metadata, and software tools—will be made available for reuse by third parties, including after the project ends.

Health data will remain under the control of the original data providers. Access to these data will require approval through governance mechanisms established in the project.

For reusable outputs, licensing terms will ensure that metadata and documentation are available in formats and under conditions that support their continued discoverability and reusability.

The provenance of data in INDICATE will be documented using appropriate standards to ensure transparency, traceability, and reproducibility. This includes the use of structured Data Contracts, following the latest version of the Data Contract Specification, which describe the full lifecycle of the data, from source systems to transformation processes and quality rules.

Data provenance information will include:

- Source systems and data origin.
- Transformation steps, including ETL processes.
- Versioning information and update history.
- Data custodianship and responsible entities.

These elements will be integrated into metadata records, following FAIR principles and community-endorsed standards where applicable.

INDICATE will implement a comprehensive data quality assurance (QA) framework to ensure that all data used or generated meet acceptable levels of accuracy, completeness, consistency, and reliability.

Relevant QA processes include:

- Definition of a data quality framework aligned with FAIR principles and based on OHDSI community best practices.
- Implementation of validation rules and automated checks during the ETL process to detect anomalies and data integrity issues.
- Use of standardised vocabularies (e.g. SNOMED CT, LOINC) to ensure semantic consistency across datasets.
- Monitoring and auditing tools to evaluate the quality of transformed datasets.
- Documentation of data quality metrics and issue resolution logs.

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

The data quality strategy and procedures will be described in detail in the Data Provider Handbook (WP2).

In addition to data, INDICATE will generate other research outputs such as software tools, workflows, technical documentation, and training materials. These outputs will be considered in the DMP, and their management will align with FAIR principles to the extent possible.

Resource Allocation: Responsibilities for data and research output management are distributed across WP2, WP3, WP4, WP5, and WP6. WP2 will define data provider onboarding, transformation, and quality processes. WP3, WP4 and WP5 will provide the technical infrastructure and governance mechanisms to ensure sustainability and reuse. WP6 will carry out the INDICATE use cases. The allocation of human and technical resources will be monitored within the project's management framework.

Data Security: All data processing activities will comply with GDPR and relevant national regulations. The federated architecture adopted by INDICATE avoids centralising sensitive health data. Instead, data remains at the provider site and only authorised, privacy-preserving federated queries are performed. WP4 will develop and enforce security protocols for authentication, authorisation, and access control.

Ethical Aspects: The project handles sensitive health data and operates under strict ethical and legal oversight. Each data provider is responsible for ensuring local ethical approvals are in place. No patient-level data will be transferred centrally. WP1 and WP3 oversees the compliance with ethical requirements as described in the Grant Agreement.

The management of these aspects will be documented and updated in coordination with the project's evolving needs and reported in the respective work package deliverables.

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

4. Other research outputs

In addition to health data, the INDICATE project will generate and manage several types of digital research outputs, which include:

- Software tools and scripts: Developed for ETL processes, data quality assessment, and transformation between FHIR and OMOP.
- Workflows: Standardised ETL workflows that are reproducible and documented in the Data Provider Handbook.
- Protocols: Methodological frameworks for data transformation, de-identification, and semantic harmonisation.
- Information models: Extensions to the OMOP-CDM for ICU time-series data and mappings to HL7 FHIR resources.

These outputs will be managed according to the same FAIR principles applied to data, particularly:

- Findable: Documentation will be included in project deliverables (e.g. D2.1, D2.3, D2.4) and deposited in open repositories when applicable.
- Accessible: Outputs will be made available to partners and, where applicable, to the broader research community under appropriate licensing.
- Interoperable: Code and models will follow community-adopted standards (e.g. OHDSI conventions, HL7 guidelines).
- Reusable: Metadata and licensing terms will be clearly documented to ensure that outputs can be reused and adapted by other initiatives.

Further planning for long-term preservation and sustainability of these outputs will be aligned with WP5 and the Knowledge Base platform architecture.

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

5. Allocation of resources

INDICATE project is co-funded by the European Commission under Grant Agreement no. 101167778. INDICATE acknowledges the importance of allocating resources to support the application of FAIR Principles and other research outputs. However, the precise costs remain under discussion and will require coordination across the consortium and with WP3, responsible for business model of the data space, in particular.

The following cost categories are currently being considered:

- Direct costs: data collection, data curation, transformation to OMOP CDM and FHIR, metadata creation, documentation, and quality assurance processes.
- Indirect costs: infrastructure development and maintenance, secure storage systems, long-term archiving, and technical support for federated access.

Both project-specific and general costs (e.g., for shared infrastructure and storage services) will need to be defined in collaboration with the INDICATE coordination team. These will be detailed in project-level planning and budget reporting as the technical implementation progresses.

The support of WP1, WP2, and WP4 is essential to provide accurate estimates and align them with sustainability and compliance strategies defined in the Grant Agreement.

Costs related to the FAIRification of data and other research outputs in INDICATE - such as data curation, metadata creation, documentation, infrastructure, storage, and governance - will be covered during the co-funding project, provided they meet the eligibility criteria and conditions set out in the Grant Agreement.

These include:

- Direct and indirect costs incurred for ensuring data quality, accessibility, and reusability.
- Expenditures related to technical infrastructure, security, long-term access, and sustainability measures.

Final allocations and reporting of these costs will be defined during the project's financial planning in coordination with WP1.

In INDICATE, data management responsibilities are distributed in alignment with the federated architecture and privacy-preserving principles of the project:

- Each data provider is responsible for managing the health data hosted at their site, as well as the data governance and privacy. This includes ensuring data quality, compliance with local ethical and legal requirements, implementation of data transformations, and participation in federated analysis.
- WP2 oversees the onboarding of data providers and defines the technical and methodological frameworks for data transformation and quality assurance.
- WP3, WP4 and WP6 participants will define and enforce governance rules for data access, security protocols, and coordination of federated analysis queries, while respecting data privacy and non-centralisation of patient-level data.
- Overall data governance is coordinated with support from WP3 to ensure consistency with ethical, legal, and project-wide policies defined in the Grant Agreement (WP1).

This distributed model ensures that data remain under the control of the providers, while enabling

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

collaborative analysis through secure and standardised processes.

In INDICATE, long-term data preservation will be addressed in alignment with the federated infrastructure and the principles of data privacy, security, and interoperability. The Data Provider Handbook (D2.2), to be developed under WP2, will define requirements for data lifecycle management, including guidelines for retention, archiving, and eventual disposal.

Each data provider remains responsible for the preservation of their local datasets. Decisions about what data to preserve and for how long will consider the scientific value, legal obligations (e.g., GDPR), and institutional policies. These decisions will be made in coordination with WP3 (governance and access) and WP1 (overall project coordination and compliance with the Grant Agreement).

The project will promote best practices to ensure data remain accessible and findable for as long as they are useful for research and policy purposes. When appropriate, documentation and metadata will be preserved to support reproducibility and reuse, even after the data itself is no longer available.

Resources required for preservation (including secure storage infrastructure, technical support, and legal oversight) will be discussed and coordinated with WP1, WP3, and WP4, particularly for identifying shared infrastructure needs and estimating associated costs. The potential value of retained data will guide prioritization efforts, with a focus on supporting future research in intensive care and machine learning model development.

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

6. Data security

Data security in INDICATE will be managed by WP3 and implemented through a federated infrastructure to ensure sensitive health data remain under the control of the data providers. The following provisions will be in place:

- Secure storage: Each data provider is responsible for implementing secure storage solutions compliant with national regulations and GDPR.
- Federated architecture: No raw patient-level data will be transferred to a central repository. All analyses will be performed through federated queries on local instances.
- Secure transfer protocols: WP4 will define and implement secure communication channels for transferring limited data or derived results.
- Access control: WP4 will define authentication and authorisation protocols to ensure access is restricted to authorised personnel.
- Data recovery: WP4 will provide guidance on institutional-level backup and disaster recovery plans to be maintained by each data provider.

These security mechanisms will be specified and coordinated through WP4 and documented in the Data Provider Handbook and relevant technical deliverables.

Data security in INDICATE is guided by the federated architecture of the project, which avoids centralising sensitive health data and ensures that it remains under the control of the data providers to ensure the data privacy.

Key provisions include:

- Federated infrastructure: Health data is stored and processed locally at each institution. There is no central repository of patient-level data, reducing the risk of unauthorised access or breach.
- Secure results transfer: All communications and results transfer during the federated analysis and AI model training will use encrypted channels. WP4 will define secure APIs and federated query protocols that prevent data leakage.
- Access control for federated analysis: WP3 and WP4 will define role-based access controls, authentication protocols, and logging mechanisms to ensure that only authorised users can access specific data elements.
- Storage and archiving: Health data will be stored in secure, access-controlled environments, in compliance with GDPR and institutional security standards for each clinical setting. Long-term archiving solutions will be aligned with institutional and national guidelines.

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

7. Ethics

The INDICATE project aims to provide cross border federated data access to ICU data for secondary use. The data of individual patients does not leave the control of the hospital by design, so that we will comply with data security, privacy, ethical principles, and relevant regulations and legislation. The data management policies will be defined in WP3. This includes the rules and principles for data access in accordance with EU regulations and Directive (1982/2006/2014/2016 EC), General Data Protection Regulation 2016/679 (GDPR) and the ePrivacy Directive (2002/58/EC), and respective national legislation of participating countries. Within WP3, INDICATE will also monitor ethical aspects of research with sensitive personal data, including the Declaration of Helsinki, International ethical guidelines for health-related research involving humans of the Council for International Organizations of Medical Sciences (CIOMS) and the WHO, and the Council of Europe Convention for the Protection for Human Rights and Dignity of the Human Being with regards to the Application of Biology and Medicine.

The INDICATE project, aimed at implementing a federated infrastructure for ICU datasets, places a strong emphasis on ethical considerations and data governance. To achieve this, we are implementing data standardisation using Common Data Models, developing adequate governance, business models, and legal and ethical framework for the sustainable operation of the infrastructure, are developing a sustainable governance framework, deploying technical infrastructure, and engaging in widespread dissemination and validation through clinical demonstration projects. A critical aspect of this is ensuring the security and privacy of health data: during clinical demonstrations, the data remains under the control of the providing hospital.

In preparation for research activities, ethical permits will be obtained, and all data processing will be conducted in strict accordance with European regulations and national legislation. To oversee these aspects, INDICATE has established an Ethics Advisory Board. This board is tasked with monitoring ethics issues in the project, focusing on data representativeness, diversity, explainability of AI, and emerging ethical issues related to AI in medical research.

The Ethics Advisory Board aligns with the European Ethics Guidelines for Trustworthy AI and the World Health Organization (WHO) guidance on Ethics and Governance of Artificial Intelligence for Health. It also keeps track of legislative developments such as the AI Act. The establishment of the Ethics Advisory Board ensures that INDICATE not only commits to technological innovation but also upholds ethical excellence in AI and data analytics in healthcare. This structure ensures that the project's advancements are ethically responsible, align with international standards, and foster trust. The Ethics Advisory Board has regular interaction with the steering committee.

Ethical issues related to data privacy, informed consent, and compliance with GDPR are detailed in the WP3 deliverables.

Ethics and legal considerations are critical to the INDICATE project, especially as it involves the health data processing (anonymised in all case). The project has designed a federated infrastructure that ensures health data remains within the control and governance of data providers, minimizing privacy risks. In any case, ethical and legal requirements will be analysed in INDICATE WP3 and will be included during the developments in WP4 and implementation during WP6 use cases:

- Informed Consent for prospective use of data will be included in the process for

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

questionnaires dealing with personal data. The informed consent process is designed to ensure that patients are fully aware of the use of their data and that their rights are respected, particularly in the context of GDPR compliance.

- Consent for Data Processing: Each data provider is responsible for ensuring that the use of patient data in the project complies with applicable legal requirements. This consent will cover the federated data analysis for research purposes and the long-term preservation of the data in line with the project's federated infrastructure approach.
- Federated Infrastructure: The federated infrastructure ensures that data remains under the control of the provider, meaning that data is not centralised. However, for analysis, aggregated data results will be shared without disclosing any patient-level data, thus ensuring that consent is respected in the data processing procedures.
- Ethical Approval: Each data provider is responsible for ensuring that the use of patient data in the project complies with applicable legal requirements.
- Data Privacy: Ensuring patient confidentiality and adhering to data protection laws (GDPR) is essential in the project. This is particularly important as the data will not be physically transferred to a central repository but analysed through federated queries.
- Ethical Oversight: The project will follow ethical guidelines established by the Ethics Advisory Board of the project, which will monitor the handling of health data and ensure compliance with relevant laws.
- Legal Jurisdiction: Data sharing across borders may encounter legal barriers depending on local data protection laws, which must be considered in the setup of the federated system.
- Long term preservation: In terms of the measures that will be in place to effectively plan for the exploitation of the Project results should be addressed by the tasks defined at WP3. It implies:
    - To define and test different business models that could be used. The final goal will be to select and fully define the business model for INDICATE by the end of the project.
    - To design and fully develop the long-term sustainability plan for INDICATE, including selection of the suitable legal model and definition of the operational model as a base for the future activities (after the project).

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union

8. Other issues

The INDICATE project will comply with national and institutional procedures for health data management in all participating countries. Given the sensitive nature of ICU health data and the federated architecture of the project, each Data Provider is responsible for ensuring compliance with applicable national regulations, sectorial guidelines, and internal institutional policies regarding data processing, access control, retention, and ethical oversight.

These procedures will be fully aligned with the principles and requirements of the General Data Protection Regulation, and will be integrated into the project's governance and technical implementation.

The INDICATE project receives funding from the European Union's Digital Europe Programme under grant agreement number 101167778

Co-funded by the European Union