IMPROVE

Framework to IMPROVE the Integration of Patient Generated Health Data to Facilitate Value Based Healthcare

D4.1: Additional data collection: needs and methodology

** This deliverable is titled "**Additional data collection: needs and methodology**". In the DoA, deliverable D4.1 is incorrectly referred to as "**Use case Definition and Key Performance Indicators (KPIs)**", which duplicates the title of deliverable D5.2 ("Use case detailed study definition and Key Performance Indicators (KPIs)"). This is an editorial error in the DoA. The present deliverable has been prepared in line with its correct scope and objectives, corresponding to additional data collection needs and methodology. The correction of the deliverable title will be formally requested in the next amendment to the Grant Agreement, to be submitted after the 2nd periodic report.*

Version 2.0

Authors:
Frans Folkvord (PBY)
Nicolò Ferriani (PBY)
Davide Guerri (Dedalus)
Laura Pinna (Dedalus)

Co-funded by the European Union

Document Control Sheet

Deliverable Number	D4.1
Deliverable Responsible	PBY
Work Package	WP4
Lead Editor	Frans Folkvord
Internal Reviewer(s)	Ernst Hermens (PMS), Hans Peeters (PMSN) and Giuseppe Fico (UPM)

History of Changes

Date	Version/Page	Change
24.10.2025	0.1	ToC of the deliverable
10.11.2025	0.2	First draft of the deliverable
21.01.2026	0.2	Review feedback provided
24.01.2026	1	Final draft of the deliverable
26.01.2026	2	Improved version of the deliverable

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise.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

Legal Disclaimer

The information in this document is provided “as is” and as it has been collected according to the inputs provided by the different partners. The above referenced consortium members shall have no liability to third parties for damage of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law. This data management plan is a living document and will evolve with the advancement of the project.

Abbreviations and Acronyms

D	Deliverable
EC	European Commission
KPI	Key Performance Indicator
PGHD	Patient Generated Health Data
PPI	Patient Preference Information
PREMs	Patient-Reported Experience Measures
PROMs	Patient-Reported Outcome Measures
T	Task
M	Month
VBHC	Value-Based Health Care
WP	Work Package
RWD	Real-World Data
GDPR	General Data Protection Regulation
EHDS	European Health Data Space
EHR	Electronic Health Records
DGA	Data Governance Act
HTA	Health Technology Assessment
CDE	Common Data Element
FHIR	Fast Healthcare Interoperability Resources
OMOP	Observational Medical Outcomes Partnership
FAIR	Findable, Accessible, Interoperable and Reusable
DPO	Data Protection Officers
PoC	Proof of Concept
DTA	Data Transfer Agreement
PwMS	People with Multiple Sclerosis
NKI	Netherlands Cancer Institute
ETL	Extraction, Transformation and Loading
bDFS	Biochemical Disease-free Survival
dMFS	Distant Metastasis-free Survival
OS	Overall Survival
HRQoL	Health-related Quality of Life
SNOMED CT	Systematised Nomenclature of Medicine – Clinical Terms
LOINC	Logical Observation Identifiers Names and Codes
AI	Artificial Intelligence
LLM	Large Language Model
QA	Quality Assurance
SpecFr	Specification Framework
ML	Machine Learning

Table of Content

List of Figures

List of Tables

Executive Summary

The WP4 “Collection and Analyses of Historical Data” aims to define, design and implement the necessary methodological, organisational and technological solutions for additional Real World Data (RWD) collection towards the refinement of the IMPROVE Framework. The work of WP4 is strictly interlaced with the activities of WP2 (Evidence-based identification, monitoring and assessment system), WP3 (IMPROVE Living lab) and WP5 (Use Cases for Validation). In the first year of the project, WP4 worked in close collaboration with WP3 for identifying the technological infrastructure and the methodological approach and allowing different data sources, in particular retrospective clinical data and prospective clinical data, to contribute to building the IMPROVE framework data repositories. Taking into account the clinical domains and objectives of the use cases to be piloted in the WP5, starting from the beginning of the second year of the project, the WP4 analysed the preliminary results of the systematic umbrella reviews, the Screenathon-based updates and the associated data extraction processes that now populate the Knowledge Warehouse, performed in WP2. This analysis provided insights that guided the conceptualisation of the data collection requirements for historical clinical datasets. Taking into account these interactions with WP2, WP3 and WP5, the WP4 journey includes the following phases:

1. Identification of functional and non-functional requirements for collecting historical data and sharing them with the IMPROVE Framework, design of the appropriate technological infrastructure and identification of the methodological approach for collecting data.
2. Analysis of pre-existing historical data related to the use case of WP5, by leveraging existing RWD repositories to minimize redundancy and maximize value from already available data sources
3. Mapping and definition of additional data needs, based also on the preliminary results and insights of the WP2 systematic umbrella reviews, screenathon and data extraction.
4. Identification of possible retrospective studies based on requirements defined above and, design of protocols for data collection, ensuring consistency, interoperability, and compliance with ethical and legal standards.
5. Identification of Proof on Concepts for testing the technical specifications and infrastructure, the methodological approach and evaluate the needed organizational aspects.
6. Starting of systematic data collection and gap analysis, for supporting framework testing and iterative improvement of the model system developed in WP3 through targeted data enrichment.

The deliverable D4.1 “Additional data collection: needs and methodology” reports the results related to the first four phases of the WP4 journey. In particular, the deliverable describes the background and the rationale of the work: by well explaining the relationship between WP4 and WP2, WP3 and WP5, by focusing on the relevance of the data gap analysis during the WP4 journey and by underlining the importance of collecting the RWD in the IMPROVE framework. Then the deliverable focuses on the methodological approach and the implementation plan adopted by WP4 for performing the phases 1-3 above, taking also into account that the IMPROVE project has to operate on a robust, representative, and high-quality data foundation, enabling the generation of actionable, generalizable insights across clinical and operational contexts. Finally, the deliverable describes the results regarding phases 3 and 4 by identifying the possible data sources and how they can contribute to building the IMPROVE framework. These activities provided results, insights, and demonstrations for finalizing and formalizing the retrospective studies and for starting with the systematic collection of historical data.

1. Background and Rationale

The work conducted within WP4 builds upon the foundational analyses and design activities carried out in the earlier phases of the project, with the objective of defining the requirements and specifications necessary to enable the collection and integration of historical clinical data within the IMPROVE framework. The rationale behind this deliverable is therefore to document the methodological pathway through which these requirements were derived, ensuring coherence with the overall project architecture and alignment with the needs of the subsequent clinical studies.

The starting point for this work was the state of the art established through WP2. WP2 conducted the systematic umbrella reviews, the Screenathon-based updates, and the associated data extraction processes that now populate the Knowledge Warehouse. This work has delivered a cross-disease synthesis of evidence on Patient-Generated Health Data (PGHD) in oncology, ophthalmology, cardiovascular disease, neurology, and chronic inflammatory conditions. The resulting database includes thousands of individual studies and hundreds of systematic reviews and meta-analyses. Thanks to the comprehensive literature review performed in WP2, the consortium obtained an in-depth understanding of the landscape of PGHD across the various topic areas addressed by the clinical studies in WP5. This literature analysis enabled the identification of relevant data types, common data gaps, existing standards, and methodological approaches used in comparable initiatives. These insights constituted the evidence base that guided the conceptualisation of the data collection requirements for historical clinical datasets.

In parallel, close collaboration with WP3 allowed the definition of the overall architecture of the IMPROVE framework. Through this joint effort, we mapped how different data sources, ranging from literature-derived evidence to datasets from previous projects, policy-related information, retrospective clinical data, and prospective clinical data, could contribute their information in a harmonised and interoperable manner. WP3 provided the architectural perspective necessary to understand how each data stream should be ingested, transformed, and made usable within the framework, thereby establishing the technical boundaries and expectations for WP4.

Within this context, WP4 focused on translating the conceptual and architectural foundations into concrete specifications for the acquisition and integration of historical clinical data. This required conducting a detailed analysis of data modelling approaches, harmonisation strategies, standard terminologies, and interoperability requirements. Special attention was given to terminology normalisation, data structure definition, metadata requirements, and the conditions necessary to ensure consistency between retrospective and prospective datasets. The resulting specifications aim to guarantee that historical clinical data can be systematically collected, standardised, and made available to the IMPROVE framework without loss of meaning or usability.

Overall, the rationale for this deliverable is to demonstrate how the interplay between literature evidence (WP2), architectural design (WP3), and technical analysis (WP4) has enabled the definition of robust and actionable requirements for historical clinical data collection. These specifications form a critical building block for the functioning of the IMPROVE framework and ensure that downstream activities, particularly those in WP5, can rely on a coherent, interoperable, and well-structured data foundation.

The iterative interaction among the findings of WP2, WP4 and WP5 allows, in fact, WP3 to build and consolidate an integrated conceptual model for PGHD within the IMPROVE framework, for mapping the definitional and structural facets of PGHD across the literature and for consolidating insights from the Science, Policy, and Practice Trackers and from stakeholder consultations to identify which PGHD domains are most relevant and actionable for Value-Based Healthcare (VBHC) and Health Technology Assessment (HTA).

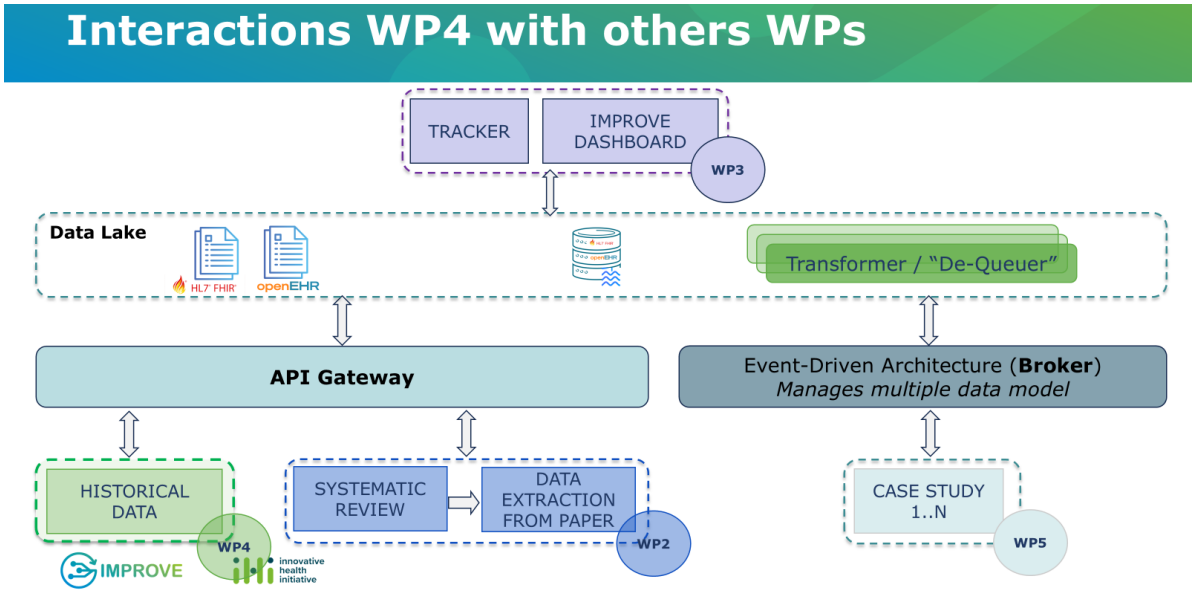


Figure 1: Interactions WP4 with other WPs

From the WP4 perspective, the interaction between these outputs is not merely background information; they constitute an iterative work for defining additional data needs and planning new data collection**.** In order to describe the work done for defining the needs and the methodology for performing the additional data collection, it is important to underline the journey of the WP4 activities, the need to cover the data gaps, related then to the strategic role of RWD in the IMPROVE framework.

1.1 The WP4 journey and data gaps

For a better understanding of the activities performed in the WP4 during the first two years of the project and the results reported in this deliverable, this section summarizes the journey of the WP4 activities according to the structure of the Document of Action.

WP4 includes four tasks:

- The activities of WP4 started with the task "T4.2. Development of Specifications Framework for the data collection technologies tailoring" at M3, with the objective of this task is to design a Specification Framework (SpecFr) including the functional, non- functional, legal and regulatory and business requirements of the data collection technologies. This task, as described above, has been carried out by strictly collaborating with WP3.

- At M10 started the task “T4.3. Data collection and (gap) analyses historical data”. In particular, the first year of this task has been very important for the definition and the realization of Proof of Concepts (POCs). Again, this was in strict collaboration with WP3 for defining and iteratively assessing both the technological infrastructure of the IMPROVE project and the methodology for collecting retrospective and then prospective data.
- At M12 started the task “T4.1. Additional data needs definition and design for new data collection and existing RWD studies”. With this task, started also the interaction with WP2, as described above.
- In parallel with the previous tasks there is the task “T4.4. Complementary methods”, the recurrent and very proactive action of the IMPROVE project able to continuously introduce insights in the framework.

To undertake this journey and address the challenges of WP4, according to T4.2, the first activities focused on the retrospective standardisation of historical datasets by mapping previously collected data onto a common data model, harmonising variables, and converting different scales into consistent formats.

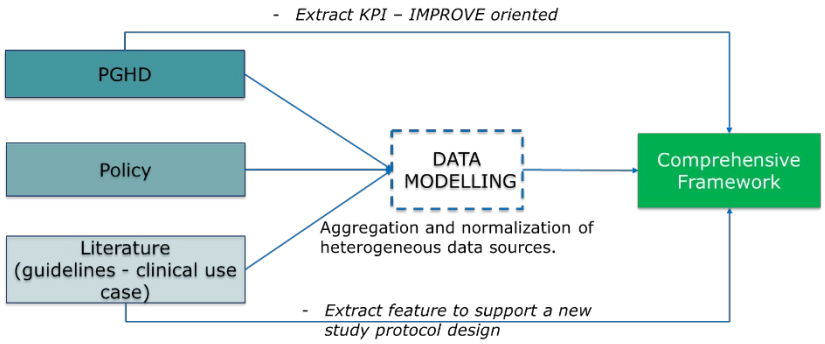


Figure 2: Extracted KPIs that are IMPROVE oriented

In parallel, WP4 promoted semantic integration through the alignment of data elements with shared vocabularies, ontologies and coding systems, ensuring consistent meaning and interpretation across datasets and sources. This ensures that historical PGHD can be integrated, interpreted, and reused across studies, reducing inconsistencies due to heterogeneous data collection practices. Additionally, WP4 emphasised the integration of multiple data sources including electronic health records, clinical notes, patient-reported outcomes, and, where available, wearable or app-derived data, to reconstruct a more complete patient trajectory. By combining these complementary sources, the work package aims to fill gaps in temporal coverage and enhance the quality and richness of both historical and prospective datasets, providing a robust foundation for subsequent longitudinal analyses and predictive modelling.

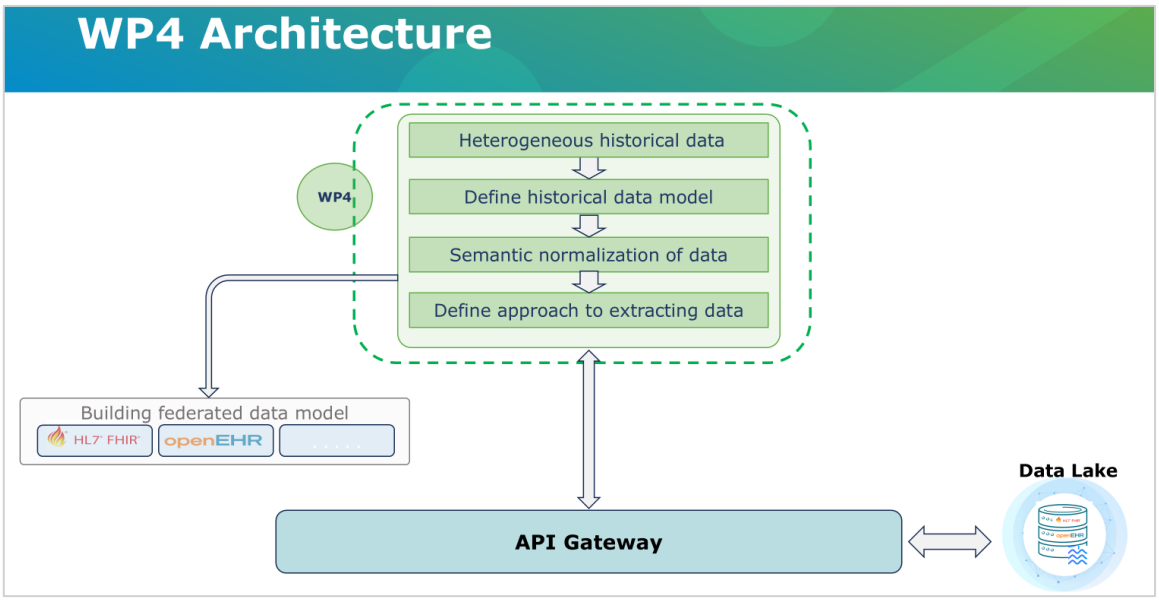


Figure 3: WP4 Architecture

The activities focused on some of the IMPROVE use case areas, such as multiple sclerosis, breast cancer, ophthalmology and heart failure. And according to T4.3 as well as by defining the POC#1 about the use case multiple sclerosis, the WP4 started also with the preliminary data collection activities. Other retrospective studies have been started regarding breast cancer, ophthalmology and prostate cancer.

After the first year of the project, the preliminary evidence of the retrospective studies, together with the setting of the needed knowledge base in the WP3 framework and the preliminary results of the systematic work in WP2 started to reveal a series of important data gaps that directly guide the planning of additional data collection and integration activities in WP4, by activating in these case the actions in the task T4.1. First, there are disease-specific gaps. Several of the IMPROVE use case areas, such as breast cancer, heart failure, and multiple sclerosis, are supported by a substantial body of PGHD-related evidence, especially in the form of PROMs. However, even in these relatively well-covered fields, PREMs, PPIs, and longitudinal, pre-treatment data are often missing. In contrast, other target diseases, atrial fibrillation, aortic stenosis, macular degeneration, and chronic rhinosinusitis, show sparse or highly fragmented PGHD evidence. For WP4, this means that new or supplementary data collection must be targeted to those underrepresented conditions and to the underexplored PGHD dimensions within the better-documented diseases.

Second, there are geographical and population gaps. Most of the evidence synthesised in WP2 originates from high-income Western countries. Data from Eastern Europe, Africa, Latin America and several parts of Asia are scarce or absent. This limits the generalisability of existing findings and the ability to build models that are robust across different health systems, socio-economic contexts, and digital maturity levels. WP4 therefore needs to prioritise RWD and PGHD collection from regions and populations that are underrepresented in the literature, leveraging the diversity of the IMPROVE consortium partners and their regional ecosystems.

Third, the analysis has revealed temporal and structural gaps in PGHD. Many studies collect data at single time points, usually post-treatment, with limited coverage of pre-treatment baselines, intra-treatment dynamics, or long-term follow-up. Continuous or high-frequency PGHD, for instance from wearables or app-based monitoring, are still the exception rather than the norm. At the same time, there is a lack of harmonised instruments and common data elements: different studies use distinct questionnaires, scales, and formats, with inconsistent metadata and variable reporting quality. For WP4, this translates into a need to design longitudinal and standardised data collection protocols, ensuring that future datasets cover the full patient journey and are recorded in a way that supports comparability and reuse.

Finally, there are integration and governance gaps. PGHD are often collected in isolation from clinical RWD sources such as EHRs, registries, imaging or administrative data, and are rarely linked to contextual policy or practice information. In addition, legal and ethical considerations, especially around secondary use and cross-border data sharing, have often been addressed only superficially in the literature. These gaps indicate that WP4 must not only acquire additional data, but also address linkage, interoperability, and governance from the outset, aligning with GDPR, the EHDS, the Data Act, and the DGA. Taken together, these findings define the data needs for WP4: more representative, longitudinal, standardised and integrated RWD and PGHD, with explicit attention to disease, geography, stakeholder perspective, and regulatory context. The subsequent task of WP4 is to translate these needs into concrete data collection strategies, partner contributions, and technical specifications.

1.2 Importance of RWD in the IMPROVE Framework

RWD plays a central role in the IMPROVE framework, both as the substrate for evidence generation and as the bridge between research, practice, and policy. While WP2 and WP3 have focused primarily on synthesising and conceptualising the evidence base, WP4 is where RWD becomes operationalised as the fuel for model development, framework testing, and use case implementation. Within IMPROVE, RWD is understood broadly, encompassing not only traditional sources such as EHRs, registries, and claims data, but also PGHD collected through apps, wearables, sensors, online platforms, and patient-reported instruments. The integration of these sources is essential to move beyond isolated datasets and to construct a holistic, longitudinal view of the patient journey, covering clinical events, daily life, experiences, preferences, and interactions with the health system. This integrated RWD layer enables the project to quantify value in a way that is consistent with VBHC: linking outcomes that matter to patients with the resources and pathways used to achieve them.

RWD is also indispensable for developing, calibrating, and validating the AI/ML components of the IMPROVE framework. Models that are trained only on trial data or narrow cohorts have the risk being brittle, biased, and poorly generalisable. The use of heterogeneous RWD from multiple regions, disease areas, and care settings ensures that the framework reflects real heterogeneity in patient characteristics, behaviours, and health system performance. It also allows the project to test how PGHD can effectively complement existing clinical data, for example by improving risk stratification, treatment selection, monitoring, or shared decision-making.

Furthermore, RWD underpins the project's ambition to be aligned with and useful for regulatory and policy contexts, including VBHC strategies, HTA processes, and the emerging EHDS infrastructure. High-quality, well-governed RWD is a prerequisite for generating credible, reproducible evidence that can inform reimbursement decisions, guideline development, and service design at local, national, and European level. By designing WP4 explicitly around interoperable, legally compliant, and ethically sound RWD flows, IMPROVE positions itself as a practical demonstration of how PGHD and RWD can be mobilised within the new European data governance landscape.

In summary, RWD is not an accessory element in IMPROVE but a core structural component of the framework. It connects the insights from WP2 and WP3 to the real-world implementation activities of WP4 and WP5, and it enables the project to operationalise its conceptual model in everyday practice. The work in WP4 will therefore focus on making RWD available, reliable, standardised, and linkable, so that the IMPROVE framework can deliver on its promise of person-centred, data-driven, and value-based healthcare.

2. Methodology and Implementation Plan

2.1 Data Needs Definition and Data Collection Design

The data needs for WP4 are defined through a structured, evidence-based process that builds directly on the activities of WP2 and WP3 and are guided by the use cases of WP5. These work packages are iteratively providing a detailed map of the current PGHD and RWD landscape, identified gaps in the scientific evidence base, and established a conceptual framework that specifies the essential variables, domains, and metadata required to support the IMPROVE model. WP4 translates these findings into a concrete data collection design that ensures the availability, quality, and representativeness of the data needed to support framework development, model testing, and use case implementation.

2.1.1 Defining Data Needs Based on Evidence and Conceptual Priorities

The first step in WP4 is the systematic definition of data needs, grounded in three primary inputs:

1. **Scientific Evidence Gaps (WP2):** The umbrella reviews and Screenathon updates revealed significant gaps in PGHD coverage across diseases, geographies, measurement domains, and phases of care. WP4 therefore prioritises data elements that are underrepresented in the literature, including pre-treatment PGHD, longitudinal monitoring data, PREMs and PPIs, and data from Eastern European, Southern European, and non-EU populations represented by consortium partners.
2. **Conceptual Model Outputs (WP3):** WP3 produced an integrated PGHD model specifying key constructs (e.g., actors, data types, contexts, integration pathways), which informs the types of variables required for model development. These include patient-reported outcomes, patient experiences, behavioural and environmental data, temporal data sequences, and linkable clinical indicators. The model also highlights governance and metadata requirements, such as provenance, device characteristics, and contextual descriptors.
3. **Use Case Requirements (WP5):** The 10 IMPROVE use cases outline specific analytical and clinical needs that further refine the data requirements. Each use case specifies the relevant patient journey stages, disease-specific indicators, and VBHC outcomes. WP4 ensures that the data collected are sufficient to support these targeted implementations.

Based on these inputs, WP4 defines a structured set of Common Data Elements (CDEs) and associated metadata categories that guide data ingestion from consortium partners and external sources.

2.1.2 Design Principles for Data Collection

The design of data collection activities in WP4 is guided by several principles to ensure robustness, scalability, and compliance:

- **Patient-Centredness:** Data collection strategies recognise the central role of PGHD in capturing daily-life experiences, treatment preferences, and functional outcomes that reflect patient priorities.

- **Longitudinality:** Recognising the gaps in temporal data identified in WP2, WP4 prioritises longitudinal PGHD and RWD that span pre-, peri-, and post-treatment phases.
- **Interoperability:** Data structures are designed to align with FHIR, OMOP, HL7, FAIR principles, and the emerging EHDS interoperability standards, enabling future integration with HealthData@EU and national health data access bodies.
- **Modularity and Flexibility:** Data pipelines accommodate different formats and modalities, including PROMs/PREMs, wearables, app-generated metrics, registry data, EHR extracts, and contextual socio-demographic data.
- **Governance by Design:** Data collection integrates GDPR, Data Act, DGA, and EHDS compliance from the outset, including the definition of lawful bases, consent models, pseudonymisation strategies, and data sharing agreements.

2.1.3 Partner Data Contributions and Data Flow Design

The data collection design incorporates structured contributions from multiple consortium partners (e.g., IBIDELL, Dedalus, MEDS, AReSS, MME, UDUS, IER, PMSN, NKI). Each partner provides datasets based on their institutional capacity, disease focus, and data access rights. WP4 documents:

- dataset availability,
- data structure and formats,
- data quality,
- population characteristics,
- temporal coverage,
- linkage potential with PGHD,
- legal and ethical constraints, and
- update frequency.

These datasets are integrated through harmonised pipelines into the Knowledge Warehouse, where automated and semi-automated validation is performed before they will be made available for WP3 modelling and WP5 implementation.

Data Flow

The data flow designed for the collection, processing, and integration of historical clinical data within the IMPROVE framework ensures full compliance with privacy requirements, interoperability principles, and project-wide analytical needs. The process consists of several coordinated steps, implemented consistently across all contributing partners.

Local Data Extraction and Privacy Compliance

Each partner begins by extracting the retrospective clinical data from their local systems in full compliance with GDPR, institutional privacy policies, and the consent conditions under which the data was originally collected. Only data that is legally and ethically authorised for reuse is carried forward into the subsequent stages of the process.

Identification of Relevant Parameters Based on KPIs

Following extraction, partners identify the specific parameters that are relevant for the project. This selection is driven by the KPIs defined both at the use-case level and at the overall IMPROVE project level, ensuring that the collected data aligns with the evaluation needs and analytical priorities shared across the consortium.

Data Modelling According to the Shared Data Model

Once the relevant variables have been identified, the data is modelled according to the shared IMPROVE data model. This modelling process includes structural transformation, terminology harmonisation, and the enrichment of metadata to ensure interoperability. During this step, partners also assign standardised tags that characterise and classify the data, facilitating efficient indexing and subsequent processing within the framework.

Secure Data Transfer to the Project Data Lake

After modelling, the harmonised datasets are transferred to the project Data Lake through an API-based mechanism. Authentication is handled via Keycloak, ensuring that data transfer is secure, traceable, and accessible only to authorised entities within the project infrastructure.

Framework Ingestion and Validation

Once received by the infrastructure, the IMPROVE framework performs an ingestion process that includes structural and semantic validation, metadata alignment, and classification based on the previously assigned tags. Through this ingestion step, the datasets become part of the project's internal repositories and are prepared for analytical processing.

Data Availability for Dashboard-Based Analytics

After successful ingestion, the data becomes available to the analytical layer of the IMPROVE framework. Here, it is processed according to the project's KPIs and analytical workflows, and the resulting insights are visualised through the IMPROVE dashboard. This ensures that standardised, high-quality retrospective data supports monitoring, evaluation, and decision-making across the various use cases.

2.1.4 Integration of PGHD and Traditional RWD

A key methodological feature of WP4 is the combined integration of PGHD and traditional RWD. Data collection design ensures that:

- PGHD from m-health, e-health, and digital tools can be linked (with safeguards) to clinical indicators and outcomes;
- structured and unstructured data can be transformed into harmonised, machine-readable formats;

- contextual information-policy measures, clinical practices, care pathway characteristics-can be tied to patient-level data for enriched analyses.

This integrated design supports the development of robust, multi-modal datasets that reflect the complexity of real-world patient experiences.

2.1.5 Iterative Refinement and Stakeholder Inputs

Data needs and collection plans are not static. WP4 employs an iterative refinement process in which:

- feedback from WP3 model testing
- insights from WP5 use case piloting
- stakeholder consultations with patients, clinicians, and regulators

lead to continuous adjustment of CDEs, data standards, and collection strategies. This ensures that WP4 remains responsive to practical realities, evolving regulatory requirements, and emergent opportunities.

By systematically defining data needs and designing a harmonised, patient-centred, legally compliant data collection strategy, WP4 ensures that the IMPROVE framework is built on a solid empirical foundation. This approach provides the high-quality, representative, and interoperable RWD/PGHD required to support model development, framework testing, and real-world implementation across Europe.

2.2 Overview of Data Collection Strategy

The overarching objective of the project is to collect and integrate a wide spectrum of healthcare data to support evidence-based decision-making by policymakers, clinicians, and other healthcare stakeholders. This includes PGHD, PREMs, PROMs, and PPI, as well as data derived from comprehensive literature reviews, and economic datasets. Such multi-source data integration is essential to enable a holistic understanding of patient health, treatment effectiveness, and healthcare service quality. Within this context, the leader of WP4 is specifically responsible for the collection, integration, and management of historical data. Historical data are defined as datasets collected retrospectively, often covering extended time periods, and provide critical baseline and longitudinal insights. To evaluate the availability, nature, and quality of historical data across project partners, a comprehensive survey was conducted involving five key contributors: AReSS Puglia, FISM (Italian Multiple Sclerosis Federation), CatSalut, UDUS, and VHIR. The survey aimed to identify the types of data held, the clinical domains covered, data sharing capabilities, and technical formats. Survey results revealed an equal split between partners offering historical and real-time data sources, with no partners indicating an inability to share data.

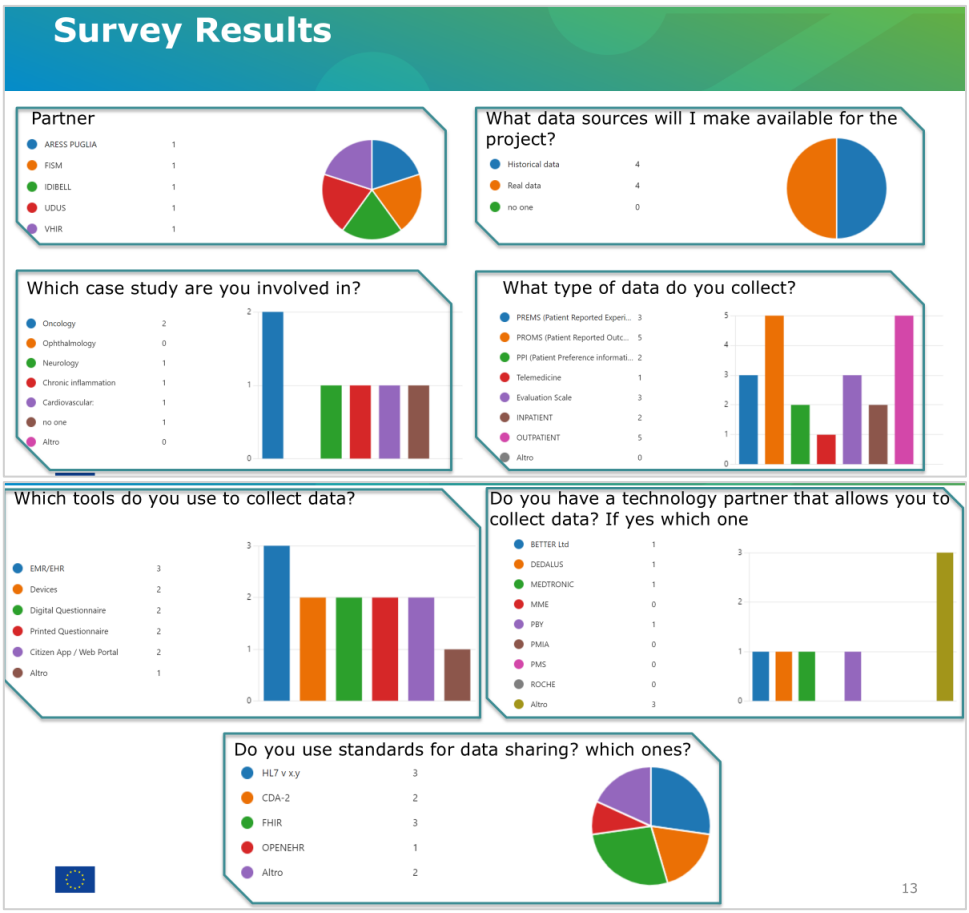


Figure 4: Survey results concerning data points within the IMPROVE consortium

The partners are engaged in various clinical case studies covering oncology (notably prostate, breast, cervical, and head and neck cancers), neurology (multiple sclerosis), chronic inflammatory diseases (including chronic rhinosinusitis), and cardiovascular diseases. These studies collectively involve the collection of PROMs, PREMs, PPIs, telemedicine records, standardized clinical evaluation scales, and both inpatient and outpatient data. Of particular note, PROMs and outpatient data were reported as the most widely available data types, reflecting the project's patient-centric focus. The diversity of data types and sources necessitated a flexible but robust data integration strategy.

		Oncology				Ophthalmology	Cardiovascular				Neurology	Chronic inflammation
	TOPICS	Prostate cancer	Cervical cancer	Neck and head cancer	Breast cancer	Macular degeneration	Heart failure	Coronary artery diseases	Artial fibriliation	Severe aortic stenosis	Multiple sclerosis	Chronic rhinosinusitis
Historical Data	PGHD										PREMS	
				PROMS							PROMS	PROMS
											PPI	
	General Data			INPATIENT								INPATIENT
				OUTPATIENT	OUTPATIENT						OUTPATIENT	OUTPATIENT
					Evaluation Scale						Evaluation Scale	
Real Data	PGHD				PREMS		PREMS				PREMS	
				PROMS	PROMS		PROMS				PROMS	PROMS
					PPI		PPI				PPI	
	General Data			INPATIENT			INPATIENT					INPATIENT
				OUTPATIENT	OUTPATIENT		OUTPATIENT				OUTPATIENT	OUTPATIENT
					Evaluation Scale		Evaluation Scale				Evaluation Scale	

Figure 5: Historical and Real Data across Use Cases

Operationalising the collection of historical data was formalised via a series of Proofs of Concept (PoCs), with PoC #1 dedicated to the acquisition and harmonisation of retrospective datasets. The data collection timeline was significantly influenced by the involvement of Data Protection Officers (DPOs) and the requirement to negotiate Data Transfer Agreements (DTAs) with each partner. These agreements, mandated by GDPR and the sensitivity of patient data, introduced complexity and extended timelines but ensured rigorous compliance with data privacy and security standards. A prominent example of historical data collection is represented by FISM's longitudinal study involving people with Multiple Sclerosis (PwMS). Data are collected every four months through one-hour evaluation sessions, encompassing 109 data points including clinical scales (EDSS, FIM, SDMT, MoCA), patient-reported outcomes (Abilhand, OAB-Q, M-FIS, HADS, Life Satisfaction Index), and other key health metrics. These data provide an invaluable resource for assessing treatment efficacy, patient quality of life, and disease progression.

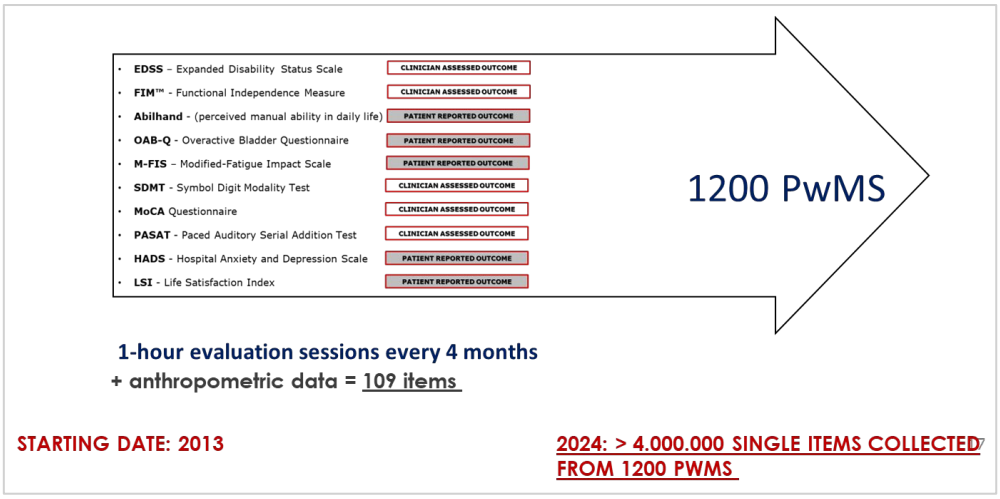


Figure 6: FISM's data collection

Similarly, Philips and The Netherlands Cancer Institute (NKI) contribute extensive datasets on prostate and cervical cancers, including clinical outcomes such as biochemical disease-free survival, overall survival, treatment toxicity, and PROMs related to urinary, bowel, and sexual function quality of life

measures. UDUS offers detailed datasets on chronic rhinosinusitis, capturing clinical symptoms, treatments, diagnostic scores (e.g., SNOT-22, nasal polyp score), and patient experience measures. For head and neck cancer, comprehensive clinical, pathological, therapeutic, and follow-up data elements are recorded, supporting detailed outcome analyses. The technical infrastructure underpinning data collection was developed under Work Package 3 (WP3). Retrospective data are shared with the project's IMPROVE framework via secure APIs developed by UPM. These APIs enable partners to upload data into a centralized repository, structured according to a standardized data model currently based on the HL7 FHIR specification. The data model is evolving to support a federated architecture, allowing for distributed data storage with centralized querying capabilities. Data transformation from source formats to FHIR resources is facilitated by Dedalus's proprietary tool "Picasso," which automates the extraction, transformation, and loading (ETL) processes and stores harmonized data in a MongoDB database. This approach allows seamless integration of heterogeneous datasets and supports scalable data management. Throughout this process, privacy considerations remain paramount. All data transfers are pseudonymized or anonymized as required, and strictly governed by partner-specific DTAs. These agreements delineate roles, responsibilities, and data handling procedures to ensure compliance with GDPR and maintain patient confidentiality.

2.3 Data Source Mapping and Partner Contributions

The consortium's historical data ecosystem comprises multiple heterogeneous datasets contributed by five core partners: FISM, UDUS, AReSS Puglia, Philips, and CatSalut. Each partner's dataset reflects their unique clinical focus and patient cohorts, necessitating careful mapping and harmonisation to align with the project's data model and analytical requirements.

FISM's contributions include longitudinal, multi-modal data from PwMS patients, encompassing clinical assessments, PROMs, and socio-demographic information. The dataset includes repeated measurements at baseline and three follow-up points, offering a rich temporal dimension critical for understanding disease trajectory and treatment outcomes.

Philips in collaboration with the NKI provides detailed prostate and cervical cancer datasets, incorporating a range of clinical endpoints such as biochemical disease-free survival (bDFS), distant metastasis-free survival (dMFS), overall survival (OS), treatment toxicity grades, and patient-reported health-related quality of life metrics (HRQoL) across urinary, bowel, sexual activity, and sexual functioning domains. Philips' data are primarily aggregated, facilitating analysis of population-level trends while respecting patient privacy.

UDUS delivers extensive data on chronic rhinosinusitis and head and neck cancers, including both clinical variables (symptomatology, diagnostic scores, treatment regimens) and patient experience indices. These datasets provide granular information on disease characteristics, therapeutic interventions, and patient outcomes, supporting comprehensive analyses of treatment efficacy and quality of care.

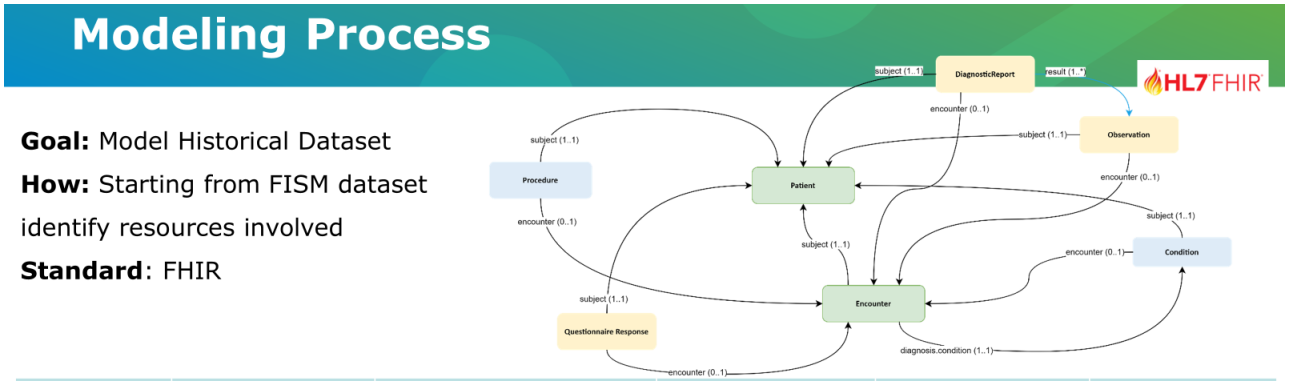
The other partners, AReSS Puglia and CatSalut, contribute complementary datasets aligned with their regional healthcare systems, further enriching the project's data repository.

A critical aspect of data integration is the involvement of clinical partners in defining Key Performance Indicators (KPIs) and selecting relevant data elements aligned with clinical questions and policy objectives. This clinical guidance is essential to ensure that technical partners can effectively apply the data model and ETL pipelines, preserving clinical meaning and analytical utility. The datasets collected are frozen historical data, with no continuous updates planned, allowing for stable retrospective analyses. The data format and granularity vary while Philips shares aggregated data, other partners supply individual patient-level data, pseudonymized to ensure privacy. For instance, FISM's MS dataset

includes three follow-up assessments per patient, whereas UDUS's datasets for head and neck cancer and chronic rhinosinusitis generally contain information related to single visits or episodes. Data ingestion adheres to the architecture defined in WP3, where source data are extracted, transformed to the standardized data model (FHIR-based), and uploaded through UPM's secure APIs to the centralized project repository. The data flow is supported by secure authentication, logging, and error handling mechanisms to ensure data integrity and traceability. Close collaboration between clinical and technical teams is maintained to address challenges in data quality, semantic alignment, and technical compatibility, with iterative feedback loops informing continuous improvements in data processing and model refinement.

2.4 Data Harmonisation and Interoperability Framework

Effective harmonisation and integration of multisource healthcare data are pivotal to the success of this project. The consortium employs a multi-layered interoperability framework leveraging established international standards and terminologies to enable semantic consistency and technical integration. The primary data modelling and exchange standard utilized is HL7 FHIR (Fast Healthcare Interoperability Resources), chosen for its modularity, extensibility, and wide adoption in healthcare IT. FHIR resources represent clinical concepts such as observations, conditions, procedures, and patient demographics, allowing structured and interoperable data exchange.



Patient	Encounter	Condition	Observation	DiagnosticReport	Questionnaire Response
NEU-FISM .identifier (0..*) .gender (0..1) .birthDate (0..1)	CLIN .identifier (0..*) .class (1..1) [inpatient | ambulatory] .type (0..*) .period (0..1)	.category (0..1). severity (0..1) .code (1..1) .bodySite (0..*) .onSet (dateTime or age) (0..1) .recordedDate (0..1) .recorder (0..1) .asserter (0..1) .stage (0..1)	PROMS PREMS .identifier (0..*) .category (0..*). code (1..1) .effective (DateTime or Period) (0..1) .value (0..1 .bodySite (0..*)	.status (1..1) [partial | final | ...].code (1..1) .issued (0..1)	.status (1..1) [in-progress | completed]

Figure 7: Modelling processes for the historical data collection

Terminology standards including SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms), LOINC (Logical Observation Identifiers Names and Codes), and ICD-9 are employed for semantic annotation, ensuring unambiguous coding of clinical findings, laboratory tests, diagnoses, and procedures.

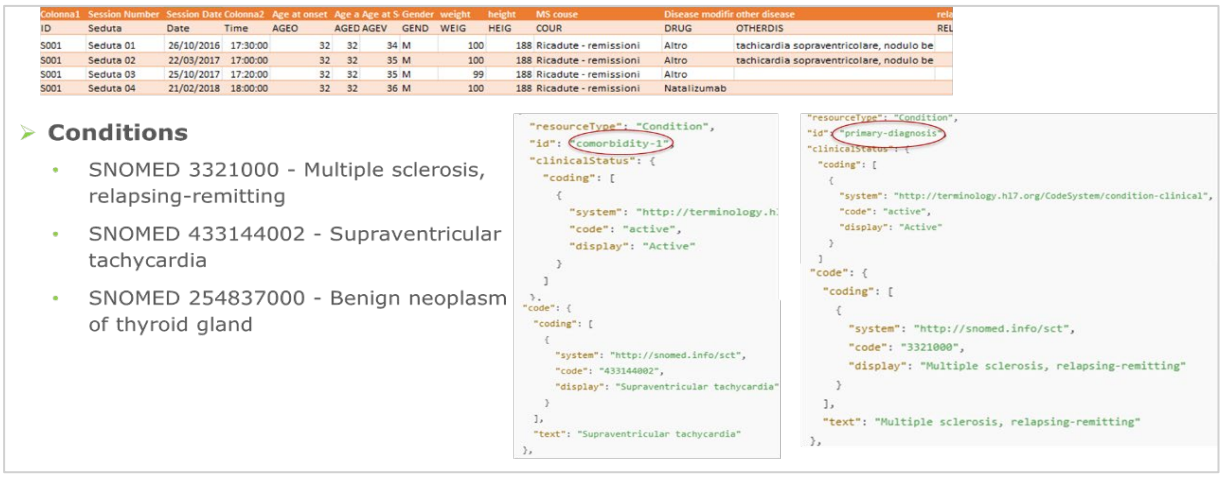


Figure 8: Different SNOMED data sets and disease areas

However, a significant interoperability challenge arises due to the heterogeneous native data standards among partners. Several partners collect clinical data natively in OpenEHR format, which differs structurally and semantically from FHIR. This necessitates a robust transformation process to convert OpenEHR archetypes and templates into FHIR-compliant resources while preserving data fidelity. The current architectural vision embraces a federated data model, wherein data reside locally within partner infrastructures but are accessible through standardized interfaces and shared metadata schemas. This approach mitigates data duplication risks, enhances data governance, and respects partner autonomy and privacy constraints.

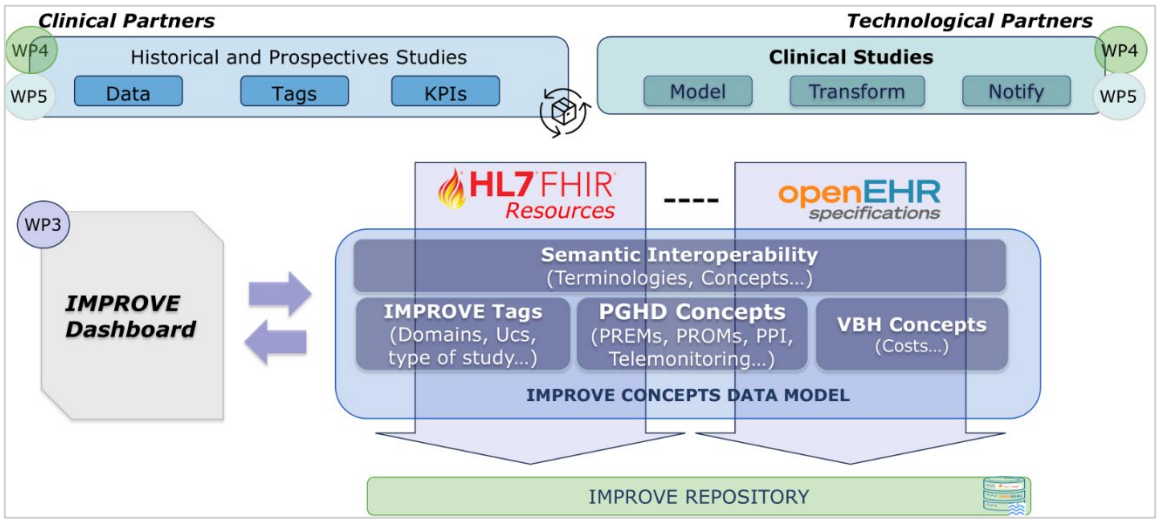


Figure 9: IMPROVE repository of RWD sets from partners

To facilitate data transformation and harmonisation for one of the key use cases, specifically the longitudinal Multiple Sclerosis dataset provided by FISM, the consortium has leveraged Dedalus' proprietary tool called Picasso. This tool automates the extraction, transformation, and loading (ETL) of heterogeneous source data into a MongoDB repository structured according to the FHIR data model. It performs key roles in data harmonisation and storage, ensuring consistency and usability for subsequent analysis

It is important to underline, however, that Picasso's use is currently limited to the FISM use case and has not yet been generalized across the entire project. Other partners follow different data ingestion and transformation workflows adapted to their specific contexts, and the development of a project-wide standardized ETL tool remains ongoing. Data validation and quality control steps are embedded to detect inconsistencies, missing values, and semantic mismatches.

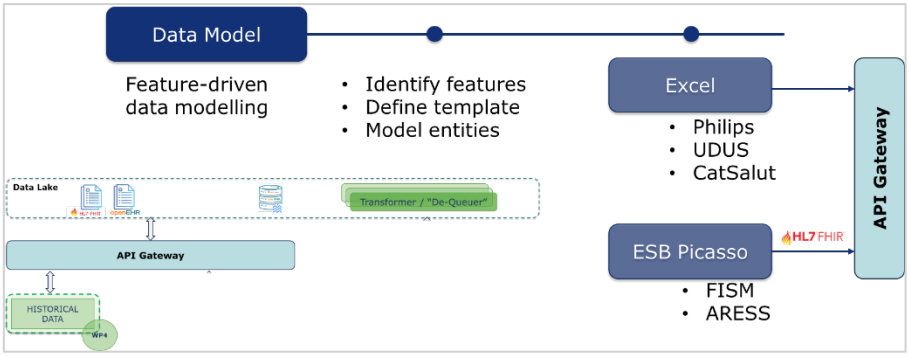


Figure 10: Data model overview

Privacy and security are rigorously enforced throughout the data lifecycle. All datasets are pseudonymized before transfer, and access controls are implemented according to the principles established in partner-specific DTAs and the GDPR regulatory framework. Secure APIs with authentication and authorization guard data ingestion and retrieval processes. The consortium continues to refine the shared data model, developing detailed metadata schemas, provenance tracking, and harmonisation rules to enable cross-partner data querying and analytics. Middleware components are being developed to enable translation between OpenEHR and FHIR standards, facilitating interoperability across differing clinical systems. Governance structures are being established to oversee data stewardship, including data quality monitoring, access policies, and compliance audits, ensuring sustainable and trustworthy data integration.

2.5 Ethical, Legal, and Governance Considerations

The ethical, legal, and governance dimensions of data collection and integration are central to WP4 and to the IMPROVE project as a whole. Because IMPROVE deals with highly sensitive health data, including PGHD from mobile applications, wearables, sensors, and patient-reported instruments, ensuring responsible, compliant, and trustworthy data handling is essential for protecting individuals, enabling cross-border research, and building stakeholder confidence in the IMPROVE framework.

Ethical Considerations: Respect for Persons, Autonomy, and Equity

PGHD carries unique ethical implications because it is collected directly from individuals in their daily lives and often outside clinical supervision. WP4 therefore emphasises transparency, voluntariness, and respect for patient autonomy throughout the data lifecycle. Participants must be informed of how their data will be used, what rights they retain, and how project safeguards ensure privacy and dignity.

Ethical considerations also extend to equity and representativeness. The gap analysis from WP2 and WP3 highlights disparities in PGHD availability across diseases, regions, and populations. WP4 addresses this by prioritising the inclusion of underrepresented groups and contexts, thereby reducing

bias in the resulting models and promoting more equitable VBHC outcomes across Europe. Care is additionally taken to ensure that the integration of PGHD does not inadvertently burden patients or exclude those with limited digital literacy, access to technology, or ability to contribute data, aligning with IMPROVE's broader mission of inclusivity and patient empowerment.

Legal Compliance: GDPR, EHDS, and Related Legislation

WP4 operates within a robust legal framework established by several key European regulations:

- **General Data Protection Regulation (GDPR)** provides the overarching rules for lawful processing of personal and sensitive health data, requiring explicit legal bases, data minimisation, purpose limitation, security, and strong data subject rights.
- The **European Health Data Space (EHDS)** introduces a harmonised structure for the primary and secondary use of health data across Member States, including interoperability requirements and Health Data Access Bodies.
- The **Data Governance Act (DGA)** supports trusted and transparent mechanisms for data sharing, including the role of data intermediaries and data altruism organisations.
- The **Data Act** establishes rights related to connected devices and reinforces fairness, portability, and interoperability in data access and re-use.

WP4 aligns all data collection, linkage, and sharing practices with these frameworks. This includes establishing clear legal bases for all processing activities (Articles 6 and 9 GDPR), implementing privacy-by-design and privacy-by-default principles, ensuring secure data transmission and storage, and using appropriate safeguards such as pseudonymisation, encryption, and controlled access environments.

Governance: Transparency, Accountability, and Data Stewardship

A key objective of WP4 is to implement a robust governance model that ensures accountability, transparency, and compliance across the consortium. This governance model includes:

- **Clear data stewardship roles and responsibilities** across partners supplying data (e.g., IBIDELL, Dedalus, MEDS, ARSS, MM, UDUS, IER, PMSN, NKI).
- **Standardised data management procedures**, including documentation of provenance, metadata, quality controls, and harmonised Common Data Elements (CDEs).
- **Access control policies** that define who can access which data, under what conditions, and for which purpose.
- **Ethical oversight mechanisms**, including compliance checks, Data Protection Impact Assessments (DPIAs), and alignment with institutional and national ethics requirements.
- **Continuous monitoring** to ensure that the data integration pipelines and AI-based processes within the Knowledge Warehouse remain compliant as regulations evolve.

Governance in WP4 is designed to be adaptive, ensuring a structured yet flexible framework that can incorporate new data sources, emerging PGHD modalities, and changes in the regulatory landscape especially as the EHDS and associated national infrastructures become operational.

Responsible Use of AI and Automated Analysis

The increasing reliance on AI, LLMs, and machine learning for data extraction, classification, and integration requires careful consideration of transparency, explainability, and bias mitigation. WP4 follows the principles of the EU Artificial Intelligence Act and the ethical guidelines for trustworthy AI by:

- Documenting algorithmic decision-making processes
- Monitoring model performance across subgroups to detect potential bias
- Ensuring human oversight in all critical processing steps
- Maintaining traceability of automated transformations and classification outcomes
- Implementing safeguards to prevent uncontrolled data inferences or unintended re-identification

These measures ensure that automated analyses remain reliable, fair, and aligned with IMPROVE's ethical commitments.

Cross-Border Data Sharing and Sustainability

As IMPROVE aims to integrate data from multiple European partners and prepare for interoperability with emerging initiatives such as HealthData@EU, WP4 builds a governance structure that supports secure, legally compliant cross-border data exchange. Standardisation, transparency, and stakeholder trust are essential prerequisites for long-term sustainability.

WP4's ethical, legal, and governance framework ensures that PGHD and RWD are collected, processed, and integrated in a way that is compliant, transparent, and respectful of patient rights. By embedding these principles into the core data architecture, WP4 safeguards participants, supports cross-border research collaboration, and establishes a trustworthy foundation for the IMPROVE platform and future European data-driven healthcare systems.

2.6 Quality Assurance and Validation Procedures

Ensuring the quality, reliability, and validity of RWD and PGHD is a central requirement for WP4. Because these data sources vary widely in origin, format, completeness, and methodological maturity, WP4 implements a rigorous quality assurance (QA) and validation framework to guarantee that all data integrated into the IMPROVE platform are fit for analytical use, compliant with legal standards, and appropriate for downstream modelling and use case development.

Data Quality Dimensions and Assessment Criteria

Quality assurance procedures in WP4 are structured around internationally recognised data quality dimensions, including:

- **Completeness** (presence of expected measurements, temporal coverage, proportion of missing values)
- **Accuracy and Veracity** (alignment with source records, consistency with expected ranges or clinical plausibility)
- **Consistency and Standardisation** (harmonised variable definitions, controlled vocabularies, uniform metadata)
- **Timeliness and Currency** (alignment with the update cycles of data providers, ability to incorporate new evidence)
- **Interoperability** (mapping to common standards such as FHIR, OMOP, HL7, and project-specific CDEs)
- **Traceability and Provenance** (ability to track the origin, transformation history, and versioning of data elements)

Each dataset contributed by partners (e.g., IBIDELL, Dedalus, MEDS, ARSS, MM, UDUS, IER, PMSN, NKI) undergoes an initial quality profiling step to evaluate its structure, completeness, and alignment with these criteria. This creates a reproducible baseline for integration, transformation, and later validation.

Validation of PGHD and RWD Integration Pipelines

Given the heterogeneity of PGHD sources, including mobile apps, wearables, sensors, PROMs/PREMs, and patient preference information, WP4 employs a multi-layered validation approach to ensure safe and reliable integration:

1. **Format and Schema Validation:** Validation of file formats, syntactic consistency, adherence to predefined schemas, and correct mapping to CDEs and ontologies.
2. **Semantic Validation:** Review of variable meaning, unit harmonisation, and alignment between provided metadata and real data content.
3. **Automated Consistency Checks:** Rule-based and machine-learning-supported checks to detect anomalies, outliers, duplicate entries, implausible sequences, or logical inconsistencies.
4. **Cross-Source Validation:** When feasible, PGHD are compared with clinical RWD (e.g., EHRs, registries) to evaluate convergence, discrepancies, and reliability across data types.
5. **Manual Expert Review:** Validation by clinicians, data scientists, and domain experts to interpret ambiguous elements, confirm automated classifications, and ensure contextual accuracy.

These steps ensure that integrated PGHD not only meet technical standards but also reflect clinically meaningful and patient-relevant information.

AI/ML Model Validation and Continuous Monitoring

Because the IMPROVE framework increasingly relies on AI, LLMs, and machine learning for data classification, extraction, and interpretation, WP4 implements dedicated validation procedures to ensure that automated processes remain reliable, transparent, and unbiased:

- **Ground-Truth Benchmarking:** Automated model outputs (e.g., LLM-based extraction) are compared against manually curated gold-standard subsets to assess precision, recall, and error types.
- **Bias and Drift Monitoring:** Regular evaluation of model performance across demographic subgroups, disease areas, and data providers to identify systematic biases or temporal drift.
- **Explainability and Traceability Measures:** Logging of model decisions, intermediate representations, and rationale outputs to ensure auditability and compliance with EU AI Act principles.
- **Version Control and Re-Training Protocols:** Clear procedures for updating models when new data become available or when performance drops below expected thresholds.

These measures safeguard the reliability of the automated analytics that power the Knowledge Warehouse and the IMPROVE framework.

Quality Control in Data Transformation and Harmonisation

As WP4 performs the harmonisation, cleaning, reformatting, and linkage of diverse datasets, each transformation step undergoes:

- Pre- and post-transformation checks
- Automated pipeline validation using reproducible scripts
- Controlled vocabularies and ontologies to enforce semantic alignment
- Documentation of all transformations to ensure full data lineage transparency

This guarantees that the data fed into WP5 and the model system in WP3 remain consistent, interpretable, and analytically sound.

Stakeholder Validation and User Feedback Loops

Quality assurance is reinforced through structured interactions with clinicians, patient groups, data providers, and other stakeholders. Feedback loops ensure that:

- data elements reflect real-world workflows and patient experiences,
- variables are clinically interpretable and actionable,
- quality issues identified during use case testing are communicated back to WP4 for iterative improvement.

This aligns QA with the user-centred design approach adopted across the project.

WP4's quality assurance and validation framework ensures that all RWD and PGHD entering the IMPROVE ecosystem meet high standards of accuracy, reliability, and legal compliance. By combining automated validation, expert review, interoperability checks, and continuous monitoring of AI/ML processes, WP4 provides a robust and transparent foundation for the downstream modelling activities of WP3 and the implementation and evaluation activities in WP5.

3. Data sources and contributions

3.1 Overview of Existing Data Providers (CatSalut, Dedalus, MEDS, ARSS, MM, UDUS, IER, PMSN, NKI)

The primary data providers involved in the project are FISM, UDUS, AReSS, Philips, NKI and CatSalut. Each partner contributes specific datasets relevant to their clinical and research domains, which are essential to support the project's objective of collecting and analysing retrospective and patient-generated health data (PGHD) to aid policymakers, clinicians, and other stakeholders. FISM provides clinical data related to multiple sclerosis, including PROMs collected at multiple follow-up points. UDUS supplies data focused on head and neck cancer as well as chronic rhinosinusitis, primarily based on single-visit clinical records. Philips contributes aggregated datasets for prostate cancer, containing both clinical and patient-reported outcomes. AReSS and CatSalut provide additional pseudonymized clinical and administrative datasets. These partners supply historical, frozen datasets that do not undergo continuous updates. The data collection and sharing processes are coordinated under a unified framework to ensure data compliance, quality, and standardization, which aligns with the project's goal to facilitate multi-stakeholder decision-making based on harmonized data.

Table 1 (see below) provides an overview of the data providers contributing to the IMPROVE project, the clinical domains covered, and the main characteristics of the datasets shared. It summarises the type and level of data supplied by each partner, the key variables and outcomes captured, and the temporal scope of the datasets. The table also highlights the specific role of each data source in supporting the project's objectives, including the development and validation of patient-centred, real-world data-driven analytical models.

Data Provider	Clinical Domain(s)	Data Type	Level of Data	Key Variables / Outcomes	Temporal Scope	Role in IMPROVE
FISM	Multiple sclerosis	Clinical data, PROMs	Individual-level, pseudonymized	Anamnesis, PROMs (3 follow-ups), treatment outcomes	Retrospective, static	Longitudinal PGHD and chronic disease modelling
UDUS	Head & neck cancer; chronic rhinosinusitis	Clinical data, PREMs	Individual-level, pseudonymized	Diagnostics, treatments, PREMs	Retrospective, single-visit	Real-world clinical pathways
Philips	Prostate cancer	Clinical outcomes, PROMs	Aggregated	Disease-free survival, toxicity, HRQoL	Retrospective, static	Outcome benchmarks

NKI	Prostate cancer	Clinical outcomes, PROMs	Aggregated	Survival outcomes, toxicity, HRQoL	Retrospective, static	Oncology expertise and RWE
AReSS	Regional healthcare services	Clinical & administrative data	Pseudonymized	Healthcare utilisation, clinical indicators	Retrospective, static	System-level context
CatSalut	Public healthcare system (Catalonia)	Clinical & administrative data	Pseudonymized	Healthcare encounters, indicators	Retrospective, static	Population-level context

Table 1. Overview of data providers and contributions

3.2 Characteristics and Scope of Data from Each Partner

- FISM (Federazione Italiana Sclerosi Multipla) provides retrospective, pseudonymized individual-level clinical data on multiple sclerosis patients. The dataset includes anamnesis data, PROMs collected over three follow-up sessions, and patient-reported outcomes aimed at evaluating multidisciplinary treatment effectiveness. These data are formalised according to the FHIR data model and stored in the shared project repository.
- UDUS contributes clinical data focused on head and neck cancer as well as chronic rhinosinusitis. The data primarily consist of single-visit patient records, including clinical observations, diagnostic details, treatment information, and patient-reported experience measures (PREMs). All data are pseudonymized and mapped to the common data model.
- Philips together with NKI supplies aggregated data related to prostate cancer patients. This dataset includes clinical outcomes such as biochemical disease-free survival, toxicity grades, and health-related quality of life metrics obtained through PROMs. These aggregated data complement the individual-level data provided by other partners.
- AReSS and CatSalut contribute pseudonymized clinical datasets relevant to their healthcare domains. Their data are retrospective, mapped to the project's data model, and shared under specific Data Transfer Agreements (DTAs).

All datasets are historical and static, with no ongoing updates expected. Collaboration between clinical partners, who define key performance indicators and relevant data elements, and technological partners, who support data transformation and integration, ensures consistent data harmonisation aligned with the project's data architecture.

3.3 Contribution to Model System Development in WP3

The outputs of WP4 play a critical role in advancing the model system developed under WP3. While WP3 establishes the conceptual foundation for how PGHD, RWD, and contextual information should be represented, interpreted, and integrated, WP4 provides the empirical data layer needed to transform these conceptual structures into functional, testable components of the IMPROVE platform. By identifying and supplying structured, high-quality datasets, WP4 enables WP3 to refine its AI- and ML-driven analytical modules, particularly those that classify PGHD types, model patient journeys, and connect data elements to relevant VBHC outcomes. The enriched datasets contribute to improving LLM-based extraction, validating ontologies, and enhancing the reliability of machine learning models by exposing them to heterogeneous real-world contexts. In practice, this means that the model system benefits from richer data variability, better representation across disease and population groups, and improved calibration of algorithms against real cases rather than abstract definitions.

Moreover, the targeted data collection addressing gaps identified in WP2 and WP3 allows WP3 to test hypotheses about the role of PGHD in clinical decisions, device design, and pathway optimization. It also enables iterative co-creation cycles, where early model outputs can be validated against real-world patterns and stakeholder expectations. In this way, WP4 ensures that the model system becomes evidence-driven, context-aware, and robust across multiple use cases.

3.4 Added Value for the IMPROVE Framework

The work conducted in WP4 provides significant added value to the overall IMPROVE framework by operationalising the core principles emerging from WP2 and WP3 and grounding them in real-world, patient-centred evidence. Through structured collection, validation, and integration of RWD and PGHD, WP4 transforms the theoretical architecture into a platform that can support actionable insights for VBHC and HTA. First, WP4 enriches the framework by broadening the data basis beyond traditional clinical sources, thereby enabling a more comprehensive picture of patient needs, experiences, and preferences. This multidimensional data layer is essential for capturing value in a way that aligns with what matters most to patients. Second, WP4 ensures that the framework is not limited to the best-documented conditions but also applies to underrepresented diseases, populations, and contexts, improving fairness, inclusiveness, and generalisability. Additionally, WP4 strengthens the IMPROVE framework by embedding interoperability and governance principles into its data acquisition workflows. By aligning with GDPR, the EHDS, the Data Act, and the DGA, the data collected through WP4 can be confidently reused in research, policy, and clinical decision-making without compromising legal or ethical requirements. This alignment not only ensures compliance but also enhances trust, making the IMPROVE framework more acceptable to stakeholders, including regulators, clinicians, and patients. Ultimately, the added value of WP4 lies in making the framework usable, scalable, and future proof, enabling meaningful integration of PGHD and RWD into real-world practice.

3.5 Long-term Utility for RWD-Based Research

The datasets, workflows, and data governance structures developed in WP4 are designed with long-term sustainability and reusability in mind, extending far beyond the immediate operational needs of the IMPROVE project. By establishing standardised, interoperable, and ethically governed mechanisms

for collecting and integrating PGHD with broader RWD sources, WP4 lays the groundwork for a durable research infrastructure that will continue to hold relevance as European data ecosystems, particularly those shaped by the EHDS, become fully operational. This forward-looking approach ensures that the outputs of WP4 can support continued learning, innovation, and collaboration, even after the completion of the project.

One of the most significant long-term benefits is the creation of a reusable and expandable data ecosystem. This ecosystem facilitates ongoing research into patient-centred outcomes, chronic disease trajectories, digital health interventions, and VBHC models. Because it is built around modular and interoperable components, it can be continuously enriched with new data sources and measurement instruments as technologies evolve and as additional RWD and PGHD streams emerge. In this way, WP4 provides a living, evolving foundation for future scientific inquiry. The Knowledge Warehouse, strengthened through the curated datasets and validated metadata contributed by WP4, is well-positioned to function as a continuously updated repository of evidence. It can support new waves of systematic reviews, comparative analyses, meta-research, and AI-driven inference long after IMPROVE has ended, providing lasting value to the broader research community.

WP4's explicit focus on semantic interoperability, technical harmonisation, and regulatory alignment positions the IMPROVE framework to integrate seamlessly with major European initiatives such as HealthData@EU, national EHDS access bodies, and emerging cross-sectoral data spaces. By meeting the technical and procedural requirements of these evolving infrastructures, the IMPROVE architecture becomes a potential reference implementation for ethically responsible, patient-centred, cross-border data reuse. This alignment enhances the ability of PGHD and RWD to travel across systems and jurisdictions, enabling multi-country collaborations, large-scale federated studies, and evidence synthesis that reflect Europe's diversity.

In addition, the long-term utility of WP4 extends directly to regulatory and HTA contexts. The structured datasets generated through WP4 can support a range of high-value analytical activities, including comparative effectiveness studies, patient preference assessments, device evaluation, pathway optimisation, and real-world monitoring of outcomes and safety. As health systems increasingly adopt hybrid evidence models combining clinical data, RWD, and PGHD, the tools and data structures created in WP4 provide a tested foundation for producing robust, reproducible insights that can inform healthcare decision-making at national and EU level. These assets also enable continuous learning healthcare systems, in which real-world feedback loops drive iterative improvements in treatments, digital tools, and clinical pathways.

The methodological innovations introduced in WP4, including LLM-based evidence extraction, automated metadata validation, harmonised PGHD ontologies, and reproducible data integration pipelines, constitute durable assets that can be replicated, scaled, or adapted by future research initiatives. They demonstrate how advanced AI techniques can be responsibly embedded into scientific workflows while maintaining transparency, quality control, and alignment with GDPR, the EHDS, the Data Act, and the DGA. These innovations help position IMPROVE as a model for next-generation data-

driven research infrastructures capable of integrating heterogeneous data types in a reliable, scalable manner.

In summary, WP4 not only advances the immediate objectives of the IMPROVE project but also contributes to the establishment of a lasting, high-quality RWD infrastructure that strengthens Europe's long-term capacity for data-driven, patient-centred innovation. The structures built here, technical, methodological, and governance-related will continue to support future research, regulatory innovation, and patient empowerment well beyond the project's lifespan.

4. Conclusions and Next Steps

4.1 Summary of Key Insights

This deliverable has documented the outcomes of the initial phases of WP4, focused on identifying additional data collection needs and defining the methodological and technical foundations required to integrate historical clinical data and PGHD into the IMPROVE framework. The work confirms that the systematic alignment between evidence synthesis (WP2), conceptual and technical architecture (WP3), and use case requirements (WP5) is essential to ensure that Real-World Data (RWD) meaningfully contributes to VBHC decision-making. A first key insight concerns the heterogeneity and fragmentation of existing PGHD and RWD. Although substantial volumes of historical clinical data and PROMs exist for certain disease areas, significant gaps remain with respect to PREMs, PPIs, longitudinal trajectories, and underrepresented diseases and populations. These gaps are not accidental but structural, stemming from inconsistent data collection practices, limited standardisation, and uneven geographical coverage. Second, WP4 activities have demonstrated the central importance of interoperability and semantic harmonisation. The mapping of heterogeneous datasets to shared data models (e.g. HL7 FHIR) and controlled terminologies is not a purely technical exercise, but a prerequisite for preserving clinical meaning and enabling cross-disease, cross-country, and cross-stakeholder analyses. Proofs of Concept developed during WP4 have validated the feasibility of this approach while also highlighting the need for iterative refinement and close collaboration between clinical and technical partners.

Third, the work confirms that governance, ethics, and legal compliance must be embedded by design. Historical data collection and reuse, particularly for PGHD, require clear legal bases, robust pseudonymisation, and transparent governance mechanisms aligned with GDPR, the EHDS, the Data Act, and the DGA. Addressing these aspects proactively has proven crucial for enabling data sharing across partners and for building trust in the IMPROVE framework. Finally, WP4 has shown that RWD is a core enabler of the IMPROVE vision, rather than a supplementary component. By operationalising RWD and PGHD integration, WP4 transforms the conceptual models developed in WP2 and WP3 into actionable, testable assets that can support AI-enabled analytics, VBHC evaluation, and real-world implementation.

4.2 Transition to Subsequent Tasks and Work Packages

The outcomes presented in this deliverable mark a clear transition point from planning and specification activities towards full-scale implementation and iterative refinement. Within WP4, the focus now shifts from defining data needs and methodologies (Task T4.1) to the systematic execution of historical data collection and gap analysis (Task T4.3), supported by complementary methods (Task T4.4).

The specifications and insights described here directly inform the next phases of WP3, where the availability of harmonised historical datasets will enable further validation and consolidation of the IMPROVE model system. The enriched data streams provided by WP4 will be used to calibrate AI and machine learning components, refine PGHD ontologies, and test the robustness of analytical pipelines across different clinical and organisational contexts. In parallel, the work supports WP5 use case

implementation, ensuring that each use case is underpinned by data that are representative, longitudinal, and aligned with VBHC objectives. The explicit identification of disease-specific, geographical, and temporal data gaps allows WP5 activities to be targeted and evidence-driven, reducing the risk of incomplete or biased evaluations. Crucially, WP4 acts as a connecting layer across work packages, translating conceptual knowledge into operational data assets and feeding real-world insights back into model development and evidence synthesis. This iterative interaction between WPs is expected to intensify in the coming project period, with continuous feedback loops guiding both technical development and stakeholder engagement.

4.3 Continuation of Data Collection Across the Project

Data collection within IMPROVE is conceived as a continuous, adaptive process, rather than a one-off activity. Building on the foundation established in this deliverable, WP4 will continue to support the project through ongoing data acquisition, harmonization, validation, and enrichment activities.

Future data collection efforts will prioritize:

- the extension of longitudinal datasets to better capture complete patient journeys;
- the inclusion of underrepresented diseases, regions, and populations;
- the integration of additional PGHD modalities, including app-based, wearable, and sensor-derived data;
- the strengthening of linkages between PGHD, traditional clinical RWD, and contextual policy and practice information.

These activities will remain closely aligned with evolving insights from WP2 reviews, WP3 model testing, and WP5 piloting, ensuring that data collection strategies remain responsive to emerging needs. At the same time, WP4 will continue to refine data governance, quality assurance, and interoperability mechanisms to ensure sustainability and future reuse beyond the project's lifetime.

In conclusion, WP4 establishes the empirical backbone of the IMPROVE framework. By continuing systematic data collection and integration throughout the project, it ensures that IMPROVE evolves into a robust, scalable, and trustworthy infrastructure capable of supporting patient-centred, data-driven, and value-based healthcare across Europe.

About IMPROVE

IMPROVE aims to be a dynamic, ready-to-use framework for seamlessly integrating patient-reported information. This adaptable system constantly evolves with the latest evidence, using PGHD and health system data to provide cost-effective solutions for diverse treatment conditions in real settings. The project follows Ontology, Epistemology, and Methodology principles. Ontology defines structures in patient-reported outcomes; Epistemology ensures valid knowledge; Methodology links techniques to outcomes, systematically addressed in its work.

IMPROVE optimizes patient-reported information in real settings, offering a deep understanding of patient behaviors. The project sets up ontology, epistemology, and methodology to minimize the burden on stakeholders cost-effectively. It adopts a scalable, data-driven approach with NLP-driven knowledge extraction. Real World Data is integrated into the Federated Causal Evidence module for comprehensive understanding. Evidence collected enables visualizing attributes affecting patient-reported outcomes through IMPROVE Engagement Factors and Indicators Knowledge Graphs.

IMPROVE's toolkit includes resources for decision-makers, featuring plausible scenarios via the Copenhagen Method. Patient engagement via the MULTI-ACT model ensures sustainable healthcare aligned with patient priorities. This project delivers a modular, open access strategy, providing a trustworthy ecosystem of evidence-based applications. Patient engagement and co-creation scenarios solidify its role in transforming healthcare research and care.

Funding Acknowledgement

This project is supported by the Innovative Health Initiative Joint Undertaking (IHI JU) under grant agreement No. 101132847. The JU receives support from the European Union's Horizon Europe research and innovation program and COCIR, EFPIA, EuropaBio, MedTech Europe, Vaccines Europe, and the contributing partners Universidad Politécnica de Madrid (Spain), PredictBy (Spain), Danish Medicine Agency (Belgium), Roche (Switzerland), Institute for Economic Research (Slovenia), Copenhagen Institute for Futures Studies (Denmark), Servei Català de la Salut (Spain), Philips Medical System Nederland BV (The Netherlands), Heinrich-Heine-Universitaet Duesseldorf (Germany), Tilburg University (The Netherlands), Dedalus (Italy), Fondazione Italiana Sclerosi Multipla Fism Onlus (Italy), AReSS Puglia (Italy), MultiMed (Italy), iserundschmidt GmbH (Germany), Better (Slovenia), The Netherlands Cancer Institute (The Netherlands), University of Applied Sciences St. Pölten (Austria), Eye Hospital, University Medical Centre Ljubljana (Slovenia), Utrecht University (The Netherlands), Medtronic Iberica SA (Spain), Fundacio Hospital Universitari Vall D'Hebron – Institut de Recerca (Spain), Splosna Bolnisnica Celje (Slovenia), ORTOPEDSKA BOLNIŠNICA VALDOLTRA (Slovenia), ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS (Greece), UDG Alliance (Switzerland).

Disclaimer

www.ihi.europa.eu

Supporters of the Innovative Health Initiative Joint Undertaking:

Project partners:

Coordinator

Associated Partner