

## D4.2. Data Inventory & Characterisation



**CLIMATE CHANGE AND AIR CONTAMINATION:  
ARTIFICIAL INTELLIGENCE APPLIED ON THE CORRELATION BETWEEN AIR  
POLLUTANTS AND NON-COMMUNICABLE RESPIRATORY DISEASES IN EUROPE**

Grant Agreement Number 101156799

Deliverable name: D4.2 Data Inventory & Characterisation  
Deliverable number: D4.2  
Deliverable type: Report  
Work Package: WP4: Framework for Data Management and Interoperability  
Lead beneficiary: IDE Team  
Contact person: Angel de los Santos; angel.delossantos@idener.ai  
Dissemination Level: Public  
Due date for deliverable: December 31, 2025



**Funded by the  
European Union**

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them.



## DOCUMENT CONTROL PAGE

Author(s):	IDE Team
Contributor(s):	Angel de los Santos
Reviewer(s):	Carlos Sánchez (IDE), Savvas Petanidis (AUTH)
Version number:	v.1.4
Contractual delivery date:	31-12-2025
Actual delivery date:	31-12-2025
Status:	Final

## REVISION HISTORY

Version	Date	Author/Reviewer	Notes
v.0	[26-11-2025]	[Ángel de los Santos /Rev name/s]	Creation, First Draft
v.1	[08-12-2025]	Carlos Sanchez	Review
v.1.2	[17-12-2025]	Angel de los Santos	Ready for external review
v.1.3	[19-12-2025]	Savvas Petanidis	Review
v.1.4	[19-12-2025]	Angel de los Santos	Final version submitted

## ACKNOWLEDGEMENTS

The work described in this publication was subsidised by Horizon Europe (HORIZON) framework through the Grant Agreement Number 101156799.

## DISCLAIMER

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them.

All the contributors to this deliverable declare that they:

- Are aware that plagiarism and/or literal utilisation (copy) of materials and texts from other Projects, works and deliverables must be avoided and may be subject to disciplinary actions against the related partners and/or the Project consortium by the EU.
- Confirm that all their individual contributions to this deliverable are genuine and their own work or the work of their teams working in the Project, except where is explicitly indicated otherwise.
- Have followed the required conventions in referencing the thoughts, ideas and texts made outside the Project.

## TABLE OF CONTENTS

DOCUMENT CONTROL PAGE.....	2
REVISION HISTORY .....	2
ACKNOWLEDGEMENTS.....	3
DISCLAIMER .....	4
TABLE OF CONTENTS.....	6
EXECUTIVE SUMMARY .....	7
1. INTRODUCTION.....	8
1.1 DESCRIPTION OF THE DOCUMENT AND PURSUE .....	8
1.2 WPS AND TASKS RELATED WITH THE DELIVERABLE .....	9
2. DATA SOURCES INVENTORY AND CHARACTERISATION PROCESS.....	10
2.1 INTERNAL DATASET INVENTORY .....	10
2.2 EXTERNAL DATASET INVENTORY.....	12
2.3 DATA SOURCES CHARACTERIZATION .....	13
2.4 DATA QUALITY ASSESSMENT.....	18
3. GAP ANALYSIS .....	21
4. INTEGRATION STRATEGY.....	22
5. CONCLUSION .....	25

## EXECUTIVE SUMMARY

This report presents deliverable D4.2 Data inventory and characterisation, which exposes an inventory with the different datasets (internal or externally gathered) used for the model training tasks and the datasets used for analysis later during the project. It also lays the common groundwork to define protocols and techniques to ensure a proper characterisation of the data collected and the improvement of quality and security of the final datasets.

The report outlines the importance of integrating characterisation and quality checking steps at different stages during data collection in order to minimize the transference of data and reduce accessibility to further maintain security and anonymity of data.

The process by which data from different sources will be integrated together is also explained, mentioning the use of a personal anonymized identifier for each patient and the use of geolocation data to create records of data for each individual, creating a new proprietary dataset with information collected from 1906 patients of 9 different medical centres.

In this last section, the management of databases is also briefly explained telling the reader who can have access to which databases and how are datasets moved around local and global databases and how are model weights aggregated. This process is critical based on the use of a Federated Learning (FL) framework, an innovating data treatment scheme that ensures privacy and minimizes data movement between systems and restricts access to raw data.

## 1. INTRODUCTION

### 1.1 DESCRIPTION OF THE DOCUMENT AND PURSUE

Respiratory NCDs, such as chronic **obstructive pulmonary disease (COPD)**, **asthma**, and **allergic rhinitis** evolve under the influence of multiple interacting factors rather than remaining static. Their progression and impact are highly dynamic and is shaped by factors like **environmental exposure** (i.e. air pollution, smoking habits, climate change), **biological variability** (i.e. genetic predispositions, immune responses, comorbidities) and even **socioeconomic factors** (i.e. lifestyle, access to treatment). The variability of these factors complicates the adoption of standardized data collection methodologies and the curation of homogeneous datasets. Furthermore, the process of curating and synthesizing the available outputs from other EU research projects or public datasets to conduct a comprehensive analysis concerning risk factors to respiratory NCDs is notably time-consuming, **potentially impeding research outcomes**.

The purpose of this document is to provide a **characterisation and analysis of the relevant datasets** provided by medical partners and climate agencies. The datasets identify **climate, pollution and health variables** that could pose a threat to the health of respiratory NCD patients. During the clinical study 1906 patients will be monitored to create internal anonymized datasets usable to conduct later analysis, using **AI techniques** and **FL** to address critical health issues, contributing to public health research and policy formulation.

## 1.2 WPS AND TASKS RELATED WITH THE DELIVERABLE

This deliverable is produced under Task 4.2, "**Identification and Analysis of Internal and External Data Sources**", which takes part of WP4, "**Framework for data management and interoperability**", within the ClimAir project. The overarching objective of **WP4** is to **establish a safe and secure data processing framework aligned with FAIR principles**. This involves the **integration of healthcare data streams** to create a comprehensive repository using FL to ensure data integrity and secure sharing while allowing for decentralized AI training.

The specific goal of **Task 4.2** is to **characterise and assess both internal and external data sources** relevant to understanding the causal relationships between environmental risk factors, specifically air pollutants like ultrafine particles and black carbon, and health outcomes. This work follows the conceptual model defined in Task 4.1 and will lay the ground for the development of Task 4.3, which continues the integration strategy defined in Task 4.2

The purpose of this document is to provide comprehensive guidance on the procedures necessary for effective **data source characterisation** and identification of the planned datasets, facilitating the successful execution of the project's designated activities.

## 2. DATA SOURCES INVENTORY AND CHARACTERISATION PROCESS

A data sources inventory is a catalogue created for all the data collected and processed during the development of a project in order to detail the origin of the data used. Efficient use of data inventories includes entries with information about the source of the data, its format, storage location, purpose of use and transferring method. Data inventories are often needed in order to accomplish compliance and privacy standards required by privacy laws like GDPR and support Data Protection Impact Assessments (DPIA). These inventories are essential as they provide a comprehensive and organized record for data assets used, ensuring transparency and reproducibility.

Data characterization is the process of summarizing the general characteristics or features of a target class of data, providing concise and precise descriptions of datasets or data classes in general terms. It typically begins with collecting data corresponding to the problem specified via queries. For this, simple data summaries based on statistical measures and plots are commonly used for effective summarization and characterization, where techniques like data profiling are used, a hierarchical process that analyses raw data to characterize the embedded information within a dataset. <sup>1</sup>

ClimAir makes use of three different kinds of data. It nourishes from pollution, climate and health records, normalizing this input to later feed the FL models developed on Task 4.4.

### 2.1 INTERNAL DATASET INVENTORY

Collecting **internal datasets** for a project involves obtaining and organizing data provided by a project partner for use in analysis, model training, or decision-making processes. Since these datasets come from **trusted partners**, they can provide valuable, **context-specific insights** that may not be available from public data, **ensuring data quality and consistency**, and addressing privacy or compliance requirements before integration. The datasets used during the development of this project will be divided in pollution, climate and health datasets:

#### Pollution

- **SILAM**: Is an open code **global to meso-scale dispersion model** developed for atmospheric composition, air quality, and emergency decision support applications, as well as for inverse dispersion problem solution. SILAM source terms include point- and area- source inventories, sea salt, wind-blown dust, natural pollen, natural volatile organic compounds, nuclear explosion, as well as interfaces to ship emission system STEAM and fire information system IS4FRIES. It includes **allergenic pollen and air quality forecasts**. It is expected to be the main source of information for the pollution variables, provided by FMI.
- **GLORIA**: To create this dataset, the SILAM system was applied to assess the **AQ over the globe**, producing and evaluating **homogeneous hourly time**

---

<sup>1</sup> ScienceDirect – Data Characterization - <https://www.sciencedirect.com/topics/computer-science/data-characterization>

- series** for O<sub>3</sub>, NO<sub>2</sub>, PM<sub>2.5</sub>, PM<sub>10</sub>, SO<sub>2</sub>, CO, NO, NH<sub>3</sub> (totally, 86 gaseous and aerosol species) for the period 1980-2015. This dataset is provided by FMI.
- **European Pollen Reanalysis (EPR):** This dataset is the European reanalysis of pollen seasons for alder, birch, and olive. Driven by the European meteorological reanalysis **ERA5**, the atmospheric composition model SILAM had calculated the **flowering and pollen dispersion patterns** from these trees for the period of 1980-2022 for Europe. The control variable of assimilation was the total pollen release during a flowering season computed independently for each year and type of tree. This dataset is provided by FMI.
  - **Local and National Pollution and climate:** Greece, Poland, France and Luxembourg. These datasets include historical and current information on pollen levels, particulate matter and other pollutants for each of the specified countries that provide them.
  - **Air pollution data (Genoa):** This dataset includes the yearly rate of the principal pollutants concentration monitored by 10 site stations, collecting records on **NO<sub>2</sub>, PM<sub>10</sub>, PM<sub>2.5</sub>, O<sub>3</sub>, SO<sub>2</sub> and CO**. Data includes the average calculation on concentrations hourly and daily.

#### Climate

- **FaD (Málaga & Berlin):** These datasets are the input for the flow and dispersion models for Malaga and Berlin, respectively. The GeoTiff files contains 4 bands: Band 1 represents the digital terrain model, Band 2 represents the buildings height, Band 3 represents the vegetation height and Band 4 represents the vegetation density. These datasets are provided by RSS.
- **ClimAir LES:** Following WPs 7-8, AQ and pollen high-resolution exposure will be created using LES-SILAM models.

#### Health

- **1906 patients' clinical data (ClimAir Platform):** This dataset will be created during the development of the project. A platform has been created to allow clinicians to enter the details of the patients of the study within their local server. This data will be **collected yearly** (during pollen season). It includes **demographic information, symptoms, comorbidities, treatment information, rhinitis symptoms details and medical testing information**. Each of the 9 medical centres within the consortium participates in the collection of this data, averaging 200 entries per medical centre. The data collected by medical partners will be normalized since the beginning, as each medical partner uses the ClimAir Platform to enter the data in the database, having a set of predefined units and values available to enter in the platform. This effort simplifies later data processing and ensures data homogenization among medical centres. Patients are asked to go to the clinics once a year over the next three years, updating patients information every year.

## 2.2 EXTERNAL DATASET INVENTORY

External datasets are sets of records publicly available (free or paid) that can offer sufficient ground for AI testing or training as well as providing enough data to perform complementary statistical analysis and incrementing the precision of the analysis performed.

The external datasets used need to be curated and characterizable to mold them to the project's necessities, so only trustworthy sources should be used, and a previous study of their potential contribution and integration must be conducted to avoid confusing datasets or useless filling data that would worsen the analysis output. Each external dataset served a function during the development of this project, including:

- **MaskAir** dataset: This dataset will come as a result of contracting the Mask Air services in the project. A personalized study has been requested to create a dataset containing most of the information included in the ClimAir tool. Users can report daily symptoms using visual analogue scales (i.e. nasal symptoms, ocular symptoms) and log their medication use. This data **is collected daily**, and it's **entered by the patients themselves**, includes **demographic information, symptoms, comorbidities, treatment information and rhinitis symptoms details**. It was identified its potential during the first stages of the project, thinking on using more data to improve the models accuracy and allowing us to use this data to make further analysis on the evolution of patient's situations.
- **ERA5**: Dataset used by FMI for several meteorological and climatological studies and research on events such as sea-effect snowfall on the Finnish coast, thunderstorm events, and wind energy applications. Its variables are included will be used in the climatological analysis tasks during this project.
- **Air pollution in Seoul (Kaggle)**: Used to test synthetic dataset creation frameworks on python. It's a free dataset including basic air pollution data from Seoul. The similarity of these datasets to the air pollution data collected on this project allowed for a dummy test to check if synthetic dataset augmentation technique would be usable.

ERA5 and MaskAir data will be used during modelling phases, introducing data collected in those datasets in the training and validation steps of the model. Meanwhile, Air pollution in Seoul and other possible similar datasets that may be included later during the project will be implemented only in testing of features that may be relevant (like synthetic data creation) and auxiliary analysis tasks.

## 2.3 DATA SOURCES CHARACTERIZATION

### 2.3.1 INTRODUCTION

Researching the effect of air pollutants on people's health requires datasets from various knowledge domains, including climatic, atmospheric, demographic and health data. The best datasets for each domain must be identified based on the specifications of the model and the research questions.

Datasets must be carefully selected to avoid wasting efforts in collecting useless or insufficient data for the intended models. Alignment with medical partners and data providers have been an essential step to avoid this issue, having proposed pre-conceptual data structures and variables to feed the model in later tasks, while avoiding a heavy duty of data cleaning and lessen the preprocessing steps needed.

Datasets may be inadequate if they differ from the operational level of the model, or if the data availability is not enough to keep up with the observations over time needed to gain relevant insights.

Obtaining data from external sources has also presented a challenge, as treatment information and health records are treated as confidential information and therefore it is not usually available as open-source data. In the other hand, climatic and pollution information is widely available, but relating that information to health records is often unfeasible for the reason beforementioned.

Another problem to tackle is the number of entries in the health dataset. Health information regarding asthma and rhinitis allergies are heavily seasonal and changes can't be detected in short ranges of time. Medical centres gather new data once a year for each patient, with a total of 2000 entries per year.

This shortage of data highlights the importance of two key aspects: i) **Data quality** and ii) need for **external sources of data** to enrich the analysis and modelling tasks. Furthermore, these constraints forbid us of using large-scale models, which are likely to require substantial amounts of data.

**Data characterization** processes serve as the first step on data quality checks and integration of data. It is essential in any project involving datasets from different sources and of a different nature, as it reveals inconsistencies in data, its structure or format. It's a critical part of data analysis, data governance and data preparation.

The result of these processes is an improvement in data quality and a final dataset collection which is ready to perform analysis and further training of AI models.

### 2.3.2 DATA CHARACTERISATION

Data characterisation processes describe the main features of a dataset to improve the understanding of the data structure before analysing it more deeply. It usually includes data type and structure identification, an initial statistical study to understand the data and some data quality measures. The following tables show the variables included after aligning with medical partners and external data providers.



Table 1 shows the demographic data taken into account during this project. This set of variables express the routines of patients and is later expected to exhibit how different habits can make a difference on asthma and rhinitis comorbidities and symptoms. Furthermore, location can be utilised to make a deeper analysis along with climatic information coming from external sources.

VARIABLE NAME	CATEGORY	DATA TYPE	ALLOWED VALUES
<b>Patient ID</b>	Other	String	-
<b>Gender</b>	Demographic	Qualitative	Male / Female / Other
<b>Date of Birth</b>	Demographic	Date	-
<b>Occupational Environment</b>	Demographic	Qualitative	20 options based on work environment (factory, indoor, rural, etc)
<b>Address</b>	Demographic	String	-
<b>City</b>	Demographic	Qualitative	-
<b>Country</b>	Demographic	Qualitative	-
<b>Workplace address</b>	Demographic	Qualitative	Full time remote / Housewife / Unemployed
<b>Remote work frequency</b>	Demographic	Qualitative	Never or number of days per week
<b>Distance to workplace</b>	Demographic	Quantitative	Number of kilometres
<b>Type of commute</b>	Demographic	Qualitative	Private car / Motorbike / Bicycle / Public transport / Walking / Full time remote or housewife or unemployed
<b>Commute duration</b>	Demographic	Quantitative	Number of minutes
<b>Pet keeping habit</b>	Demographic	Qualitative	No / Pet species
<b>Smoking habit</b>	Demographic	Qualitative	Never / Current / Former / Passive exposure
<b>Alcohol consumption</b>	Demographic	Qualitative	Approximate frequency per week or month (5 options)
<b>Physical activity</b>	Demographic	Qualitative	Approximate frequency per week or month (5 options)
<b>PA environment</b>	Demographic	Qualitative	Outdoor / Indoor / None
<b>Consumption of food not cooked at home</b>	Demographic	Qualitative	Approximate frequency per week or month (5 options)



<b>Family history of allergy</b>	Demographic	Qualitative	Yes / No
----------------------------------	-------------	-------------	----------

Table 2 lists all the variables collected by the ClimAir web form tool regarding relevant medical information, patient symptoms and treatment details.

<b>VARIABLE NAME</b>	<b>CATEGORY</b>	<b>DATA TYPE</b>	<b>ALLOWED VALUES</b>
<b>Duration Type</b>	Rhinitis severity	Qualitative	Intermittent / Persistent
<b>Seasonality</b>	Rhinitis severity	Qualitative	Seasonal / Perennial
<b>Severity</b>	Rhinitis severity	Qualitative	Mild / Moderate / Severe
<b>Sleep disturbance</b>	Rhinitis severity	Qualitative	None / Mild / Moderate / Severe
<b>Impairment of daily activities</b>	Rhinitis severity	Qualitative	None / Mild / Moderate / Severe
<b>Impact on school/work</b>	Rhinitis severity	Qualitative	None / Mild / Moderate / Severe
<b>Perceived symptom burden</b>	Rhinitis severity	Qualitative	None / Mild / Moderate / Severe
<b>Comorbidities</b>	Comorbidities	Qualitative	None / Conjunctivitis / Asthma symptoms / Atopic dermatitis / Food allergy / Eosinophilic esophagitis
<b>Symptom</b>	Symptoms	Qualitative	Sneezing, itching, cough, etc (15 options)
<b>Seasonal Pattern</b>	Symptoms	Qualitative	Perennial / Seasonal
<b>Frequency</b>	Symptoms	Qualitative	< 4 days a week / >= 4 days a week
<b>Aggravation location</b>	Symptoms	Qualitative	-
<b>Morning worsening</b>	Symptoms	Qualitative	Yes / No
<b>Triggers</b>	Symptoms	Qualitative	Pollen, house dust, etc (13 options)
<b>Type</b>	Treatment	Qualitative	Oral H1 antihistamines / Intranasal corticosteroid / Intranasal H1 antihistamine / combination / Intranasal alfa 1 agonists
<b>Dose</b>	Treatment	Quantitative	Number of tablets/puffs



<b>Frequency</b>	Treatment	Qualitative	Daily / Weekly
<b>Oral Montelukast</b>	Treatment	Qualitative	Yes / No
<b>Allergen Immunotherapy</b>	Treatment	Qualitative	Yes / No
<b>Year of Initiation</b>	Treatment	Date	dd/mm/yyyy
<b>Year of discontinuation</b>	Treatment	Date	dd/mm/yyyy
<b>Composition</b>	Treatment	Qualitative	Pollen / house dust mite / fungi / animal dander
<b>Route</b>	Treatment	Qualitative	Subcutaneous / sublingual drops / sublingual tablets

Table 3 shows the data coming from SILAM model, provided by FMI through different sources (EPR, GLORIA and general Air Quality datasets). This dataset contains the main information for climatic and pollution variables and will be later used to analyse the impact of micropollutants and environmental factors on patients' health, as well as being part of the input of the federated models.

The next variables are shown in different units and have separated ways to be shown, depending on the way these are measured. For the data coming from GLORIA dataset, we can differentiate between i) CNC, Concentration in air ( $\text{mole m}^{-3}$ ); ii) DD, dry deposition ( $\text{mole m}^{-2} \text{sec}^{-1}$ ); iii) WD, Wet deposition ( $\text{mole m}^{-2} \text{sec}^{-1}$ ); iv) AOD, relative unit; although all exposure indicators are available for each variable. For example, air pollutants are usually measured as CNC values. Meanwhile, pollen data used during the project are measured as CNC and DD, depending on availability.

VARIABLE NAME	CATEGORY	DATA TYPE	DATA UNITS
<b>SO2</b>	Pollution	Quantitative	CNC, DD, WD
<b>NO2</b>	Pollution	Quantitative	CNC, DD, WD
<b>O3</b>	Pollution	Quantitative	CNC, DD, WD
<b>PM2.5</b>	Pollution	Quantitative	CNC, DD, WD
<b>PM10</b>	Pollution	Quantitative	CNC, DD, WD
<b>Pollen (alder, birch, grass, hazel)</b>	Pollution	Quantitative	CNC, DD, WD
<b>CO</b>	Pollution	Quantitative	CNC, DD, WD
<b>Wind-blown dust</b>	Pollution	Quantitative	CNC, DD, WD
<b>Temperature</b>	Climatic	Quantitative	°C
<b>Humidity</b>	Climatic	Quantitative	$\text{g/m}^3$



<b>Wind speed</b>	Climatic	Quantitative	m/s
-------------------	----------	--------------	-----

It is important to clarify that meteorological variables are often retrieved from the pollen datasets, as it already provides the required temporal alignment. The system falls back to the air quality dataset only if meteorological parameters are missing.

An initial statistical analysis on this data will be carried out in the next deliverables, as this document aims to lay the ground for future work. Data quality measures are explained next as a part of the data governance plan and in Section 2.4.

### 2.3.3 DATA GOVERNANCE

Patients retain control over their clinical data and provide informed consent for medical centers and Mask Air to manage it according to the agreed terms of service. ClimAir clinical centers will utilize patient health data solely for the duration of the project. Upon project completion, this data will be destroyed, unless national regulations prevent it.

The following structure serves as an initial approach, presented under the frame of use of FL techniques. The biggest concern with collaborative training in healthcare is data privacy concerns, which limit data sharing and the clinical implementation of what is technically possible, resulting in an increasing focus on privacy-preserving approaches such as FL<sup>2</sup>. The main advantage of FL is the generation of a higher-quality model by leveraging larger training datasets obtained from multiple sources, beyond what could have been achieved with the data of a single device or a system, while maintaining high levels of data privacy. In our context, the use of FL ensures data integrity and usability, while minimizing data aggregation costs.

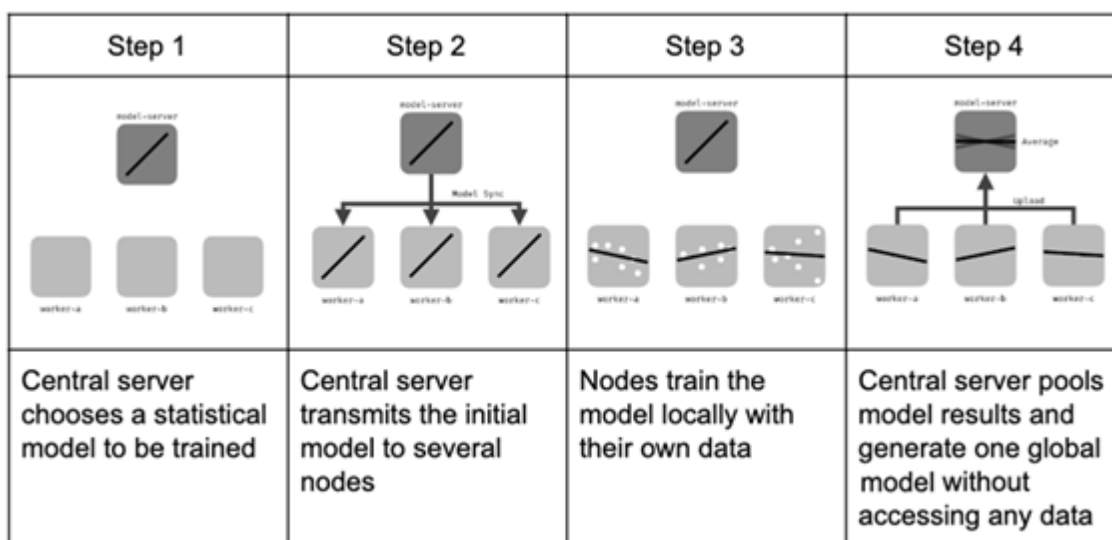


Figure 1. Federated Learning working diagram.<sup>3</sup>

FL allows a shared machine learning model to be trained locally at each data source, fundamentally separating the computation steps from data movement. It naturally reduces data exposure on multi-source, multi-type data projects where sensible data is



<sup>2</sup> PubMed Central: **Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture**

<sup>3</sup> Wikipedia – Federated Learning - [https://en.wikipedia.org/wiki/Federated\\_learning](https://en.wikipedia.org/wiki/Federated_learning)

used, as only the local models trained are shared between systems, keeping the data local to each medical partner's system. With this approach, we only share the aggregated model updates instead of the raw (or anonymized) data itself, enabling collaborative learning from the data coming from all sources. This way, data remains unavailable for external sources while making the most use of each record.

Complementary measures have been studied to avoid further risks of data leaking, restricting access to databases and creating a standard process for clinicians to input patient records, keeping a simple and secure way to collect new data and update the information in future visits. To create the ClimAir medical dataset, clinicians will have access to an online platform showing a form with fields to fill in with the patient's demographic, allergic and treatment information. The fields in this form are the same for each medical partner and fixed values are given to the clinicians to select, further improving data homogeneity. Access to local nodes is only available to the medical partners in charge of each of the nodes and system administrators who can remotely access the nodes to upload updates like database scheme changes, inclusion of new variables and other maintenance tasks.

External data can be included to the local and global models as explained in D4.1 making use of programmatic access to APIs and the flexibility of the OMOP Common Data Model (CDM) schema, which tackles most of the operational issues that arise from the use of distributed and heterogeneous data. External actors who treat with sensitive data (in our case, SILAM and MaskAir) also need to adhere to these principles, minimizing the handling of raw data to the least possible amount of people and systems. Information coming from SILAM is simpler to handle, as most of the data retrieve is information about climatic and pollution factors that affect each patient record. The geolocation information of the patient is only used to access the proper climatic and pollution data and is later anonymized or ignored.

In the case of MaskAir, the only people who have access to the uncoded data are i) system administrators, for system and technical support; ii) data protection officers (DPOs) from participating institutions for regulatory compliance. Using a role-based access system, they ensure that no sensitive data is shared with third parties or external researchers. For further security, all the information will be saved either in encrypted servers following EU-compliant cloud structure or in anonymized datasets shown to authorized researchers, following ISO 27001 information security standards. In practice, only researchers and data analyst will have access to the sensitive data in order to realise statistical and outcome analysis. The presence of external auditors can be needed if it is required by ethics committees or funding agencies.

## 2.4 DATA QUALITY ASSESSMENT

Data quality checks are critical to any data-based project which involves the adherence of ML pipelines and the development of AI models. This becomes ever more important with the implementation of a FL framework for both data collection and model training stages, making data quality an essential step during the development of the project.

Defining these procedures at an early stage of the project ensures that latter data processing steps will be minimized and makes for an improvement during the development, training and testing of the local and global models, **avoiding problems**



**like data formatting issues, missing data or the use of different scales for similar data values.**

It is essential to **define a scheme to follow for all the partners** working in data related tasks. This scheme will need to take into account the differences between data origins and how each of these will be integrated in the general workflow defined.

Data collected in different origins will naturally present small differences on the quality of the data. To make sure that quality requirements are met for each dataset, there will be slight changes on the procedures of quality control and data collection for each origin.

- **ClimAir medical partners** enter health records through each local node's online platform created for this project, the **ClimAir Online Form Platform**, which ensures a **high** level of **alignment** within the different sources the data is collected from. Nonetheless, certain variables are filled and need to be checked in order to remove empty or useless values.
- **MaskAir datasets** will be pushed to the pipeline using a secured standard API communication protocol request through ClimAir local nodes. This data shows a high similarity with the ClimAir Online Form Platform datasets collected by the medical partners but shows three key differences: i) The frequency of data collection is different; ii) the options shown by the MaskAir application to fill in the variables are not aligned in some cases with the options of the clinicians; iii) geolocation data is only available if patients agree on it, so some records are expected to have missing data in that regard.

The integration of MaskAir data heavily depends on users of the mobile application. In order to minimize missing data and errors from the beginning, clinicians a user guide to explain each person how to download and use the MaskAir app. Medical staff should guide each patient to enter the study code necessary to participate in the ClimAir study. Clinicians are encouraged to guide each patient and help them fill in the first day questionnaire.

- **SILAM model produce great quality climatic and pollution data.** This fact is backed up by the **peer-reviewing process** that the model has gone through **and regular evaluations** against observations and established systems that compare the model output with measurement data like **dust events via WMO dust warning systems**. These tests ensure that the **simulations data** produced by SILAM is **reliable and robust**, being **validated in real world applications**. The datasets produced by the model present a high spatial resolution and a great completeness, covering many pollutants and climatic conditions over a 20km global grid, offering fine-grained pictures on the regions of analysis.
- **FaD datasets** are currently being collected through communication with local authorities and checking satellite data and undergoing preprocessing procedures to stablish a high level of quality. These datasets depend on the LES-SILAM high-resolution simulations and fusion-based downscaling to produce quality datasets and models (Task 7.2 and 7.3 respectively) which can be further analysed and will be integrated into the IDSS later in the project.

With these differences in mind, the following procedures will be implemented when needed.



1. Data profiling and validation: Examining data structure and patterns, applying rules to make sure that the information gathered by every source of information follows the characterization scheme defined.
2. Data cleaning: Basic filtering of data through measures like missing data control, removing duplicate information and fixing possible formatting issues. The use of the ClimAir Platform ensures the minimization of missing data as well as homogenization. Nonetheless, we expect at most a completeness of 95% for most variables, leaving a 5% missing rate for future problems that may arise during data collection phases. Medical partners have been asked during the dataset definition step to ensure the minimization of edge cases and missing values due to inconsistencies in the predefined values of the ClimAir platform.
3. Data standardization: Homogenize naming in datasets and improve consistency across records. This step is mainly solved by early adopting a consistent common data scheme.
4. Data Matching: Datasets from different sources will be joined using a patient ID created after anonymizing the information of the patient. This ID is created right after a patient first visits the clinician's office and has its first check-up, creating a pseudonymized identifier for data processing and data linkage in next procedures.

This procedure maximizes data quality, ensuring that the final dataset is resilient and accurate and making it usable for the following model training tasks.

### 3. GAP ANALYSIS

When data is properly aligned, meaning it is cleaned, standardized, and mapped to a common structure, AI systems can accurately identify patterns, integrate information, and make robust predictions. This alignment reduces bias and errors while improving both the performance and trustworthiness of the final AI solution.

The ClimAir project is being developed under a FL framework, meaning that this alignment is ever so more important than in other projects where data can be centralized and postprocessed at a single centre, ensuring consistent formats and definitions.

This framework is essential for the treatment of critical information and will improve the reliability and effectiveness of the models trained during the life of this project. But for this to happen, we need to make sure that all the heterogeneous data sources send their data in a homogeneous structure and format.

At this stage of the project, the collection of clinical data has not yet started, but in order to align with this principle, different partners have held a substantial number of meetings to understand the formats needed on data and the undergoing processes. The concept developed is backed up by the underlying architecture of the tools developed to collect data, which ensures the minimization of data formatting errors and gives a clean definition to all partners participating in the collection and processing of data, ensuring future quality in datasets collected and during AI training phases.

To this day, the only gaps found in data come from the creation of the different FaD datasets which are being collected to support future tasks for this project. These datasets are the output of task 7.1, which results in 9 different datasets, one for each of the following cities: Berlin, Malaga, Milan, Toulouse, Luxembourg, Thessaloniki, Brasov, Lodz and Chernivtsi.

These datasets include data based on spatial and temporal resolution, temporal density, population data, maps of vegetation and air pollution and aerobiological observations. Most of the information is available and has been collected properly for every city. If data is not available through local sources, it is expected to substitute it with satellite data. Only a few problems remain.

- **Temporal density:** street and road maps with traffic intensity, diurnal, weekly and seasonal variations are currently unavailable for **Milan, Brasov, Lodz and Chernivtsi**.
- No population data or air pollution and aerobiological observation for **Chernivtsi** have been identified so far.

While a common solution is being developed for missing temporal density information (affecting 4 datasets), **Chernivtsi** data availability is being studied as the **most difficult case** to resolve, as it is the only city lacking information in several areas. Missing information is currently being handled by **RSS** and should be coming in the **next few months**. As **data collection** steps have still to be initiated in **M18** by **medical partners**, we **expect** this issue to have a **low impact** during the project development.

## 4. INTEGRATION STRATEGY

The collaborative work of different frameworks and tools is not enough to ensure the preservation of data privacy across multiple organisations. An integration plan is needed to further details how these tools will be used, explaining how the data will impact the federated models without ever sharing raw data.

Data collection of clinical variables happens in medical centres at first, and then patients are given access to the MaskAir platform to keep track of their symptoms (via survey) before, during and after the pollen season in a daily basis. Climatic and pollution data is collected and processed by FMI through the SILAM model. The records coming from MaskAir or the pollution and climatic datasets are integrated using the architecture established by KEY through an API. The procedure of capturing a patient's data and linking all these variables could be summarized in the following steps.

1. Patient is registered for the first time at a certain medical centre within the consortium, and a Patient ID is automatically assigned to that person. This identifier is the primary key used for all clinical and environmental associations throughout the study.
2. Work and workplace addresses are obtained and converted into geographical coordinates.
3. The geographical coordinates of the patient are then securely transmitted from the local node to the central ClimAir server to perform environmental matching.
4. Then, SILAM data is dynamically retrieved via API based on time-geolocation pairs, building a clean and harmonized dataset for the patient.
5. After that, an enriched dataset without any clinical information is securely delivered to the hospital's local node, which ensures that the local nodes train models using standardized environmental datasets.
6. At the local nodes, the clinical and atmospheric data is combined to become the input for the local training workflow.

During this process, sensitive data is pseudo anonymized to lessen the risk of data filtering, limiting access to raw data and helping to keep the data governance structure.

Figure 2 shows the procedure taken after a patient information has been collected, showing the workflow and data treatment procedures between the patients first visit to a clinic, the retrieval of auxiliary information like climatic and pollutant data, the training on local nodes and the aggregation of the models into the global model on the central server.

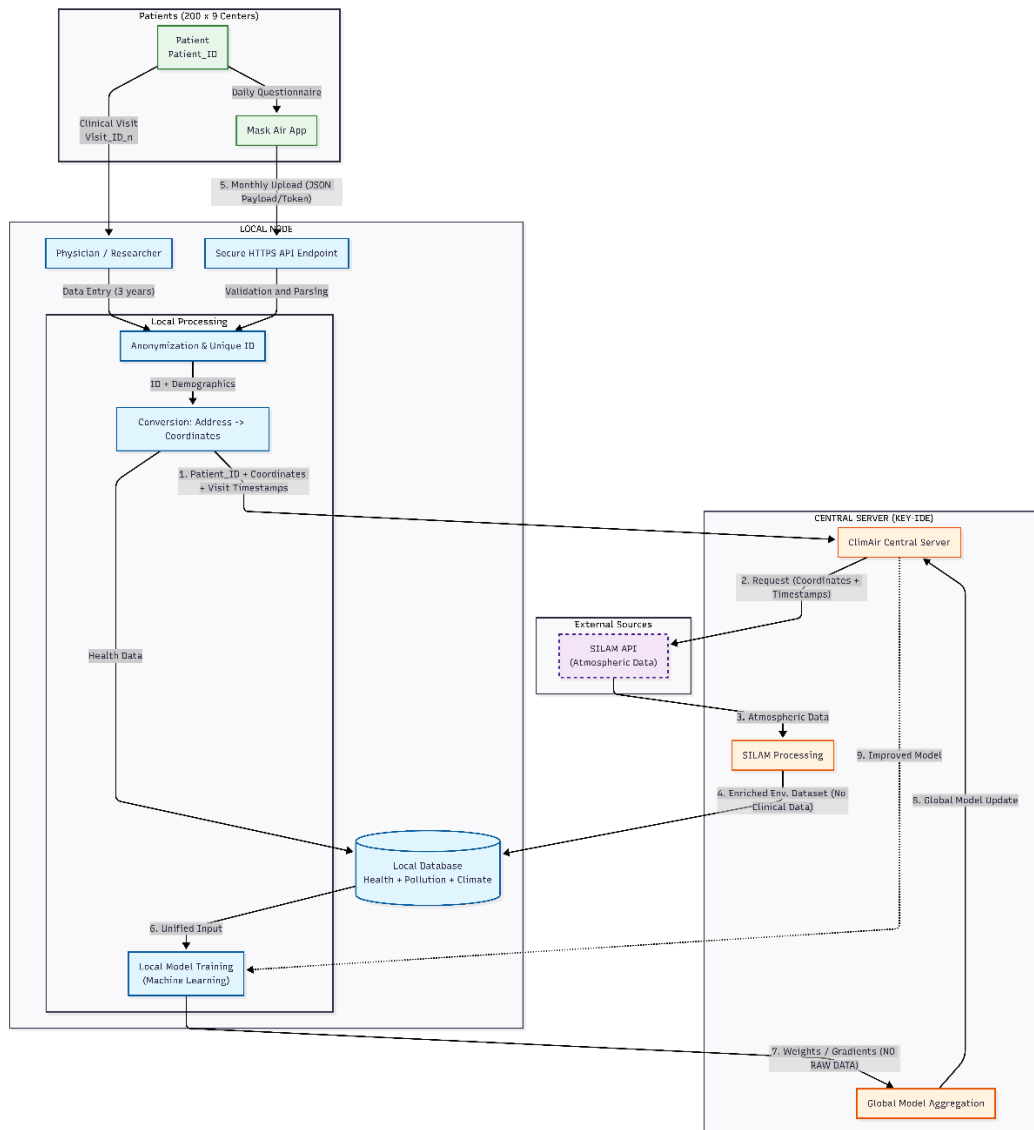


Figure 2. Patient information flow after first visit. From collection to local training and global model aggregation. A Patient ID is created when the patient is first registered in the system, and a Visit ID is created for every visit of the visit to the clinic. The visit timestamp is later used to retrieve environmental exposure data and match that data with the patient record.

Patient's data is first collected by a clinician and stored in the local nodes. The patients selected must have an annual visit for the next three years. After this data is stored in the local nodes, the processing steps explained before are applied to generate the data containing climatic and pollution data along with the clinical information for that patient, which is stored in the local database.

The local database contains the information that will be used to train each local model. By following the characterization and quality control procedures, we ensure that information across organisations is homogeneous, so all the local models can be trained in a similar fashion.

After these models are trained, the gradient aggregation step happens, where we use the weights for each model to build a larger general model (global model), which is stored and feeds on data located in the ClimAir central server.



Following this flow, we find that each origin has its own local model and proprietary datasets with the clinical, climatic and pollution information regarding the patients for each partner. Also, the global model is built thanks to every local node input, applying FL techniques.

The methodologies defined in this document will serve as a guide for Task 4.3, which focuses on the implementation of these procedures, the first stage of data collection, the harmonization of datasets and its curation. The defined workflows also ensure data privacy and security for the federated models which will be developed during Task 4.4, facilitating collaborative AI training with quality data.

The curated datasets that result from this process, along with the initial models developed during the implementation of these procedures and first training phases will serve as input for tasks in WP9 and WP10.

## 5. CONCLUSION

This deliverable demonstrates the identification process carried out for health, climatic and pollution datasets internal and external to ClimAir in the context of building a FL framework approach for effective and privacy-preserving machine learning correlation establishment.

When large-scale or complex models require the use of multiple data sources, it is even more important to employ proper characterisation methodologies for efficient data management. This is especially true when ensuring that researchers have access to the necessary model-operating variables.

During Section 2, an inventory of the available datasets during the project shows the structure of the data that will be used to train the local and global models, specifying the necessary procedures to align the data types, formats and values throughout all the origins.

This section also includes the characteristics structure of the data along with the quality measures described ensures the minimization of error during training while improving performance of future model training workflows. These procedures also help keeping anonymity on sensitive data and reducing the actors handling both sensitive and non-sensitive data.

Almost no gaps have been found in the sources of data, only having a few issues with the creation of FaD datasets. Data gaps in these datasets are already being fixed, having contacted local authorities for the issues and the cities mentioned in Section 3 and having backup strategies like the use of satellite data.

The concrete workflow to collect and manipulate patients data has also been mentioned Section 4, where the integration strategy has been explained, including external information through the use of APIs and explaining the whole process starting with the first visit of the patient to the clinic until the training of the local models and the integration of the results of the training steps into the global model using innovating FL techniques.

The results carried out by WP4 will feed and strengthen the outcomes of later developments of the project, especially in Task of WP9 and WP10, which input heavily relies in the methodology defined in this document. The definition of well-structured and characterised dataset will ease the development of models happening during Tasks 9.1 and 10.1 minimizing biases and ensuring model accuracy and robustness.