Horizon HLTH 2022 Project MISTRAL
"A toolkit for dynaMic health Impact analysiS to predicT disability-Related costs in the Aging population based on three case studies of steeL-industry exposed areas in Europe".

Research and Innovation Action
Topic: HORIZON-HLTH-2022-ENVHLTH-04-01
GA n. 101095119

Duration: 48 months from 01/01/2023
Coordinator: ISTITUTO SUPERIORE DI SANITÀ

Deliverable ID.:	**2.1**	
Deliverable title:	**Data Management Plan**	
Planned delivery date:	30/06/2023	
Actual delivery date:	30/06/2023	
Deliverable leader:	APS Public Health Environment and Social Equity (PLANET – 2)	
Contributing partners:	Polytechnic University of Bari (POLIBA – 9)	
Dissemination Level:	PU	PU = Public;
		CO = Confidential
		Cl = Classified

This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101095119.

This deliverable reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

Document information and history

Deliverable description (from DoA)
Development of data management plan

Please refer to the Project Quality Handbook for guidance on the review process and the release numbering scheme to be used in the project.

Version N.	Date	Author [Person and Organisation]	Reviewer [Person and Organisation]	Milestone*	Notes
01	20/06/2023	Ilaria Bortone (PLANET)	Angela Lombardi (POLIBA)	TOC	
02	27/06/2023	Ilaria Bortone (PLANET)	Giuseppe Campanile (PLANET)	Proposed	
03	28/06/2023	Ilaria Bortone (PLANET)	Angela Lombardi (POLIBA) Domenico Lofù (POLIBA)	Revised	

** The project uses a multi-stage internal review and release process, with defined milestones. Milestone names include abbreviations/terms as follows:*

- *TOC = "Table of Contents" (describes planned contents of different sections);*
- *Intermediate: Document is approximately 50% complete – review checkpoint;*
- *ER = "External Release" (i.e. to commission and reviewers);*
- *Proposed: document authors submit for internal review;*
- *Revised: document authors produce new version in response to internal reviewer comments approved: Internal project reviewers accept the document.*

Table of Contents

1 Introduction 3
2 Data Summary 4
3 FAIR data 7
3.1 Making data findable, including provisions for metadata 7
3.2 Making data accessible 8
3.3 Making data interoperable 9
3.4 Increase data re-use 10
4 Allocation of resources 11
5 Data security 14
6 Ethics 16
7 Other issues 17
Annex 1. Informed Consent 18

List of Tables

Table 1. Zephyr Study: cross-sectional multi-center observational study to siZe dEterminants associated with black carbon concentration in biological and Predictors of quality of life in different subpopulations distinguisHed bY age group and Residence (Primary Data Collection) 4

Table 2. Tramontana Study: multi-center reTRospective study to select the determinAnts in the dose-effect function, and to forecast Models for QALY, DALY and pOllutants contaminatioN through deTerministic and stochAstic methods from 30.000 citizeNs distributed in the three europeAn countries (Secondary Data Collection) 5

1 Introduction

The MISTRAL project aims to develop a technological toolkit for dynamic, intelligent HIA toolkit to predict the health impact of health-related features, forecasting the trajectories of disability and quality of life reduction. This method will use environmental, socio-economic, geographical, and clinical characteristics, managed, and elaborated with a federated learning architecture. The generated models will be adjusted for lifestyle and individual conditions data sourced from large population-based digital surveys. The models will be trained and validated on three different exposures to the steel plants' pollution: Taranto in southern Italy, Rybnik in Poland, and Flanders in Belgium.

This document is the MISTRAL data management plan (DMP). The aim of the MISTRAL Data Management Plan (DMP) is to identify the project's research data and to describe how to make them findable, accessible, interoperable, and re-usable (FAIR). The DMP describes the data management life cycle for all datasets to be collected, processed and/or generated by the research project. The HE DMP describes, among others:

- the handling of research data during and after the project
- the type of data that will be collected, processed, or gathered
- what methodology and standards will be applied
- whether and how the data will be made (openly) accessible
- how the data is stored

Following the HE DMP Template available on the Participant Portal, all partners involved in the research's activities related to data collection were asked to provide detailed information about the data generated during the entire project as reported in 1. Data Summary.

In the Deliverable, we report then an initial analysis of how we intend to manage the amount of data produced in the project them. The first checkpoint of the whole architecture of the DMP is the release of the first scientific/technological publications that will be published within the MISTRAL project: indeed, the data reported in the article will be made available and interoperable to the larger typologies of stakeholders. To avoid issues related to IP rights and their access, as a first step in the strategy of development of DMP only data related to publications available to the public will be released. In MISTRAL Project DMP is intended to be a living document in which information can be made available on a finer level of accuracy and details through updates as the implementation of the project progresses and when significant changes occur.

Further updates on Data Management will be provided in the second reporting period. Indeed, the DMP is intended to be a living document in which information can be made available on a finer level of granularity through updates as the implementation of the project progresses and when significant changes occur.

2 Data Summary

- *Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.*

Yes, it is encouraged to make existing data available for research within MISTRAL. WP3 will provide data templates and lists of pre-defined values, to be able to harmonize the different datasets that are provided.

- *What types and formats of data will the project generate or re-use?*

Hereafter we report the tables where the information about the contents and the collection/generation of the data provided by MISTRAL's Studies are reported. At this early stage of development of the project, the reported information will be processed to determine the general specifications and structures of the metadata that will be generated within the DMP.

Table 1. Zephyr Study: cross-sectional multi-center observational study to siZe dEterminants associated with black carbon concentration in biological and Predictors of quality of life in different subpopulations distinguisHed bY age group and Residence (Primary Data Collection)

#	Data Description	Data Type
1	socio-demographic context, lifestyle (e.g. dietary habits and physical activity), medical history and chronic diseases, functional limitations and dependencies	General and Specific Health Anamnestic Assessment
2	Framingham Study Birth History Questionnaire	Pre-Natal and Peri-Natal Health assessment
3	socio-administrative data related to unemployment, illiteracy, housing, and poverty	Social and Wealth Assessment
4	Urine and Blood Sample	Measurements in biological matrices/tissues
5	Food Frequency Questionnaire (FFQ), (IPAQ), short form International Physical Activity Questionnaire (IPAQ-SF), Physical Activity Questionnaire for Older Children (PAQ-C)	Nutrition and Lifestyle
6	willingness to pay (WTP), willingness to accept (WTA)	Environmental\Occupational Questionnaire
7	Estimates of ambient exposure (black carbon, nitrogen dioxide [NO2], and PM ≤2.5 µm [PM2.5]) will be constructed based on the residential addresses of the individuals, using a spatial-temporal interpolation method	Residential exposure estimates
8	Data on dust particles, toxic & odor gases, particulate matter (PM1, PM2.5, PM10), Sulfur oxide (SOx), Nitrogen oxide (NOx), Carbon dioxide (CO2), Carbon monoxide (CO), Wildfire Flux of Formaldehyde (CH2O), TVOCs, cloud diffraction, along with meteorological parameters like wind speed and direction, relative humidity, ambient temperature, pressure, light intensity, rainfall, UV Radiations and more …)	Copernicus Atmosphere Monitoring Service Air Pollution Analysis

#	Data Description	Data Type
9	Climate Data	Copernicus Atmosphere Monitoring Service Air Pollution Analysis
10	Data acquired from the sensors will analyse dust particles, toxic & odour gases, particulate matter	Indoor Air Sensors (individual)

Table 2. Tramontana Study: multi-center reTRospective study to select the determinAnts in the dose-effect function, and to forecast Models for QALY, DALY and pOllutants contaminatioN through deTerministic and stochAstic methods from 30.000 citizeNs distributed in the three europeAn countries (Secondary Data Collection)

#	Data Description	Data Type
1	Electronic Health Records (EHRs): anthropometric data, smoking habit, and dietary habits (included the derived quantity of foods), vaccination coverage, functional limitations, substances addiction, diagnosis, and mortality	Health indicators
2	Social deprivation indicators, education levels, labour force and employment, health care services and resources, poverty. Family\Addresses Wealth Data: family component, remote family medical history, declared prevention	Socio-economic information
3	Estimates of ambient exposure (black carbon, nitrogen dioxide [NO2], and PM ≤2.5 µm [PM2.5])	Residential exposure
4	Analysis of physical and financial infrastructures, the structure, and dynamics of the real estate market (hedonic price model), and the analysis of urban heritage. Analysis of the real estate market. Analysis of the different models of urban regeneration, observing the main paradigms (financial - technological - social - architectural) and outlining the main challenges related to sustainability.	Demo-anthropologic Urban and Real Estate

Formats of data:

- Data and metadata will be requested, stored and transferred (across partners and in Copernicus) in a comma-separated values (CSV) format.
- MS Excel-compatible files including comma-separated and .xls(x) format will also be accepted to facilitate the data exchange.
- For statistical purposes, other formats include .JSON, .RData (R), .mat (matlab).
- Where applicable data formats may be migrated when new technologies become available and are proven robust enough to ensure digital continuity and continued availability of data.

- *What is the purpose of the data generation or re-use and its relation to the objectives of the project?*

MISTRAL aims at developing a technological toolkit for dynamic, intelligent HIA toolkit to predict the health impact of health-related features, forecasting the trajectories of disability and quality of life reduction. This method will use environmental, socio-economic, geographical, and clinical characteristics, managed and elaborated with a federated learning architecture. The generated models will be adjusted for lifestyle and individual conditions data sourced from large population-based digital surveys. The models will be trained and validated on three different exposures to the steel plants' pollution: Taranto in southern Italy, Rybnik in Poland, and Flanders in Belgium. Data will be then generated from cross-sectional cohort studies (Zephyr Study) in the three settings and from retrospective data available in the three countries (Tramontana Study). Thus, MISTRAL's studies aim at promoting a framework that collects data on human environmental exposures as well as human biological samples that will be collected in a biobank and may serve for future investigations of time and lifestyle exposure trends as well as early warning markers of exposure and disease.

- *What is the expected size of the data that you intend to generate or re-use?*

To be evaluated during the project. The expected size depends on the extent and the nature of the data that are made available.

- *What is the origin/provenance of the data, either generated or re-used?*

The MISTRAL Project will both generate and re-use data. Zephyr and Tramontana Study will be the source of information. The Zephyr Study is a multi-center cross-sectional cohort study running in Taranto in Italy, Gent/Hasselt in Belgium and Rybnik in Poland aiming at collecting primary data on lifestyle, socioeconomic status, environmental exposure and quantitative measurements of pollutants in biological matrices. This study is essential to identify QALY and black-carbon contamination using census data we will provide a score using an ensemble learning approach to predict the probability of the two outcomes starting from a set of features assessed directly in the cross-sectional survey (Zephyr Study).

The Tramontana study is essential to select the determinants in the dose-effect function. In particular, the QALY and BC scores from Zephyr Study will be trained and tested in the longitudinal cohort (Tramontana Study) using the secondary data features classified as the best predictors in the cross-sectional survey. In addition, starting from the deterministic models, forecasting time series analysis will be implemented to predict stochastic trajectories of the health-related outcomes, QALY, and DALYs. That analysis will be replicated in a third cohort (the Genk-Hasselt) as new calibration and sensitivity analysis to compare the accuracy of the models. The study will also benefit from the existing ENVIRONAGE cohorts and FLEMISH database and biobank in Belgium.

- *To whom might your data be useful ('data utility'), outside your project?*
    - MISTRAL consortium
    - European Commission services and European Agencies
    - EU National Bodies
    - The general public including the broader scientific community.

3 FAIR data

3.1 Making data findable, including provisions for metadata

- *Will data be identified by a persistent identifier?*
- *Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.*
- *Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?*
- *Will metadata be offered in such a way that it can be harvested and indexed?*

Data Description

The description of procedures to generate data is associated with a dataset (i.e., collection of data).

At this stage of development of the project, the specific typology and total number of variables in a single dataset table (see Data Summary) cannot be defined as a priori.

The procedures for the identification of data are defined as follows:

- Each dataset is initially assigned to a unique ID, automatically generated through a Universally Unique Identifier (UUID) application.
- Each dataset is also associated with a Digital Object Identifier (DOI). The service is provided by the DOI (www.doi.org) community through a request to a local Registration Agency (RA).

The use of a DOI guarantees, at the same time, unique identification of the single dataset and the possibility of automatic data web retrieval.

After this step, the dataset is univocally associated with an identifier.

The implementation of the data description depends on the typology of the datum considered.

In most cases, a text description is appropriate. In this case, data are described by compiling a form (data description template), available to all users.

Moreover, the search keywords are aimed to enhance discoverability. Their aim would be to describe the content of the data, context, and purpose. Based on the broad themes presented across the data, here are seven key keywords that could optimize discovery and potential re-use, socio-demographic, medical history, environmental exposure, nutrition and Lifestyle, biological Sample Measurements, and air pollution analysis.

Metadata

One or more metadata files are generated for each dataset. The metadata are identified by the same unique ID of the related dataset, with a different suffix/extension.

Each metadata file is uploaded in a standardized format, depending on the dataset considered.

Appropriate templates will be available for download to all MISTRAL's partners in the Virtual Infrastructure.

The metadata files are linked to the data descriptor, and directly accessible through a web link.

The link is realized through the definition of an URL address, related to the metadata ID, which is also associated with a web resource on the cloud correlated to the MISTRAL Virtual Infrastructure.

Throughout the project, we will use a standardized approach to make it accessible so it could be easily used by other systems, e.g., in a searchable index. We use the common protocols used for metadata from digital resources.

3.2 Making data accessible

Repository:

- *Will the data be deposited in a trusted repository?*
- *Have you explored appropriate arrangements with the identified repository where your data will be deposited?*
- *Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?*

Data:

- *Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.*
- *If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.*
- *Will the data be accessible through a free and standardized access protocol?*
- *If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?*
- *How will the identity of the person accessing the data be ascertained?*
- *Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?*

Metadata:

- *Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?*
- *How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?*
- *Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open-source code)?*

By default, and as a first case study of data management, only data related to publications will be made openly available In general, the Steering Committee will decide on a case-by-case basis which data can be released in order to avoid issues related to IP rights protection or access.

The MISTRAL Virtual Infrastructure will be set up for this project as a distributed infrastructure (Local and Cloud) with modular architecture. The entire software architecture will allow non-destructive read access to implement data integration solutions based on a Global As View logic. The solution will include a configuration adaptable to the different data sources in the federated processing. The federated logic will be adopted using interoperable PaaS for every data frame in each country and a connector or ESB to collect a single IaaS that will be the workstation. The resources will be coordinated using the same virtual machines on different connectors. The federated learning will start from the design of the data collection platform. The automation software and management tools will be added to assign these resources and perform the provisioning. In parallel, the protection architecture will be pervasive and touch all aspects of data processing. Authentication and authorization processes will mediate access to the data. Differential privacy techniques will be adopted to eliminate the possibility of reverse engineering techniques on the models to go back to the original data. Finally, through a Global As View approach, Federated access to different data sources starts from the assumption that it is possible to design an overall data schema that acts as a semantic intermediary toward query interfaces to local data.

In addition to local storage, public metadata and aggregate datasets will be made available to users once publications are finalized through MISTRAL's website and through the Open AIRE sharing web platform. Relevant MISTRAL metadata and aggregated dataset will be uploaded by the involved researchers to the ZENODO (https://zenodo.org/) platform, compiling project-related information.

This will enable automatic data extraction from the Open AIRE platform, thus ensuring accessibility through a standard platform for Open Data access.

3.3 Making data interoperable

- *What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?*
- *In case it is unavoidable that you use uncommon or generate project-specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?*
- *Will your data include qualified references1 to other data (e.g. other data from your project, or datasets from previous research)?*

All data will be made available in standard/open formats compliant with commercial/open software to allow as much as possible data exchange between researchers and institutions. Standard vocabulary for metadata description will be used, in case this will not be possible a mapping of more common ontologies (i.e.,

[1] *A qualified reference is a cross-reference that explains its intent. For example, X is regulator of Y is a much more qualified reference than X is associated with Y, or X see also Y. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data. (Source: https://www.go-fair.org/fair-principles/i3-metadata-include-qualified-references-metadata/)*

diagnostics, module, validation, demonstration…) will be provided. In this case, specific technical contributions from specialists in semantics and logic will be considered.

3.4 Increase data re-use

- *How will you provide the documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?*
- *Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?*
- *Will the data produced in the project be useable by third parties, in particular after the end of the project?*
- *Will the provenance of the data be thoroughly documented using the appropriate standards?*
- *Describe all relevant data quality assurance processes.*
- *Further to the FAIR principles, DMPs should also address research outputs other than data, and should carefully consider aspects related to the allocation of resources, data security and ethical aspects.*

This section will be compiled throughout the course of the project, when we get more information on the datasets that are made available for MISTRAL.

4 Allocation of resources

- *What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.)?*
- *How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)*
- *Who will be responsible for data management in your project?*
- *How will long-term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?*

The POLIBA team (sisinflab@poliba.it), and APS PLANET (info@planet-npo.eu), respectively leaders of WP1 and WP2, are co-managers of the data management, with these specific responsibilities.

The costs for making data FAIR include the costs of the cloud facility and of personnel involved in collecting and managing data:

- Set up the layers in the Virtual Infrastructure
- Implementation of the UUID generator
- DOI registration request
- Preparation of templates for:
    - Data descriptor (general, text format, pdf output)
    - Metadata: text template, spreadsheet template
- Data collection
- Generation of the data descriptor
- Generation of the metadata
- Upload to the private cloud server
- If public, upload data to ZENODO/Open AIRE

The overall cost of DMP according to the reported cost items is considered and covered by the Open Access cost of the project budget (see Annex 2 in Grant Agreement).

The POLIBA is responsible for:

a) The initial set-up of the hardware and software components of the Virtual Infrastructure
b) Maintenance of the hardware and software components of the Virtual Infrastructure
c) Carrying out the initial Security assessment of the Virtual Infrastructure
d) Perform Security Assessment on a regular basis (e.g. one year) to guarantee the agreed security level
e) Reporting and blocking any possible security threat, taking appropriate measures accordingly
f) Creation and management of the internal User Group Account database (LDAP), as one of the components of the Virtual Infrastructure

The POLIBA responsibilities include:

- setting up and upgrading, when needed, of the hardware and software components of the Virtual Infrastructure

- creation, maintenance and upgrading of the User Group Account database
- capacity management of hardware and software components

The POLIBA responsibilities do not include the definition of:

- the user groups
- the list of members belonging to one or more user groups
- the access, upload and download rights for each user group

The POLIBA is not responsible for the interruptions of the data services that are due to force majeure.

The POLIBA is not responsible for the content (of data and documents) reported into the local repository, which must be compliant with GDPR Data Policy.

In the event of an ICT Incident, the POLIBA will follow its internal ICT Incident Response procedure, which may entail granting access to the platform to EC Security staff, including the generation of a forensic copy of the server for further analysis; EC Security Service will act in compliance with GDPR.

The POLIBA obligations

The POLIBA shall:

- Not distribute the list of members for each user group, except to ISS (Project Coordinator) and PLANET.
- Inform ISS and PLANET about any scheduled interruptions due to Virtual Infrastructure services upgrading or technical interventions at the POLIBA

And in general terms comply with personal data protection rules (Regulation EC 45/2001)

PLANET is responsible for a-b:

a) Collecting the users request for access to and download of data
b) Preparing, and checking the list of User groups and members for each user group
c) Providing to the POLIBA the list of User groups and members for each user group that will be transferred into the internal User Group Account database (LDAP)

PLANET responsibilities include:

- Ensuring that the list of members is constantly up to date and consistent with the assigned rights
- Providing instructions to the POLIBA about the data repository structure
- Timely communication with the POLIBA of any change in terms of groups (e.g. name, rights) and related membership
- Communication to the user any scheduled interruptions due to data repository services upgrading or technical interventions at the POLIBA.
- Communication to the users of any unplanned interruption of services

The APS PLANET obligations

- Verify that the data and documents uploaded into the repository are following and compliant with what is reported in this GDPR Data Policy
- Timely communicate any possible compliance issue to the Project Coordinator
- Timely communicate any possible in compliance with this GDPR Data Policy to the Data Providers

Assistance to data providers and data users is coordinated via the MISTRAL helpdesk on data management that is accessible via the internal pages of the MISTRAL website. The helpdesk is managed by PLANET for MISTRAL consortium partners. Policymakers can address their questions to info@planet-npo.eu.

5 Data security

- *What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?*
- *Will the data be safely stored in trusted repositories for long-term preservation and curation?*

With respect to Privacy and Data Protection, the forthcoming EU legislation - the General Data Protection Regulation - imposes several new obligations upon the consortium partners being data processors. Moreover, several new rights are granted to data subjects and significant fines are introduced in case of a data breach. Apart from this legislation, the consortium partners regard privacy and data protection as the fundamental principle and hence apply a strict policy on this matter.

Two privacy-aware techniques will be integrated through the application of diverse machine-learning solutions throughout this method. The primary aim of these techniques is to safeguard the sensitive data, whether personal data (e.g., gender, name) or quasi-private data also known as proxies (like a user's past historical locations). These two techniques have distinct implementation methods, but both primarily address the aspect of preserving the privacy and/or security of the user's sensitive data.

Federated learning (FL), the first approach considered here, belongs to the family of distributed learning techniques commonly referred to as "secure-multi-party computation" (SMPC). The innovative principle behind Federated Learning (FL) or SMPC is that users, instead of sharing their data with other parties (clients or a central server), retain their data on their devices (e.g., smartphones, IoT applications, etc.). In this setup, each model trains a local ML model using the user's local data. To enhance prediction quality, a centralized server collects the machine learning model parameters (not the actual data) and performs some form of aggregation, aiming to enrich the final prediction results. These aggregated model parameters are then sent back to the clients, each of which updates its local weights. In the end, clients receive an improved version of the ML model prediction without needing to share their data. FL also has the added advantage of leveraging computational resources from other clients, like other distributed learning techniques.

The second approach developed through this research involves the application of differential privacy. This technique revolves around the central idea of adding a controlled amount of noise to the data, thereby protecting the user's privacy. Differential privacy techniques have the main advantage of removing/reducing the possibility of applying reverse engineering techniques on the models to retrieve the original data.

In general, in accordance with GDPR and recent EU legislations, any user data that needs to be collected is done so with explicit consent. Moreover, the adoption of the above techniques, such as FL and DP used in the context of the developed ML tasks, would provide the opportunity to improve the quality of service in personalized tasks while maintaining local data).

A DPIA (Data Privacy Impact Assessment) will be delivered as Deliverable D2.2 at Month 12, and other obligations as stated in the GDPR (e.g., a register of processing activities, data breach notification procedures, etc.) will be carried out. The Deliverable D2.2 was delayed because due to the complexity of the issue concerning access to the retrospective data and after a preliminary assessment with the Data Protection Officer of the Project Coordinator, it appeared more convenient to start the ZEPHYR Study and to deliver the DPIA along with the design of the virtual infrastructure to define the impact on the protection of secondary data (TRAMONTANA Study).

6 Ethics

- *Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).*
- *Will informed consent for data sharing and long-term preservation be included in questionnaires dealing with personal data?*

Informed consent forms for participation in research and for sharing data will be included in each data collection task dealing with personal data. In Annex I, we have reported the informed consent forms prepared for English-speaking participants. These will be previously approved by the hosting institution's Ethics committee and by the Ethics Officer and reviewers who performed the ethics screening during the grant agreement preparation. Equivalent forms will be drafted in Italian, for Italian participants, in Dutch for Flemish participants and in Polish for subjects from Rybnik. The information included in the forms needs to be accepted by all participants before the data collection may start.

Acceptance will be pursued by means of the signature of participants or legal guardians, for those tasks that involve data collection in person, which is envisioned for tasks involving minors, or by means of clicking on the relevant boxes for accepting terms and conditions of use (in the case of the data collected through the app). For those tasks that involve minors as participants, in collaboration with the independent ethics advisor, we are working on the development of the best communicative (visual and verbal) strategies for informing minors about the scope of the research and the data treatment policy to be used in the app.

7 Other issues

- *Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?*

The procedures for data sharing and requesting access to use the data are described in the MISTRAL data policy (separate attachment).

Annex 1. Informed Consent

INFORMATION SHEET AND CONSENT FORM

FOR SUBJECT PARTICIPATION IN THE ZEPHYR (MISTRAL) STUDY

cross-sectional multi-center observational study to siZe dEterminants associated with black carbon concentration in biological and Predictors of quality of life in different subpopulations distinguisHed bY age group and Residence.

SUBJECT INFORMATION SHEET FOR PARTICIPATION IN THE STUDY

- The study involves the collection, preservation, and analysis of coded, therefore identifiable and non-anonymous biological samples -

You have been asked to participate, for research purposes, in an observational study conducted by _____. We believe that you can take part in this study as you fall into one of the age groups envisaged by the study (children: 5-15, adults: >18, elderly: >65) and live in one of the areas covered by the study (_____), identified according to the ordinal level of theoretical exposure to the areas closest to the pollutant sources.

The following study will presumably involve 3648 participants from the 3 European countries, of which 1216 participants are in Italy. The study was approved by the Local Ethics Committee identified for the Apulia Region.

Your participation will be on a totally voluntary basis. Before deciding whether or not to participate, you should read the information below and ask the experimenter for clarification on anything you do not understand.

PURPOSE OF THE STUDY

1. Collection of urine and blood samples to measure determinants associated with the concentration of black carbon (BC) in urine samples and other pollutants (PCB, Benzene) in blood and urine. The collected exposure data will be used to create a practical score (dose-response function) of environmental pollution contamination.
2. Measuring the clinical, social, and environmental determinants and predictors of quality of life (QoL), measured using a specific validated multinational questionnaire the EQ-5D-5L, and to create a probabilistic score, using feature selection algorithms in machine learning, that can predict different classes of QoL using clinical, social and environmental determinants.

PROCEDURES

If you decide to voluntarily participate in this study, we will ask you to undergo certain tests and/or procedures:

a) Anamnesis (family, physiological, remote, and forthcoming pathological) using specifically structured questionnaires
b) Assessment of dietary and lifestyle habits (e.g., physical activity) through validated questionnaires
c) Collection of a urine sample for the assay of black carbon and other metabolites
d) Collection of a blood sample by peripheral venous sampling for the assay of ferritin, folic acid, IL-6, high-sensitivity c-reactive protein, homocysteine, cholesterol, cystatin C, insulin, TSH, FT4, FT3, Vitamin D, oestradiol, Sex Hormone Binding Globulin, Immunoglobulin E, gamma-glutamyl transferase, glucose and haemochrome
e) Questionnaires for standardized measurement of health-related quality of life (EQ-5D- 5L) and environment
f) A questionnaire on the history of occupational exposure to pollutants
g) Three questionnaires on socio-economic status (HERO questionnaire, willingness to pay (WTP), willingness to accept (WTA)).

POSSIBLE RISKS AND INCONVENIENCES

Participation in the study does not entail any risks and/or side effects related to the planned diagnostic and control investigations.

The evaluations in this study were designed to minimize pain, discomfort, and any other foreseeable risks in relation to his health status.

EXPECTED BENEFITS FOR STUDY PARTICIPANTS

This study is not aimed at improving your health condition. You have the right to refuse to participate in it. In this case, you will, however, receive all the standard therapies for your condition, without any penalties, and the doctors will continue to follow you with due care.

EXPECTED BENEFITS FOR THE COMMUNITY

The aim of this study is to create a cloud-based IT platform for Health Impact Assessment (HIA) analyses that can use artificial intelligence to simulate different clinical, economic, and social scenarios, before and after the introduction of environmental risk mitigation policies, based on data and evidence derived from case studies (several European cities) with different levels of exposure to environmental, health, geographical and socio-economic factors.

Developing formulas to calculate the individual's risk of developing certain chronic and acute (non-oncological) diseases, using not only clinical but also social and environmental data, and simultaneously projecting the social and economic cost to the healthcare system (measured by QALY and DALY). The models will draw from both historical clinical data (e.g., registers) and new-generation data, i.e., collected from real

populations, organizing a real cross-sectional observational study on healthy volunteers in three case-study cities: Taranto, Rybnik, and Genk-Hasselt.

Create a collection of biological samples to validate an innovative, fast, and inexpensive method for the quantitative measurement of pollutant molecules in organic tissues, which can provide real data on individual exposure in the coming years and help prevent associated disorders.

SAMPLES COLLECTED FOR THE CONDUCT OF THE STUDY AND ANY RESIDUAL SAMPLES

Urine samples will be collected on the first day of examination using special carbon-free metal and black specimen jars (Yvsolab, Turnhout, Belgium) and stored at -48° C until final storage at -80° C. To avoid external contamination by carbon particles, we aliquot the urine samples in a clean room with filtered air (Genano310; Genano Oy, Espoo, Finland).

The study involves the collection of blood samples (approx. 35 ml). The blood samples collected will be centrifuged and aliquoted (if necessary) and frozen at -80°C until analysis. The tubes will be labelled with the unique code assigned to the participant. The samples will be analyzed at the KCM laboratories at the University of Hasselt in Belgium under the direction of Prof. Tim Nawrot. Ferritin, folic acid, IL-6, high-sensitivity c-reactive protein, homocysteine, cholesterol, cystatin C, insulin, TSH, FT4, FT3, Vitamin D, Estradiol, Sex Hormone Binding Globulin, Immunoglobulin E, gamma-glutamyl transferase, glucose and blood count will be measured in plasma.

Further use of the samples (if any) remaining from the investigations in the study will only be for scientific purposes directly related to those of the main study. The use will be made clear to you, and you will always have the right to withdraw your consent at any time and without explanation, resulting in the destruction of the samples.

For any use - of data and/or samples - for different purposes, a new authorization will be requested from the Ethics Committee, a new information sheet produced and your possible consent to participate in the study collected.

INFORMATION ON THE RESULTS OF THE STUDY

In the check list at the end of the consent form, you will be asked to indicate whether you wish to receive information about the results of this study. You can also choose not to receive any information If you have chosen to be informed, you will be informed of all information acquired through the study that is relevant to your state of health.

RIGHT TO CONFIDENTIALITY AND PROTECTION OF PERSONAL DATA

Pursuant to European Regulation 2016/679 (GDPR) and D. Lgs. no. 196 of 30 June 2003, "Code for the protection of personal data", we inform you that your personal data will be collected and archived (in hard copy and electronically) and will be used exclusively for scientific research purposes.

The personal data (e.g., name, age, gender, address) will be separated from the clinical data using a subdivision matrix with a translation code (unknown to the evaluators) that links the personal data to the subject code.

You have the right to request the up-to-date status of the data recorded concerning you and the correction of any errors, as well as to know who is responsible for keeping the data and who has access to it. Your samples and the clinical and personal data associated with them will be processed by the researchers and staff in charge in such a way as to ensure respect for your confidentiality and may be shared in coded form with other researchers, on the basis of your consent and in compliance with the regulations in force (European Regulation 2016/679 (GDPR)and Legislative Decree no. 196 of 30 June 2003, Code on the Protection of Personal Data and subsequent updates).

The only people who will be aware of your participation in the study are the investigators, doctors and nursing and technical staff involved. None of the information acquired during the study or provided by you will be disclosed to others without your written permission, unless:

- necessary to protect your rights or well-being (if, for example, you require emergency treatment); *or*
- required by the regulations on clinical trials.

In fact, the authorized personnel of the Local Ethics Committee and of the Taranto Local Health Authority - who are obliged to verify the proper conduct of the study - may have to access the original clinical records of the individual subjects, and thus become aware of your name, but they are bound by the rules on confidentiality and professional ethics not to reveal your identity to others.

When the results of the study are published or disclosed at the conference, there will be no information revealing your identity, as the data are presented in aggregated and therefore anonymous form.

If photos, videos or audio recordings are used, your identity will be protected or concealed.

For further clarification, please find attached the information and consent form for the processing of personal data in accordance with Articles 13 and 14 of the European Regulation **2016/679** (**GDPR**).

PARTICIPATION AND WITHDRAWAL

Your participation in this study is on a VOLUNTARY basis. Failure to participate will have no effect on your relationship with ASL Taranto, or on your right to receive treatment or other services provided by this organization.

The consent freely given by you may be revoked at any time, without any disadvantage or prejudice to you and without you having to provide any explanation, with the associated right to request that all previously collected samples be destroyed or permanently anonymized unless they have already been fully used.

Should you decide to withdraw your consent, no new information will be collected and added to existing data or databases.

WITHDRAWAL OF THE SUBJECT BY THE INVESTIGATOR

The investigator may decide to withdraw her from the study if conditions arise that make it necessary, for reasons of clinical safety. If certain conditions arise that make you no longer eligible (e.g., you are facing a therapy that could cause a loss of sensory function or cognitive decline), you may have to drop out, although you may wish to continue. The investigator will decide on this and let you know whether it is possible for you to continue participating in the study. The decision may be aimed at safeguarding your health and safety, or it may be forced by the study protocol, which may indicate that subjects who develop pathologies must be excluded.

NEW FINDINGS

During the study, you will be informed of any new findings (positive or negative), which may change your decision to participate in the study. If you are provided with new findings, you will also be offered a new information sheet and asked to renew your consent to participate in the study.

IDENTIFICATION OF EXPERIMENTERS

If you have any questions about the study, please contact us:

- _____

RIGHTS OF SUBJECTS PARTICIPATING IN A STUDY

You have the right to withdraw your consent at any time, without explanation, and stop participating in the study without being penalized. By participating in the study, you do not waive any of your rights to care and treatment including compensation for any damages resulting from the clinical practice. If you have any questions regarding your rights as a subject participating in a study, please contact your local study contact person:

- _____

CONSENT FORM

I have read (or someone has read to me) the information provided on the previous pages. I was given the opportunity to ask questions and received satisfactory answers. I was also given a copy of this form.

I, therefore, agree to participate in the proposed clinical trial and undergo the following procedures:

- Anamnesis (family, physiological, remote, and forthcoming pathological) through the use of specifically structured questionnaires
- Assessment of dietary and lifestyle habits (e.g., physical activity) through validated questionnaires
- Collection of a urine sample for the assay of black carbon and other metabolites
- Collection of a blood sample by peripheral venous sampling for the assay of
- Questionnaires for the standardized measurement of health-related quality of life (EQ-5D- 5L) and environment
- A questionnaire on the history of occupational exposure to pollutants
- Three questionnaires on socio-economic status (HERO questionnaire, willingness to pay (WTP), willingness to accept (WTA))

In addition,

- ☐ I consent
- ☐ do not consent to the general practitioner being informed of my participation in this clinical study

- ☐ I would like to receive information on the results of the study, any new findings and/or diagnostic/therapeutic possibilities
- ☐ I do not wish to receive any information on the results of the study

Subject name

Subject's signature *Date*

Signature of the Experimenter *Date*

I have explained the study to Mr/Mrs.

and/or his legal representative and I have answered all his questions. I believe that he has understood the information provided and contained in this document and has voluntarily expressed his consent to participate.

Name of experimenter

Date (must be the same as the subject's signature)

DECLARATION OF CONSENT TO THE PROCESSING OF PERSONAL DATA PURSUANT TO

AND FOR THE PURPOSES OF THE EUROPEAN DATA PROTECTION REGULATION

I, the undersigned _____noted the information provided by the test center

EXPRESSES CONSENT

to the processing of their personal data, for the purposes and in the manner indicated in the information notice itself, and also for the purposes of communication and dissemination of the same, within the limits indicated above.

INTERESTED

- ☐ consent
- ☐ withholds consent

Place and date Signature

FILL IN THE PART BELOW IN CASE THE PERSONAL DATA AND IMAGES CONCERN MINORS,

INCAPACITATED PERSONS, BANNED PERSONS

I, the undersigned _____ born in _____ on _____

I, the undersigned _____ born in _____ on _____

Aware of the civil and criminal liability, as set out in Articles 75 and 76 of Presidential Decree 445/2000 in the event of false declarations or the production of false documents or data containing untrue information,

IN THE ROLE OF

- ☐ parent
- ☐ tutor / curator

About Mr./Mrs. _____ born in _____on _____

Having taken note of the information provided by the A.O.U. Policlinico di Bari

EXPRESSES/EXPRESSES CONSENT

to the processing of the personal data and images of the person concerned, for the purposes and in the manner indicated in the information notice, including for the purposes of their communication and dissemination, within the limits indicated above.

FATHER

- ☐ Expresses consent
- ☐ Denies consent

MOTHER

- ☐ Expresses consent
- ☐ Denies consent

TUTOR/CURATOR

- ☐ Expresses consent
- ☐ Denies consent

Subject's signature *Date*