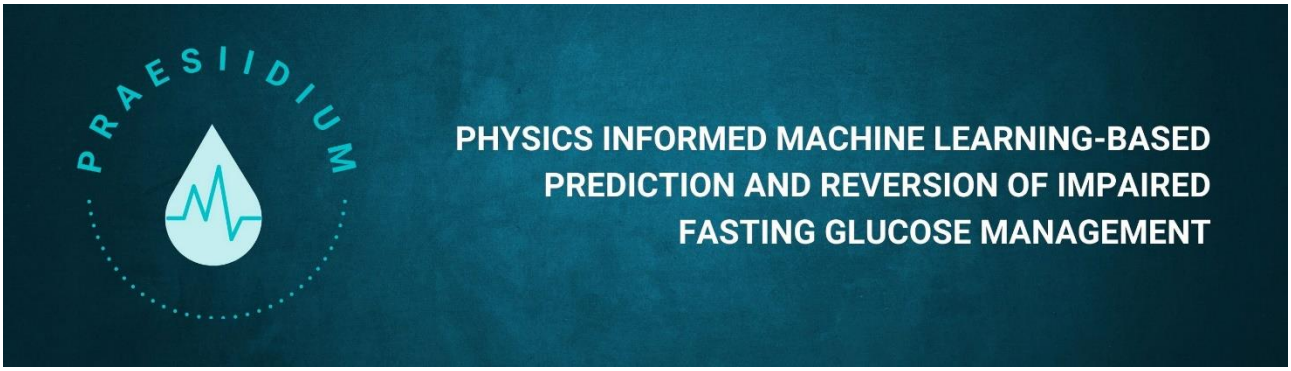

D1.6 – Data Management Plan v2

Due date of deliverable: 31/12/2024

Responsible partner: SPXL

DOCUMENT DETAILS

Title	**Data Management Plan v1**
Deliverable No.	1.6
Work Package	WP1 – Project Management and coordination
Task	T1.4
Deliverable Type	R – document, report
Lead Partner	SPXL
Contributing Partner(s)	all
Due date of deliverable	31/01/2024
Actual submission date	30/12/2024
Abstract	This document is an updated version of the Data Management Plan (DMP) submitted on M6, which establishes the data management procedures to be followed within the project by all partners. It has been updated to better address the needs emerged during the project's evolution, i.e. data management responsibility within the consortium, data transfer agreement, etc. We finally included considerations on gender fairness within the project and compliance with future regulations were also discussed.
Keywords	Data Management Plan, Data security, GDPR, privacy, high sensitivity data

Dissemination level:		
PU	Public	X
RE	Restricted to a group specified by the consortium (including Commission Services)	
CO	Confidential, only for members of the consortium (including Commission Services)	

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

CHANGE HISTORY

Version	Date	Changed by	Changes made
V1.1	15/04/2024	Giulia Masetti (SPXL)	Integration of 5.2.3 on DTA and MTA and par 6.4 on collaborative working
V1.2	04/11/2024	Giulia Masetti (SPXL), Fiorella Operto and Gianmarco Verruggio (SdR)	Integration of paragraph 4.1 and 4.2 on gender fairness and compliance with future regulations
V1.3	15/12/2024	Giulia Masetti (SPXL)	Completed draft for consortium revision
V1.4	30/12/2024	Paolo Tieri (CNR), Thomas Justa (EURO), Roberto Larcher and Giulia Masetti (SPXL)	Corrections of typos and partner revision. Preparation of the final version for submission.

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

TABLE OF CONTENTS

Document Details 1
Change history 2
Table of Contents 3
List of Figures 5
1. Overview of the deliverable 8
1.1. Scope 8
1.2. Audience 8
1.3. Summary 8
1.4. Structure 8
2. Introduction 9
2.1. PRAESIIDIUM in pills 9
2.2. Aim of the project 11
3. Data summary 12
3.1. Origin and purpose of data 12
3.2. Types of the data within PRAESIIDIUM 12
3.2.1. Real-life data 12
3.2.2. Synthetic data 13
3.2.2.1. Provision and generation of synthetic data via the MT2D model 13
3.2.2.2. ML-based trained models and predictions 14
3.2.3. Open-source data 14
4. Ethical regulatory framework 14
4.1. Gender fairness 15
4.2. Compliance with future regulations 15
5. Ethical considerations for 'high sensitivity' data management 16
5.1. Types and purpose of 'high sensitivity data' within PRAESIIDIUM 16
5.1.1. Data from retrospective clinical studies 16
5.1.2. New highly sensitive data collected within the project timeframe 17
5.1.3. Metadata generated from high sensitivity data 18
5.1.4. Purpose of the high-sensitive data within PRAESIIDIUM 18
5.2. Main risks to 'high sensitivity data' and mitigation practices 19
5.2.1. Risk Assessment 19
5.2.2. Mitigation practices for high sensitivity data handling and sharing 19
5.2.3. Data Management Responsibilities within consortium 20
5.2.4. Repositories for data collection 20

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

5.2.5. High sensitivity data quality and standards 21
5.2.6. Storage and data curation 21
5.2.7. Restrictions or delays to sharing, with planned actions to limit such restrictions ... 21
5.2.8. High-sensitivity data sharing infrastructure 21
6. Data controller and data security 22
6.1. Data Controller 22
6.2. Data Security 22
6.3. Repository and Recovery 23
6.4. Collaborative working 23
6.4.1. Project manuscript elaboration 23
6.4.2. Collaborative code development platform 24
7. Conclusions 24
8. References 24

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

LIST OF FIGURES

Figure 1- PRAESIIDIUM's project details 9

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

Explanation of acronyms and abbreviations

Acronym	Full name
AI	Artificial Intelligence
BMI	Body Mass Index
BW	Body weight
CGM	Continuous glucose monitoring
CHK	CheckHealth Sweden AB
CNR	Consiglio Nazionale delle Ricerche
CO_2	Carbon dioxide
CPCSSN	Canadian Primary Care Sentinel Surveillance Network
CSV	Comma-separated values
CTL	Cytotoxic T-lymphocytes
DC	Dendritic cells
DMP	Data Management plan
NDPA	National Data Protection Authority
ECG	Electrocardiogram
EMR	Electronic Medical Records
ESA	Ethic Self Assessment
EU	European Union
FIF	Fondazione Italiana Fegato Onlus
FFQ	Food Frequency Questionnaire
FHIR	Fast Healthcare Interoperability Resources
GDPR	General data protection regulation
GEO	Gene Expression Omnibus
HbA1c	Glycated haemoglobin
HKL	HK3 Lab Srl
hsCRP	High-sensitivity C-reactive protein
IC	Immune Complex
IFN-g	Interferon-gamma
Ig	Immunoglobin
IL	Interleukin
ISO	International organisation of standardisation
LPS	Lipopolysaccharide
JSON	JavaScript Object Notation
min	Minutes (time)
ML	Machine learning
MT2D	MISSION Type-2 Diabetes (project and model)

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

MUG	Medical University Graz
O_2	Molecular oxygen
OGTT	Oral Glucose Tolerance Test
NK	Natural Killer cells
PIML	Physics-informed machine learning
SPXL	Spindox Labs
SQL	Structured Query Language
SUPSI	Scuola Universitaria Professionale della Svizzera Italiana
T2D	Type-2 Diabetes
TGF-b	Transforming growth factor-beta
TH	T-Helper cells
TNF-a	Tumor necrosis factor
Treg	Regulatory T-cells
UNIL	University of Latvia
VO_{2max}	Maximal oxygen consumption
WP	Work package

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

1. OVERVIEW OF THE DELIVERABLE

1.1. Scope

The Data Management Plan (DMP) of the PRAESIIDIUM project demonstrates the data management procedures to be followed during the project duration by all partners.

The DMP will also define potential issues associated to data management, in the most understandable manner to be accessed anytime by all project partners to avoid possible data mishandling and misuse. This document is intended to be considered by all project partners within all project activities and procedures dealing with data management, as well as during the elaboration of all project deliverables and other relevant documents.

1.2. Audience

The PRAESIIDIUM DMP version 2 is public. We have pre-identified some targeted audience who may benefit from the description here provided:

- **Technology providers** such as researchers in academia, SMEs or other institutions working on artificial intelligence (AI)-based algorithms for preventative and personalised medicine which act as *data controller* in using data and "high-sensitivity" data. AI algorithms are also becoming restricted under European regulations such as the AI-Act.
- **Clinicians and healthcare professionals** e.g., in the fields of endocrinology, hepatology, metabolism, etc. who conduct clinical studies and act as *data provider* of "high-sensitivity" data.
- **Patients, caregivers, and healthy people** involved in clinical studies will be the actual *data owner*. The provision of biological samples and data are solely on a voluntary basis prior to the Informed Consent, which can be withdrawn at any time.
- **Industry** including (but not limited to) pharma companies, wearable devices manufacturers, patients remote control manufacturers, etc.

1.3. Summary

This document presents the updated version of the Data Management Plan (DMP) of the PRAESIIDIUM project, which establishes the data management procedures to be followed within the project by all partners. Such measures are meant to avoid any possible issues related to data misuse and breaches of security. The present version includes detailed data management responsibilities within consortium, consideration on gender fairness and compliance with future regulations.

1.4. Structure

The deliverable D1.6 will be structured in the following sections:

- Introduction of project aims and rationale.
- Data summary: this section describes the data collected and generated within the project, their types and purposes related to the PRAESIIDIUM objectives.
- Presentation of the Ethical framework on which this document is based on and gender fairness principles.

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

- Ethical considerations for 'high sensitivity' data collection, handling, and management, including potential ethical and privacy issues. A list of responsibilities identified within the consortium related to data/metadata collection and usage is also presented.
- Data controller and security: procedures to be followed to ensure a secure storage and management of the general data of the project.

2. INTRODUCTION

2.1. PRAESIIDIUM in pills

- In 2021, 1 in 11 European adults were living with diabetes (61M people); by 2024 its incidence is expected to increase to 69M people,
- 36% European adults live with **undiagnosed diabetes**, which if not properly cure can cause life-threatening outcomes,
- Impaired glucose tolerance (IGT) and Impaired fasting glucose (IFG) are precursors of the overt diabetes called "**prediabetes**" and – if early diagnosed – can be reversed by lifestyle modifications (e.g., healthy diet and physical activity),
- Therefore, there is an urgent need of personalised, early diagnosis tools,
- The PRAESIIDIUM project will develop a tool for the **real-time prediction** of the **prediabetic risk** of an individual based on "physics-informed machine learning" (PIML) model accessible through a user-friendly web-based application,
- The PIML will make use of multi-scale, integrated mathematical models (MT2D integrated model) coupled with explainable AI trained on large-scale, real-life, and already existing clinical databases, plus data from two newly dedicated clinical studies with continuous ingestions by wearable sensors.

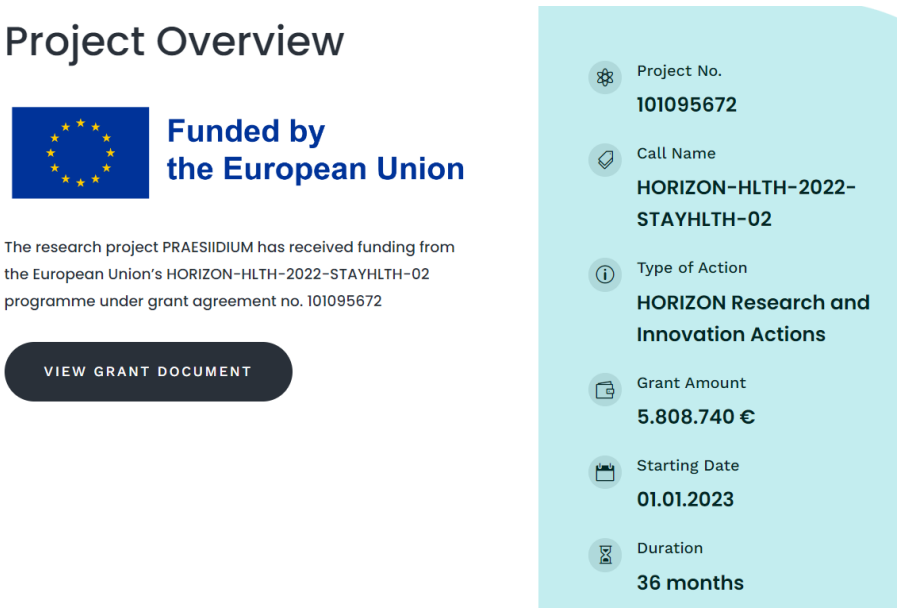

Figure 1- PRAESIIDIUM's project details

Diabetes is a chronic disease which occurs when the pancreas no longer produces insulin (or the body is not able to correctly use the secreted insulin[1]), leading to an increased glucose blood level (i.e., *glycaemia*). European prevalence of diabetes in 2021 was of 1 in 11 adults (61M)[2], and its incidence is predicted to increase to 69M by 2045. Diabetes is also the third leading cause of death, responsible of a total 6.7M deaths (1.1M only in Europe).

Amongst others, **type-2 diabetes** (T2D) accounts for 90% total diabetes cases with an adulthood onset. It is characterised by insulin resistance (i.e., the insulin hormone cannot effectively regulate the blood glucose levels) and/or a failure in compensatory mechanisms for insulin secretion (i.e., the body does not correctly use the insulin produced) [1]. Besides genetic predisposition and family history, metabolic syndrome and obesity increase the risk of T2D development, along with physical inactivity, age, and ethnicity. Despite a management of diabetes is still possible considered all the challenges, taking **effective prevention measures** are mandatory.

According to the IDF Diabetes Atlas (10th edition)[2], 36% European adults are living with undiagnosed diabetes, and such incidence is much higher in low/middle income countries, where 1 in 2 people live with undiagnosed diabetes. Also, 541M adults (20–79-year-old) worldwide have *Impaired Glucose Tolerance* (IGT), an important risk factor for further T2D development. Both IGT and/or the *Impaired Fasting Glucose* (IFG) are intermediate condition within the "healthy-to-T2D transition" and are manifestations of the so-called prediabetes condition. For the sake of completeness, the definitions of IGT and IFG are:

- **Impaired Glucose Tolerance** (IGT) occurs when "glycaemia is higher than normal, but not high enough to be classified as diabetes" [2]: blood glucose level of 140 to 200 mg/dL 2 hours after a 75g oral intake of glucose (during oral glucose tolerance test),
- **Impaired Fasting Glucose** (IFG) happens when "glycaemia level is consistently higher than normal ranges in the fasting state, but not high enough to be classified as diabetes" [2]: fasting blood glucose level of 100 to 125 mg/dL5. In 2021, around 25,600 European adults had IFG, and this number is set to increase close to 27,000 by 2045.

As for T2D, family history and ethnicity are important risk factors for prediabetes. Moreover, prediabetes can occur because of the metabolic syndrome (MetS), i.e., presence of high body fat (central obesity), non-alcoholic fatty liver disease (e.g., steatosis) and insulin resistance.

Unlike overt T2D, **prediabetes** is still an early-stage, intermediate condition that **can be reversed**; in fact, "*the rate of T2D progression within 3 years can be decreased by approx. 58% with a lifestyle modification*" [2]. **Physical activity** should be performed for minimum 30-54 minutes at least 2-5 days/week, and a **healthy diet** is also strongly recommended (2014 WHO and IDF recommendations7)[3,4].

[1] Diabetes (who.int)
[2] IDF Diabetes Atlas | Tenth Edition
[3] Prevention (idf.org)
[4] Framework-for-Action-on-Sugar-010615.pdf (idf.org)

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

2.2. Aim of the project

The "physics-informed machine learning-based prediction and reversion of impaired fasting glucose management" (PRAESIIDIUM) project will exploit and optimize some of the outcomes of the multi-scale, hybrid computational model MT2D of the human metabolic and inflammatory status (developed by CNR during the EU-FP7 project **MISSION-T2D**[5]), along with the integration of a new research area called "*physics-informed machine learning*" (**PIML)** [3] to develop an easy-to-use, reliable, and efficient tool for the prediction of the prediabetes risk. Thus, the aim of the PRAESIIDIUM project (i.e., in Latin "*praesidium*" means "protection") is:

To develop a tool aimed at providing a **real-time prediction of the prediabetic risk of an individual**. The prediction algorithm will be based on a "*physics-informed machine learning*" approach: a rich dataset of real-life data, obtained from already existing previous and a new clinical trial with continuous data ingestion through wearable sensors, will be combined with mathematical models and explainable artificial intelligence (XAI) techniques, to overcome the limits of "*black-box*" machine learning approaches while improving the prediction performances and reducing the computational time of the risk calculation based on the simulations of the mathematical model. The final algorithm will be implemented in a web-based platform, where medical doctors and patients can inject data from several sources (acquisition form connected sensors and manual insertion) and obtain a real time analysis of the risk to develop the prediabetic condition over time.

[5] www.mission-t2d.eu/MISSION-T2D/index.html

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

3. DATA SUMMARY

3.1. Origin and purpose of data

The PRAESIIDIUM project aims to develop a tool for the real-time prediction of the prediabetes risk of an individual based on a "*physics-informed machine-learning*" (PIML) algorithm, which will combine trajectories from a multi-scale mathematical model (from the previously EU-funded MT2D project), experimental datasets and information coming from the dynamics of specific variables involved in the human metabolism, in an explainable Machine Learning prediction algorithm, finally integrated into a user-friendly web-based platform for doctors and patients, able to generate a personalized prediabetes risk score. The PIML will be trained and tested on three large-scale real-life clinical databases and on two newly dedicated clinical studies; the latter will also provide a continuous ingestion of data from wearable sensors to monitor the physical activity of the individual.

Within the PRAESIIDIUM project:

- **'Data'** refers to information – datasets and/or user findings – collected, examined, and considered for the reasoning, discussion and/or calculation (e.g., statistic results, measurements, observations from questionnaires, surveys, wearable sensors, etc.) according to the project implementation.
- **'Metadata'** refers to information about the data itself and the basic characteristics of the data (not limited to e.g., who created the data, when/how/where the data were generated etc.).

The project foresees two main phases whereas data and metadata will be used and/or generated:

- The **'analysis and development phase'**, in which *in-silico* and retrospective clinical datasets are used to train the PIML model and to generate the prediabetes risk score,
- The **'clinical study phases'**, in which three retrospective databases and two newly developed clinical studies will collect or generate more data to ameliorate the risk prediction score.

3.2. Types of the data within PRAESIIDIUM

The types of data gathered and used within PRAESIIDIUM can be summarized into three main categories:

- **Real-life data** collected or generated during the project timeframe from already existing large-scale clinical databases and newly developed clinical studies.
- **Synthetic data** created in an "artificial manner" because of realistic simulations (e.g., from MT2D in-silico data) or from machine-learning algorithms.
- **Open-source data** retrieved from open source/open access published databases or articles (e.g., Human Protein Atlas, Expression Atlas, etc.)

3.2.1. Real-life data

Real-life data are initially gathered from the PRAESIIDIUM retrospective clinical case studies (owner: CNR, FIF and MUG) and intended to be representative of real-life scenarios of prediabetes insurgence in adulthood: e.g., healthy individuals with family history, metabolic risk factors and

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

obese/bariatric cohort with the highest risk of prediabetes onset. During the project timeframe, at a later stage, two newly dedicated clinical studies will be conducted within the project to collect more data. Real-life data also includes the observations from clinical visits, e.g., Food Frequency Questionnaire (FFQ), physical activity questionnaires and continuous data ingestion from wearable sensors (e.g., Fitbit).

A more detailed description of these data and guidelines for collection, sharing and usage are provided in following chapter.

3.2.2. Synthetic data

Synthetic data will be gathered and generated within PRAESIIDIUM from the following identified sources:

3.2.2.1. Provision and generation of synthetic data via the MT2D model

CNR will provide quantitative, *in-silico*-generated synthetic dataset from 46170 simulations.

The **input** of the simulations includes the following discrete starting parameters:

- Gender (M/F);
- weight (underweight, normal, overweight);
- height (short, average, tall);
- number of sessions of physical activity (0, 1, 2, 3);
- duration of the bout of physical activity (30, 60, 90min);
- intensity in terms of $\%VO_{2max}$ (40, 60);
- 3 meals per day; in each meal are specified the carbohydrates (low, medium, high), proteins (low, medium, high) and fats (low, medium, high).

Output numerical values of the simulations include:

- **inflammation markers** (recorded every 8h): B-cells (B-1, B-2), PLB, TH (Th1, Th2, Th17, Treg), CTL, Treg, NK, MA, DC, EP, ADIP (number, volume), IgM, IgG (IgG1, IgG2), IC, IL-2, Danger, IL-12, IFN-g, IL-4, TNF-a, TGF-b, IL-10, IL-6, IL-10, IL-18, IL-23, IFN-b, IL-1b, LPS, Leptine;
- **metabolic outcomes** (recorded every minute):
  - Arterial concentrations (glu, pyr, lac, ala, gly, ffa, tgl, O_2, CO_2)
  - Organs (22 metabolites)
  - Hormones (insulin, glucagon, epinephrine)
  - Fasting glucose
  - Rate of appearance (glucose, alanine, triglycerides)
  - Total daily energy balance
  - $\%VO_{2max}$
- **anthropometric measures**: BW, BMI, fat mass, fat free mass

Further datasets from a coherent number of data will be produced and provided with the same characteristics (number of simulations to be decided).

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

3.2.2.2. ML-based trained models and predictions

Technology partners (e.g., SUPSI, CNR, HKL) involved in the training of machine learning methods will handle data from various sources, mainly, but not restricted to, surveys, clinical measurements, medical records, and electronic health records; made available from other partners within the project. Public datasets and/or additional simulations might be included to enrich the given data if approved by the partners.

3.2.3. Open-source data

Partners will also integrate data collected within the project with previously published data in open-source databases or open-access publications. More specifically FIF will integrate datasets from open databases like the Human Protein Atlas (proteinatlas.org), GEO (https://www.ncbi.nlm.nih.gov/gds), and Expression Atlas (https://www.ebi.ac.uk/gxa/home). Detailed information on open-sourced data and their purposes within the project are described elsewhere (Deliverable D2.2). In any case, involved partners will ensure the compliance of such data with regulations in terms of data management.

4. ETHICAL REGULATORY FRAMEWORK

The PRAESIIDIUM Consortium tailored the project according to an **Ethics-By-Design (EbD) methodology** and the ethical analysis and requirements are outlined elsewhere (Deliverable D2.3), encompassing all aspects of the overall process. For adequately addressing any ethical issues that could arise in the project execution related to privacy and consent, the data collected and processed during e.g., focus groups, and during retrospective data collection and from the observational clinical studies will only be processed in the way described in the following paragraphs.

All participants in PRAESIIDIUM clinical prospective observational studies will be informed of all aspects of the research that might reasonably be expected to influence willingness to participate. Ethical approvals were obtained from local ethical commissions (MUG, UNIL) before the initiation of the prospective data collection, as well as for the retrospective data analysis (if not already applicable with the actual consent). It is worth noting that for which concerns the clinical study provided by FIF, recruitment has been already started in a still ongoing clinical study and therefore there is not the need of a new ethical protocol request. All participants, in any case, have the right to withdraw from the research at any time, and to withdraw retrospectively any consent given and to demand that their personal data be destroyed. The project will avoid situations where participants feel pressurized to take part in the observational studies.

All the information collected during the study is considered confidential and is treated in compliance with:

- Reg. 2016/679/UE General Data Protection Regulation (GDPR)
- Data Protection Act (DPA)
- Research Organisation Act (FOG, Federal Act on General Matters)
- Legislative Decree No. 196 of June 30, 2003, "Personal Data Protection Code"

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

- Deliberation No. 52 of July 24, 2008, "Guidelines for the Processing of Data in the Context of Clinical Trials of Medicines by the Privacy Authority"

For which the data collected through wearable sensor devices is concerned, such device data will be formatted according to device type and manufacturer and may be based on Continua/ISO/IEEE 11073 Personal Health Data Standards, JSON text format and HL7/FHIR. The CHK security policy is based on the framework recommended by the Swedish Contingency Agency, following SS-EN ISO/IEC 27001/27002. CHK is however not formally ISO certified.

4.1. Gender fairness

The PRAESIIDIUM ethic committee is chaired by ethics and gender experts from School of Robotics (SdR), to guide gender-fair decisions regarding the data management process. After a detailed analysis of ethical requirements for both retrospective and prospective clinical studies (Deliverable D2.1 and D2.3), coupled with outcomes from the consortium ethics-self-assessment (ESA), the consortium has been made aware that:

- There might be a gender bias present in clinical databases, which may reflect a gender prevalence of the disease itself or willingness to participate.
- A gender bias in the training dataset may result in biases of the trained algorithms.
- The quality of retrospective data can impact on gender biases, effects on patient diverse socio-economic backgrounds, and relevant demographic factors.

Therefore, we recommend that:

- A careful tracing of data cleaning and preprocessing should be implemented for all the retrospective datasets.
- Harmonisation procedures have been already implemented to enhance reproducibility and control for biases and unseen stratification aspects.
- Targeted recruitment should be established to obtain a participation of individuals with different demographics, socioeconomics, and relevant variability.
- Informed consent and non-discriminative selection criteria should be implemented.
- Respect of gender fairness to ensure that data collection, analysis, and dissemination are inclusive, equitable, and avoid reinforcing gender biases.

Finally, gender fair language is recommended in data descriptions, reports, and dissemination materials to avoid misrepresenting gender or social groups. Similarly, in visualizations and publications, we make sure both genders are fairly represented and depicted in ways that do not perpetuate stereotypes.

4.2. Compliance with future regulations

The current DMP is well aligned with the main EU regulation on data protection, namely the GDPR, and will be also reviewed based on future regulations such for example the AI Act.

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

The AI Act[6], introduced in April 2021, has been recently approved by the EU Parliament in March 2024, and is a significant legislative initiative aimed at regulating the development and deployment of artificial intelligence systems within the European Union. As of latest update (March 2024), the official enter into force of the Act will occur in several months, but it will delineate new guidelines, which might be crucial for the development and further exploitation of AI-based systems such in PRAESIIDIUM. We will keep the implementation of AI-Act monitored and will take due measures to get aligned with it.

5. ETHICAL CONSIDERATIONS FOR 'HIGH SENSITIVITY' DATA MANAGEMENT

5.1. Types and purpose of 'high sensitivity data' within PRAESIIDIUM

According to EU guidelines, "high sensitivity data" are personal data subjected to the following conditions: *"[…] genetic data, biometric data processed solely to identify a human being, health-related data…"*[7]. Such type of data fits in the "real-life data" category presented in the general overview of the PRAESIIDIUM data types (Par. 3.2.1.).

As already summarised, PRAESIIDIUM will collect and use mainly three groups of "high sensitivity data":

- Data retrieved from already existing large-scale clinical databases analysed in a **retrospective** manner (WP2, see D2.1),
- Data collected during the project in two dedicated **prospective** clinical studies (WP6),
- Metadata generated from high-sensitivity data handling, usage and storage.

5.1.1. Data from retrospective clinical studies

High sensitivity data from retrospective clinical studies are collected and analysed from already-existing large-scale databases, in particular:

- **Case study 1: Healthy individual with a family risk of prediabetes** (MUG). The 'BioPersMed project' [4] is designed as a single-centre, prospective, observational study. Only asymptomatic subjects without diagnosed of cardiovascular diseases (CVD) but with at least one traditional CV risk factor were eligible to participate, besides of age and gender, comprising smoking, elevated total cholesterol levels and arterial hypertension. Moreover, sedentary lifestyle, obesity, social environment, or T2DM, cholesterol, increased triglyceride levels, familial hypercholesterinaemia, etc. At baseline and at regular 2-years follow-ups, an in-depth diagnostic CVD and MD work- up was carried out, laying emphasis on standardisation and reproducibility of history taking, questionnaires for health, psychological and sleep issues, physical examination, ECG, laboratory/blood sampling with biobanking, exercise testing (6 min walking test, grip strength, spiro-ergometry), echocardiographic analysis of cardiac structure and function, pulmonary function testing, carotid intima/media-thickness measurement, as well as

[6] Regulation - EU - 2024/1689 - EN - EUR-Lex
[7] What personal data is considered sensitive? (europa.eu)

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

body composition, bone density including bone and hormonal biomarkers, and oral glucose tolerance testing (OGTT).

- **Case study 2: Liver-disease and obese patients with risk of prediabetes** (FIF). The clinical-biochemical and histological data at baseline from N=250 morbid obese patients will be included in the retrospective dataset. The clinical parameters have been obtained from blood samples after 8 hours of fasting (in average two weeks before of the surgical intervention) and liver biopsies performed during the abdominal surgery. Apart for morphometric parameters, clinical and biochemical parameters, also histological evaluation of liver biopsies (semi-quantitative) is available, as long as plasma biobanking from already-enrolled patients.
- **Case study 3: Trajectory (longitudinal) data for health-to-prediabetes-T2D transition** (CNR). The data source of this case study is the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): a nation-wide database that contains de-identified information from Electronic Medical Records (EMRs) belonging to patients that undergo primary care visits in Canada[8]. The portion of the CPCSSN database that will be available for this case study includes EMRs collected between 01/01/2000 to 31/06/2015 from overall 1,283,154 patients (44% males, 56% females), 90,278 of which diagnosed with diabetes (including T2D, Type-1 diabetes, and gestational diabetes; 51% males, 49% females). Each patient is associated with several records in the database including, for example, data from medical encounters (e.g., diagnosis, prescribed medical examinations), laboratory results (e.g., biomarkers collected from blood samples), results of physical examinations (e.g., blood pressure measurements, weight), and drug prescriptions. As the database contains data on primary care encounters, the number of records for each patient can vary from a few records to several tens of records, as it depends on how often patients visit their general practitioner and on their general health status. It is worth noting that the CPCSSN version used in this case study is provided through an agreement with Toronto Metropolitan University (formerly, Ryerson University), as mentioned in the Letter of Support attached to the Project Grant Agreement. Specifically, CPCSSN are provided to Toronto Metropolitan University via an agreement that was exempt from IRB approval as the University Review Ethics Board provided a waiver of review since the data is de-identified and cannot be linked to other databases for re-identification of patients. Access to the database has been granted to researchers at CNR-IEIIT through earlier agreements (individual visiting researcher agreements) that are regularly renewed since August 2019.

5.1.2. New highly sensitive data collected within the project timeframe

The newly generated high-sensitive data collected within the PRAESIIDIUM project timeframe belong to the following studies (WP6):

- **Clinical study 1**: Patients with low risk of prediabetes conducted at Medical University of Graz (MUG) registered at the German Register of Clinical Studies DRKS-ID: DRKS00033540
- **Clinical study 2**: Patients with metabolic risk factors conducted at University of Latvia (UNIL), Registered at ClinicalTrials.gov with ID: NCT06195566

[8] For Clinicians - Canadian Primary Care Sentinel Surveillance Network (cpcssn.ca)

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

- **Clinical study 3**: Plasma bio-banking and clinical data collection at the Italian Liver Foundation (FIF) is an already ongoing patients/samples recruitment, the same aforementioned data and samples will be collected (see par 5.1.1). The study is registered on ClinicalTrials.gov public website under identifier NCT06098417.

A non-exhaustive list of data/variables to be collected during the clinical study #1 and #2 are here summarized.

I. **Demographic Data**: demographic information such as age (no birthdate allowed), gender, ethnicity, and family history will be collected at the enrolment visit.

II. **Health-related Data:** such as BMI, HbA1c, glycaemia (fasting glucose level), insulin, previous medications, ketones, heart rate, ECG, and hsCRP, inflammatory markers, lipids, etc.

III. **Lifestyle Data**. Lifestyle factors such as mealtimes, food frequency questionnaires (FFQs), meal nutrition values, physical activity frequency data.

IV. **Continuous ingestion through wearable sensors:**

   a. **Continuous Glucose Monitoring** (CGM) as a mini-invasive real-time glucose measurements to represent a continuous tracking of glucose fluctuations and trends.

   b. **Tracker of physical activity** e.g., Fitbit.

5.1.3. Metadata generated from high sensitivity data

Associated metadata generated from the usage and collection of high-sensitivity data have been identified and are (not-limited to):

- A comprehensive data dictionary to provide detailed descriptions of all data fields, codes, units of measure, and any special formats (It will be updated regularly and stored alongside the main data files),
- Information about different parameters used by the instrumentation (temperature, absorbance wavelength, time of measurement), range of standard curve used, and dilution of the sample, amongst others,
- Training procedure of the ML models and subsequent validation of those predictions using dedicated scoring methods and standard practices (e.g., cross-validation),
- Characteristics of the wearable devices (type, manufacturer, etc.) used for remote data collection, individual observations (such as timing, sources, metrics). Support for additional meta data for remotely collected data (e.g., non/wear-time, accuracy) wearable devices will be investigated.

5.1.4. Purpose of the high-sensitive data within PRAESIIDIUM

Within PRAESIIDIUM, the collection and usage of all high-sensitive data can be summarized into three main purposes:

1. Input Data for the training of the MISSION-T2D model (MT2D, CNR)
2. Data for training of the physics-informed machine learning (PIML) algorithms

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

5.2. Main risks to 'high sensitivity data' and mitigation practices

5.2.1. Risk Assessment

Collection, generation, usage, and storage of high-sensitivity data may raise ethical concerns mostly related to breaches of informed consent and privacy issues. Within PREASIIDIUM the following areas have been investigated to raise awareness on the possible risks related to high-sensitivity data collection and usage:

a) Identified risks associated strictly related to the "high-sensitivity data" collected within both retrospective and prospective clinical studies are:
   - unauthorized access,
   - data breaches,
   - data/sample loss or corruption,
   - insider threats,
   - non-compliance with informed consent conditions.

b) Identified risks concerning remotely collected data (wearable sensors, questionnaires, etc):
   - patient data measurements lost or inaccurate due to system technical fault,
   - lack of alarm of patient data measurement outside reference interval,
   - data from another individual reported for a specific patient.

c) Identified risks concerning machine learning models using high-sensitivity data for train and validation processes:
   - personal-related information leakages,
   - unattended access,
   - exploitation of a trained model to extract personal-related information.

5.2.2. Mitigation practices for high sensitivity data handling and sharing

To avoid and mitigate the above identified risks, collection, usage and sharing of high-sensitivity data must comply with the following conditions, in line with the principles of integrity and confidentiality mentioned in the Art. 5 of the GDPR:

- Mandatory compliance to written informed consent.
- Mandatory **pseudonymization** of all information and samples complying with the EU-GDPR requirements. The key for re-identification of the involved participants remains at the Data Provider. Therefore, the Data Provider is best able to manage requests from participants for access, rectification, restriction of processing, data portability, objection to the processing or erasure of their personal data. The original identity of the participant will be retained as confidential at the data collection point (e.g., hospital, clinic). In fact, storage of personal data and contact information (e.g., name, surname, phone number, email address) at the data collection points will be necessary due to the longitudinal nature of the clinical study, with necessity to stay in contact with subjects involved in these.

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

- Data directly or indirectly related to people involved in the PRAESIIDIUM clinical studies will never be provided to subjects not directly involved in the consortium.
- Data provided to third parties, included the ones included in project deliverables or publications, will be duly pseudonymized and presented as **aggregated data** as frequent possible (e.g., via statistical inference, grouped variables, etc.).

5.2.3. Data Management Responsibilities within consortium

In compliance with good data management practice and DPA, we proceeded with the institution of data management responsibilities at different levels within consortium:

I. ***Data Owner:*** are patients and/or participants in the clinical studies who donate their biological samples, data and time. They can request access only to their own data, can withdraw the informed consent and ask for their (personal) data deletion.
II. ***Data Provider***: is usually the clinical site/institution responsible of the data collection, which occurs compliant with the Informed Consent. They display full access to data archives and can access/move/delete (only if necessary and upon notification) data, can provide/share the data only in a pseudonymised manner,
III. ***Data Controller***: are consortium members that can access only the pseudonymised data needed for their current analyses within this project and not for any other purposes. Consortium members are considered *controllers* for the processing of personal data regarding the Project and will act in accordance with all applicable data protection laws (including but not limited to GDPR, "Applicable Data Protection Law"). They shall maintain the personal data in such pseudonymized form and shall not attempt to re-identify any donor(s). In the event any donor, for whatever reason, becomes identifiable to the controller, the controller agrees to immediately notify the Provider, to immediately and irretrievably remove the reference to that donor and to preserve, always, the confidentiality of information pertaining to such donors. In case of patient informed consent withdrawal, the controller is then obliged to return to the Provider, if possible, or to demonstrably destroy with required care all data received, or a separate written agreement shall be entered into between the Provider and the controller, if the personal data shall be used for another project.

Prior to data or biological samples sharing within consortium, members have established either *Data Transfer Agreements* (DTA) or *Material Transfer Agreements* (MTA), respectively. These are legal documents signed by each party, detailing duties and responsibilities for which data handling, sharing and storage are concerned.

5.2.4. Repositories for data collection

- High-sensitivity data collected from retrospective studies: data will be structured and formatted in password-protected Excel formats, CSV, or other formats which may include different database file types with the aim of ease of use and interoperability with various data analysis tools. The expected scale of the data lies in the range of megabytes or gigabytes.
- High-sensitivity data generated within the project clinical studies will be collected via secure and private electronic medical records (EMR), lab tests, along with data collected via

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

questionnaire and wearable sensors, all collected through the LinkWatch platform (JSON/FHIR, SQL storage format stored in the LinkWatch internal database – Heidi SQL).

5.2.5. High sensitivity data quality and standards

Data collection tools will be standardized and validated whereas possible. Data will be checked for completeness, consistency, and accuracy at the time of collection and before analysis. Quality assurance procedures will be put in place to minimize errors and missing data. This will be complemented by unit testing and by the risk management procedures from the CHK Quality System. Specific quality controls for data from wearable devices (e.g., validity, outlier/missing values) should be considered in the trial protocol design.

5.2.6. Storage and data curation

The data collected from clinical studies will be stored securely in a GDPR-compliant system with strict access controls and conform the local institution standards (e.g., FIF, MUG and UNIL). Data curation will be performed by dedicated data managers at each data collection point (‘*Data Providers*) who will regularly check the data for errors, update records when necessary, and ensure proper (pseudo)anonymization. The data, along with metadata and data dictionary, will be archived securely according to the institutional standards. The datasets will be backed up regularly and the backup copies will be kept in a geographically separate location to ensure data safety in case of disasters.

Collected data from wearable sensors and remote-assisted questionnaires will be managed by the LinkWatch cloud-based storage for the entire duration of the project. This includes regular back-up procedures and security provisions in accordance with the GDPR and the CHK security policy. The security policy is based on the framework recommended by the Swedish Contingency Agency, following ISO/IEEE 27000.

Following the end of project data will be maintained or possibly archived for at least **5-year period**, to be agreed by the consortium and per ethics commissions regulations.

5.2.7. Restrictions or delays to sharing, with planned actions to limit such restrictions

Restrictions on data relies on the need to protect participant privacy in accordance with GDPR regulations. Sharing of collected data will be also subjected to informed consent from users. In case there will be need to limit data access or sharing, full (pseudo)anonymisation and data encryption will be employed to facilitate restrictions where appliable.

For which the case study 3 (CNR-CPCSSN) is concerned, access right outside CNR may be granted within specific agreements to be defined with the data Provider (Toronto Metropolitan University).

5.2.8. High-sensitivity data sharing infrastructure

For which the wearable sensors and remote patient monitoring is concerned, the LinkWatch internal database (developed within the project timeframe) will be the main repositories for authorized users (e.g., clinicians, technologists) within the project.

This project has received funding from the European Union’s Horizon Europe research and innovation program under grant agreement **No 101095672**

6. DATA CONTROLLER AND DATA SECURITY

The following paragraphs will present and explain management guidelines for usage, storage and sharing of data collected or generated within the PRAESIIDIUM which are NOT identified as to 'high sensitivity data' (for which the previous chapter was dedicated to).

6.1. Data Controller

The PRAESIIDIUM project will collect and process data according to the required scientific needs and in compliance with EU and national regulations, General Data Protection Regulation (GDPR).

The Consortium agreed on the definition of the '*Data Controller*' as being the PRAESIIDIUM partner(s) responsible for the collection, handling, standardisation/curation and secure the storage of all the data collected. Each partner will therefore be responsible to adopt all the procedures proposed for data management.

As part of the Data management plan, project-wide guidelines will be developed, monitored, and discussed by project coordinator, consortium members, ethical board members in relation to:

- how to handle project-related data within the confines of the project (i.e., whether to disseminate/share, or exploit/protect our results for commercial exploitation),
- aspects concerning shareability of any parts of datasets/metadata (e.g., for technical or research publications). In the case of special disclosures of project data (i.e., outside the consortium or in presence of high sensitivity data), the project's aforementioned data governance mechanisms will evaluate the potential data sensitivity issues and make a "*to share or not*" decision accordingly,
- datasets shared inside the consortium will remain consortium-confidential and can only be shared outside the consortium if the partner(s) owning the datasets agree.

The Consortium members shall agree on the Data Management Plan (DMP) and shall each maintain a written record of all processing activities that are carried out under the Consortium Agreement as required by Applicable Data Protection Law. Furthermore, each Consortium member shall maintain appropriate administrative, technical and organizational measures to meet the requirements of the Applicable Data Protection Law to protect the personal data from misuse and unauthorized access or disclosure.

6.2. Data Security

The following guidelines have been proposed to ensure data security within PRAESIIDIUM:

- Data will be stored in at least two independent and separate locations to avoid loss of data,
- Data might be encrypted if needed by consortium members,
- Avoid or at least limit the usage of USB flash drivers,
- Share high sensitivity files through password-protected or encrypted files,
- Label files in a systematically structured way to ensure the coherence of the final dataset.

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

We will exploit state-of-the-art methods which are appropriate both to PRAESIIDIUM's circumstances and the potential risks that loss or theft data or operational processing data might pose. The measures we will deploy in PRAESIIDIUM project must enable the consortium to access data in a timely and straightforward manner for processing needs.

6.3. Repository and Recovery

A dedicated private SharePoint repository has been created for sharing and storage of all PRAESIIDIUM-related documents (**not containing high-sensitivity data**, for definition see par 5.1), where only consortium members have access to. Such repository allows access to members upon different rights (e.g., read-only, editing…) appointed and – if needed modified – by the Project Coordinator, which is in charge also of all accessibility issues. The SharePoint repository keeps a frequent backup procedure to avoid losing any project information. In case of an incident, the data can be recovered according to the procedures of the data repository owner, in the name of the Project Coordinator (Spindox Labs).

In parallel, a private Git repository has been created and shared amongst consortium members specifically for algorithms/ML methods sharing.

Below are listed and identified the main data security risks and mitigation actions foreseen:

- *Accidental data loss due to software or hardware failure*: a backup copy of the data is available. The backup copy is not in the same physical location as the original one.
- *Data access by unauthorized third party*: access to the data repository is allowed only for Consortium members and requires secure authentication mechanisms. Additionally, the access is restricted by rights to each partner contact, depending on its activity.
- *Data breach*: access to data is restricted to consortium members. When a data breach is suspected to happen, a procedure of communication and mitigation measure is put in place. The project coordinator gathers the information with the responsible ones and notify within 72 hours to the National Data Protection Authority (DPA). Then the subjects that are affected with this breach will be notified explaining the violations of their data.

According to the GDPR regulations, data must be stored by consortium members for the shortest time possible, especially those regarding high-sensitive data. As for the information deemed to be published, we will follow what is requested by EU-guidelines in terms of dissemination and exploitation we will keep (non-sensitive) data/information for at least **5 years after the project ends**.

6.4. Collaborative working

6.4.1. Project manuscript elaboration

There are two official tools to produce manuscripts and deliverables: Office365 and Overleaf. In the first case all the people taking part to the PRAESIIDIUM project have been provided with the access to the project SharePoint where it is possible to download the document templates and start editing a new document. If a new person needs access to the SharePoint, an email with the request must be sent to the coordinator technical lead that will grant access to the new person. The same procedure applies in the case of Overleaf usage.

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**

6.4.2. Collaborative code development platform

If two or more partners need to collaborate on the same code, the PRAESIIDIUM project makes available a git repository hosted on Google Cloud. Each person that wants to have access to the newly created repository must write an email to the project's coordinator asking to create a new user for them and specify the name of the repository they want have access to.

7. CONCLUSIONS

This updated version of Data Management Plan (DMP) describes the procedures to be followed by the project partners about the management of data and data having "high sensitivity" for privacy and ethics reasons. This document has been updated regards the gender fairness procedures, a more detailed definition of the responsibilities within consortium and the need for Data/Material Transfer Agreements. As a consortium, we are fully committed to following EU legislation in terms of data protection and AI trustworthiness, therefore there will be a constant oversee of these aspects during the last year of the project.

8. REFERENCES

[1] Bloomgarden Z.T. (2011) American Diabetes Association Postgraduate Meetings - 2011. Diabetes Care; 34(11): e164-e169. https://doi.org/10.2337/dc11-1557

[2] Mayans L. (2015) Metabolic Syndrome; insulin resistance and prediabetes. FP Essent. Aug; 435:11-6.

[3] Karniadakis G.E., Kevrekidis I.G., Lu L., Perdikaris P., Wang S., Yang L. (2021) Physics-informed machine learning. Nature Reviews Physics 3, 422-440. doi.org/10.1038/s42254-021-00314-5

[4] Haudum C.W., Kolesnik E.K., Colantonio C., Mursic I., Url-Michitsch M., et al (2022). Cohort profile: 'Biomarkers of Personalised Medicine' (BioPersMed): a single-centre prospective observational cohort study in Graz/Austria to evaluate novel biomarkers in cardiovascular and metabolic diseases. *BMJ Open*; **12:**e058890. doi: 10.1136/bmjopen-2021-058890.

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement **No 101095672**