4PCAN

D1.3 – Updated Data management plan

30 October 2024

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

Project number: 101104432
Project name: Personalized CANcer Primary Prevention research through Citizen Participation and digitally enabled social innovation
Project acronym: 4P-CAN
Call: HORIZON-MISS-2022-CANCER-01-01

Version number	0.6
Status	Final
Dissemination level	Public - PU
Due date of deliverable	30/04/2024
Actual submission date	30/10/2024
Project officer	Marianne Ines DA SILVA
Work package	WP 1 – Project Management
Lead partner	INOMED
Partner(s) contributing	INTERROBANG, ASE, PU, DNIPRO, INSA, IARC
	`Prevention and early detection implementation reseach' Cluster Partners
Authors	
Main author name	Adriana Boată (INOMED)
	Gabriel Hâncean (INOMED)
Reviewers	
Reviewer name	Marius Geantă (INOMED)
	Cosmin Badea (Ethics Advisor)

Statement of originality *This deliverable contains original unpublished work except where clearly indicated otherwise.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

Version Tracker

Date	Version	Author	Description
01/10/2023	0.1	Adriana Boată	Table of Contents and descriptions
11/10/2023	0.2	Viktor Semenov	DNIPRO data management
13/10/2023	0.2	Ana Santos	INSA data management
13/10/2023	0.2	Carolina Espina	IARC data management
15/10/2023	0.2	Bilyana Mileva	PU data management
19/10/2023	0.2	Răzvan Zamfira	INTERROBANG data management and data security
24/10/2023	0.2	George Stefan	ASE economic analysis data
29/10/2023	0.2	Cosmin Badea	Ethics Review
30/10/2023	0.3	Iulian Oană	DMP for Living-lab Activities Annex
31/10/2023	0.3	Adriana Boată	Review and harmonization
31/10/2023	0.3	Marius Geantă	Final Review
31/03/2024	0.4	`Prevention and early detection' Cluster Partners	Updated Data Management Plan for Prevention and Early Detection Cluster
2/10/2024	0.5	Gabriel Hâncean	Data Management in WP5 Living-lab Activities
30/10/2024	0.6	Marius Geantă	Final Review

Table of Contents

List of Tables

List of Abbreviations

Acronym	Description
CEE	Central Eastern Europe
CPP	Cancer Primary Prevention
ECAC	European Code Against Cancer
HE	Horizon Europe
MS(s)	Milestone(s)
M	Month
Del(s)	Deliverable(s)
WP	Work Package
GA	Grant Agreement
PMP	Project Management Plan
DMP	Data Management Plan

Executive Summary

The purpose of the Data Management Plan (DMP) is to outline data management strategies within the 4P-CAN project, ensuring compliance with GDPR guidelines. It encompasses data collection, processing, resource allocation, security, ethics, and updates throughout the project's lifecycle. This living document adapts to changing data needs, consortium policies, and external factors.

1. Introduction

This deliverable outlines the outcomes of the 4P-CAN Task 1.2 Data Management Plan, which complements the initiatives in Task 1.1 Management & Good Governance Protocols, Practices and Task 1.4 Quality Assurance, all situated within Work Package 1 – Project Management.

At the heart of the 4P-CAN project is the shared responsibility among citizens, professionals, and policymakers for primary cancer prevention. Spanning four years, 4P-CAN is determined to unveil the intricate interplay between individual choices and environmental contexts in determining cancer risk. These insights are critically mapped across national, community, and individual landscapes. Rooted in the European Code Against Cancer (ECAC) guidelines, which identify 11 lifestyle-related modifiable factors, the project interweaves diverse research avenues like economic, social, and behavioural sciences with cutting-edge technology. The ultimate vision is the crafting of tailored tools to champion primary cancer prevention, particularly in the Central and Eastern European context.

The insights gathered are not static; they inform personalized interventions, shaped within Living-labs, and influence the narratives of targeted communication campaigns. From addressing pivotal risk factors such as smoking, alcohol consumption, and environmental exposures, to piercing through the obstructions in policy execution and adherence to healthy choices, the project places an emphasis on Eastern European nations including, but not limited to, Romania, Bulgaria, and Ukraine.

This endeavour is driven forward by an ensemble of 18 institutions, hailing from 11 countries. While Eastern European nations form the core, contributions from Western EU nations like Ireland, and Belgium, and international organizations like IARC, enhance the project's breadth and depth. Notably, 4P-CAN marks its presence within the Cancer Mission cluster, specifically spotlighting 'Prevention and early detection' and making strides in Citizen Engagement initiatives.

2. Background

The 4P-CAN Data Management Plan (DMP) aligns with the guiding principles detailed in the Horizon 2020 guidelines for Open Access to Scientific Publications and Research Data, as well as the Horizon 2020 FAIR Data Management Plan. The DMP ensures compliance with relevant articles defined in the Grant Agreement (GA) of the 4P-CAN project.

2.1. Article 13.1 Sensitive Information

All involved parties must ensure the confidentiality of data, documents, or any other materials (in any form) labelled as sensitive in writing ('sensitive information')—both during the implementation of the 4P-CAN project and for at least the succeeding five years.

2.2. Article 13.2 Classified Information

Parties are obliged to handle classified information consistent with the relevant EU, international, or national laws on classified information. Deliverables containing classified information should follow special submission procedures agreed upon with the grant agency. Any subcontracting of tasks involving classified information requires explicit written approval from the grant agency.

Any disclosure of classified information to third parties, including participants in the project's execution, is prohibited without prior explicit written consent from the grant agency. Specific security protocols, if any, are detailed in Annex 5 of the GA.

2.3. Article 15.2 Data Processing by Beneficiaries

Beneficiaries of the 4P-CAN project are mandated to process personal data in harmony with the Agreement, following the EU, international, and national laws on data protection, emphasizing Regulation 2016/679.

2.4. Article 17 Communication, Dissemination, Open Science, and Visibility

Open Science: Open Access to Scientific Publications - Beneficiaries must ensure open access to peer-reviewed scientific publications that are derived from their results.

Open Science: Research Data Management - Beneficiaries have the responsibility to manage the digital research data created during the 4P-CAN project ('data') with utmost integrity, aligning with the FAIR principles.

2.5. General EU regulatory articles for 4P-CAN

2.5.1. General Data Protection Regulation (GDPR) - Regulation (EU) 2016/679: This is the primary regulation in the EU relating to the protection of personal data. It mandates that personal data should be processed lawfully, transparently, and for a specific purpose. It also gives rights to data subjects, such as the right to access, rectify, or erase their personal data.

2.5.2. Directive on open data and the re-use of public sector information (Open Data Directive) - Directive (EU) 2019/1024: This directive sets out the conditions for the re-use of public sector information, promoting the use of open data and its potential for innovation, growth, and transparency.

2.5.3. Directive 2002/58/EC (ePrivacy Directive): This directive concerns the processing of personal data and the protection of privacy in the electronic communications sector.

2.5.4. Regulation (EC) No 223/2009 on European statistics: Relevant if the project produces or uses official European statistics.

2.5.5. Directive 2013/37/EU amending Directive 2003/98/EC on the re-use of public sector information: This directive deals with the re-use of public sector information for purposes other than the initial one.

Data Transfer Outside the EU/EEA: GDPR contains specific provisions about transferring personal data outside the European Economic Area (EEA). In the context of Horizon 2020, if a project involves data transfer from a non-EU country to an EU member state, it is important to ensure that the non-EU country provides an adequate level of data protection. The European Commission has the authority to determine whether a country outside the EU offers an adequate level of data protection.

2.6. Partners' involvement in data generation and re-use

The 4P-CAN consortium consists of 18 organizations and 11 countries (Table 1). The table briefly describes the data that is going to be used/re-used, processed and collected by the Project's partner.

Table 1 4P-CAN Consortium and related data

Part. No.	Participant organisation name	Country	Data
1 (COO)	ASOCIATIA CENTRUL PENTRU INOVATIE IN MEDICINA (INOMED)	Romania	Organisational data; personal data; socio-demographic data; behavioural data; attitudes, perceptions, and opinions data; environmental data;
2	DNIPRO STATE MEDICAL UNIVERSITY (DNIPRO)	Ukraine	Data about air pollution from Ukrainian citizen air quality monitoring network Eco City, Public organisation Save Dnipro, Municipal institution "Centre of Ecological Monitoring", Dnipro Regional Centre of Hydrometeorology, State institution "Dnipropetrovsk Regional Centre for Disease Control and Prevention of the Ministry of Health of Ukraine". Data about cancer epidemiology from Dnipro Cancer Registry. Data about opinion and behaviour concerning ECAC recommendations in 4P-CAN countries collected during subcontracting Kyiv International Institute of Sociology
3	EUROPEAN NETWORK FOR SMOKING PREVENTION (ENSP)	Belgium	Policy data; literature data on cancer primary prevention strategies; stakeholders' opinions; impact indicators.
4	ASSOCIATION FOR	North	Policy data; stakeholders' opinions.

	HELP AND SUPPORT OF PATIENTS WITH GYNECOLOGICAL DISEASES (FEMINA M)	Macedonia	
5	ACADEMIA DE STUDII ECONOMICE DIN BUCURESTI (ASE)	Romania	Socio-economic data for costs calculation; national-European health expenditure/investments data; literature data;
6	BULGARIAN JOINT CANCER NETWORK (BJCN)	Bulgaria	Policy data; stakeholders' opinions.
7	PLOVDIVSKI UNIVERSITET PAISIY HILENDARSKI (PU)	Bulgaria	Policy data; stakeholders' opinions; personal data; socio-demographic data; behavioural data; attitudes, perceptions, and opinions data; environmental data.
8	ASOCIATIA EUROCOMUNICARE (EUROC)	Romania	Policy data; personal data; socio-demographic data; communication trends data;
9	FUNDATIA RENASTEREA PENTRU EDUCATIA, SANATATE SI CULTURA (FR)	Romania	Policy data; stakeholders' opinions.
10	DIRECTORATE FOR HEALTH AND SOCIAL ASSISTANCE (DGAMS)	Moldova	Policy data; stakeholders' opinions.
11	INSTITUTO NACIONAL DE SAUDE DR. RICARDO JORGE (INSA)	Portugal	Data from the scoping review of performance indicators obtained through public scientific databases and organizations repositories. Data from the interviews and questionnaires applied to national stakeholders of the countries within the consortium. Data from focus groups about COVID-19 and CPP synergies
12	MONTENEGRIAN SOCIETY OF MEDICAL ONCOLOGY (CUMO)	Montenegro	Policy data; stakeholders' opinions; environmental data.
13	FONDAZIONE POLICLINICO UNIVERSITARIO AGOSTINO GEMELLI IRCCS (FPG)	Italy	Policy data; stakeholders' opinions.
14	ECHALLIANCE COMPANY LIMITED BY GUARANTEE (ECHA)	Ireland	Personal data; organisational data.
15	ALL CAN INTERNATIONAL (ALL.CAN)	Belgium	Policy data; stakeholders' opinions.
16	RZ INTERROBANG GRAPHICS (INTERROBANG)	Romania	Aggregated data intended for public dissemination coming from the other project partners: e.g., national surveys collecting socio-demographic data; behavioural data; attitudes, perceptions, and opinions;

			environment data;
17	INTERNATIONAL AGENCY FOR RESEARCH ON CANCER (IARC)	France	Policy data; personal data; socio-demographic data; behavioural data; attitudes, perceptions and opinions data; environmental data.
18	University of Aveiro	Portugal	Policy data; personal data; socio-demographic data; behavioural data; attitudes, perceptions and opinions data; environmental data.

3. Data Summary

The re-use and generation of data within the 4P-CAN project is pivotal in comprehensively understanding the landscape of cancer prevention and in devising strategies and interventions that are well-informed and impactful. The data's utility extends beyond the confines of the project, serving as a valuable resource for various stakeholders in the realm of public health and cancer research.

Given the interdisciplinary and extensive nature of the project, it will re-use existing policy, demographic, epidemiological, sociological, and environmental data related to cancer. The existing data re-used will assist in establishing baseline values and in the contextual analysis essential for implementing strategies, refining methodologies, and evaluating the impact of interventions.

Table 2 Activities involving data per WP

WPs	Types of data / activities (Q – qualitative; q - quantitative)	Data format
WP1 Project Management		
Management and coordination	Q: Organizations' information (Legal; Financial); Consortium members contact data…	.docx; .xlsx; .pdf; .jpeg; .png; .ppt
Monitoring and assessment	Q-q: Progress reports; Meetings and events data (minutes, agendas, summaries)	.docx; .xlsx; .pdf;
Quality assurance	Q: External experts' data; Ethics committee data	.docx; .xlsx; .pdf;
WP2 Multi-level Assessment of Cancer Risk Factors Preventive Legislation		
Quantitative research (scientific and grey literature)	Policy data; Impact indicators.	.docx; .xlsx; .pdf;
Qualitative research (interviews; focus groups)	Stakeholders' opinions and knowledge	.docx; .xlsx; .pdf;
Stakeholders mapping (including National Standing Committees)	Q-q: Stakeholders' interactions; contacts and expertise	.docx; .xlsx; .pdf;
Environmental pollution	Q-q: Air quality data; military generated pollution data; radon data; cancer epidemiology data.	.csv; .docx; .xlsx; .pdf;
Prevention and Health Curricula Mapping	Q-q: reports; interviews and surveys	.docx; .xlsx; .pdf;
WP3 Socio-economic and informational determinants of cancer prevention		
Direct and indirect costs modelling	Q-q: Costs and budgets; expenditures; healthcare facilities data publicly available (personnel; medications; supplies); policy data (cancer programmes; screening; early-diagnosis data; treatment guidelines); workhours; individual health spending; alcohol – tobacco – junk food prices etc.; epidemiological cancer data.	.csv; .docx; .xlsx; .pdf;
Behavioural study with monetary incentives (n=100 students)	Q-q: literature review; lifestyle and socio-demographic data;	.docx; .xlsx; .pdf; .ppt;
Informational analysis	Q-q: Social Media trends analysis data	.csv; .docx;

		.xlsx; .pdf
COVID-19 – CPP synergies	Q: focus groups	.docx; .xlsx; .pdf
WP4 Targeted public health communication campaigns		
Measuring knowledge, attitudes, motivations, behaviours, and implementation issues concerning ECAC recommendations	Q-q: national surveys collecting socio-demographic data; behavioural data; attitudes, perceptions, and opinions; opinions/perceptions on implementation of cancer prevention interventions	.csv; .docx; .xlsx; .pdf
Communication campaigns	Q-q: literature review; stakeholders' interviews on implementation research outcomes; citizens' knowledge, intentions to change behaviours and adoption of recommendations, and implementation research outcomes	.docx; .xlsx; .pdf
Data visualization (CPP Dashboards)	Q-q: data generated as part of different project WPs as well as demographic, epidemiological, sociological, and environmental data related to cancer from reliable outside entities and datasets.	.csv; .Json; .xml
WP5 Living-Labs for primary prevention of cancer		
Design and indicators	Q-q: Stakeholders mapping; focus-groups; environmental data; infrastructure data; healthcare access data; cancer epidemiologic data.	.docx; .xlsx; .pdf
Personal Networks Analysis	Q-q: socio-demographic data; attitudes, perceptions, and opinions data; peer-network data; personal data – **See Annex 1 for all the details**	.docx; .csv; .xlsx; .pdf
Environmental assessment	q: Air quality; air pollution; water pollution; radon levels; mining data; soil pollution; weather; climate	.docx; .xlsx; .pdf
Data integration	Aggregated data intended for public dissemination coming from the other project partners	.csv; .Json; .xml
WP6 Multi-stakeholders' co-creation of CPP policy recommendations		
Innovation recommendations	Summary of all the data above and experts' opinions	.docx; .pdf
Inequalities recommendations	Summary of all the data above and experts' opinions	.docx; .pdf
Ukraine recommendations	Summary of all the data above and experts' opinions	.docx; .pdf
WP7 Dissemination, Evaluation, Education, Communication		
Dissemination and Communication activities	Q-q: Press releases; scientific advancements in CPP; public events; interviews; podcasts; webinars; reports	.docx; .xlsx; .pdf
Education Activities – 4P-CAN Academy	Educational materials; Project's findings and scientific articles; interviews; podcasts; webinars; summer-school sessions	.docx; .xlsx; .pdf

3.1. Re-use of Existing Data

The 4P-CAN project will gather, generate and re-use data related to cancer primary prevention, including lifestyle data (as presented in Table 2 above), from international reputable sources (like the World Health Organization, Organisation for Economic Co-operation and Development), national databases and governmental resources, literature sources - research studies, scientific publications (English and local languages), grey literature and policy reports.

Re-use of existing data can provide baseline or contextual information, facilitating a more comprehensive understanding of the current landscape and allowing more informed development of project strategies and interventions.

Data provided to 4P-CAN partners for analysis may not be re-used unless permission for data transfer to third parties is given. Any consideration of discarding re-used data would be due to issues with data relevancy, reliability, or compatibility with project objectives.

3.2. Types and Formats of Data

The project will generate qualitative and quantitative data, including epidemiological data, survey responses, workshop feedback, and policy recommendations. Formats may include structured databases, text documents, spreadsheets, and multimedia formats for educational materials.

- Qualitative Data: This includes feedback from workshops, minutes of consortium meetings, responses from engagements/interviews with stakeholders, policy recommendations, and evaluations.
- Quantitative Data: This comprises epidemiological statistics, surveys results, literature review, and numerical data from the evaluations.
- Educational Material: Multimedia content, text documents, and online resources generated for the Dissemination and Communication Activities, as well as 4P-CAN Educational Academy.

3.3. Purpose of Data Generation or Re-use

The purpose is multifold, on three levels of analysis – macro level (national and European context surrounding a citizen); meso level (the community someone lives in); micro level (the individual's closest network; attitudes, perceptions and behaviours) - to gain insights into the current state of cancer primary prevention, to understand the networks of healthy lifestyle promoting behaviours and what is the role of different stakeholders, while evaluating their feedback and engagement level, to disseminate knowledge, and to formulate policy recommendations.

The generated or re-used data is integral to achieving the project's objectives of developing strategies for dissemination, evaluation, education, and communication of 4P-CAN results and maximizing their impact. In summary:

- Contextual Understanding: Existing data will be re-used to understand the present context of cancer prevalence, incidence, and prevention strategies.
- Strategic Development and Refinement: Both generated and re-used data will inform the development and refinement of dissemination strategies, education material, and policy recommendations.
- Evaluative Insight: Generated data will be critical in evaluating the impact and efficacy of the implemented interventions, thereby aligning them closely with the project's objectives of maximizing results' impact.

3.4. Expected Size of the Data

The expected size of the data is variable and will depend on the extent of surveys, workshops, and engagements conducted. It may range from smaller datasets of structured survey responses to larger, more complex datasets of epidemiological information and workshop feedback.

Formats: Data will be stored in structured relational databases, text documents (.docx, .pdf), spreadsheets (.xls, .csv), and multimedia formats (.mp4, .mp3) as appropriate.

3.5. Origin/Provenance of the Data

Generated data will originate from the various work packages and activities conducted within the project, including surveys, workshops, stakeholder engagements, and educational academies. (See Table 2)

Re-used data will be sourced from reputable external entities and databases, with appropriate acknowledgment and citation.

3.6. Data Utility

Beyond the project, the data may be useful to a wide range of stakeholders including researchers in cancer prevention, public health policymakers, healthcare providers, educational institutions, and other entities involved in cancer research, policy formulation, and public health promotion.

4. FAIR data

4.1. Making Data Findable, Including Provisions for Metadata

All results, deliverables, publications, and reports produced within the project will be made available to all consortium partners in the project repository (i.e., Microsoft Teams and SharePoint), where results can be identifiable and locatable by means of a persistent Uniform Resource Locator (URL).

All public project results will be published on the project website and will be linked to an URL and keywords.

All public results and other relevant information will also be made available on the Open Access platforms such the European Repository Zenodo and the Horizon Results Platform (HRP) to enhance results visibility.

The Zenodo repository allows for the deposition of all kinds of digital content: publications, data, software, multimedia among others. All published data will be provided with a Digital Object Identifier (DOI): Once published, datasets remain fixed over their lifetime, while the metadata can change.

The HRP, free platform hosted under Funding & Tenders Portal, connects EU-funded research and innovation results with interested third parties: investors, policymakers, research centres, partners, and other stakeholders, to accelerate the chances of successful matchmaking.

Other repositories considered:

- European Data Portal[1]: A central gateway to public data from European countries. It harvests metadata from public sector portals and then directs users to the original data.
- European Social Survey (ESS) Data Archive[2]: Contains complete data and documentation which is available freely to all users. It includes the ESS Cumulative File which combines data from all rounds.
- EUDAT[3]: Offers a suite of services to help researchers manage their data, as well as storing and sharing datasets. Also, it is compliant with EU regulations and specifically caters to European researchers.
- **Open Science Framework (OSF)[4]**: Provides an infrastructure for researchers to store datasets, code, and other research artifacts. It also supports version control and offers persistent DOIs.

Data and Metadata Standardization: The data generated in 4P-CAN will be accompanied by

[1] https://data.europa.eu/en
[2] https://www.europeansocialsurvey.org/data
[3] https://www.eudat.eu/
[4] https://osf.io/

metadata to ensure it is easily findable. Metadata will be structured according to internationally recognized standards, taking into consideration the specific nature of the project (e.g., health data, survey data).
This will include general metadata (title, description, keywords, DOI unique identifier, publication date, language), provenance metadata (project name, data creators/contributors, license), technical metadata (file format, data type, version and any software requirements for accessing or interpreting the data), content specific metadata (sample size, demographics, variable/columns definitions, methodologies used to gather the data), accessibility metadata (any relevant access restrictions, repositories)

Data Identification: Each dataset will be given a unique identifier, ensuring traceability and accessibility. We plan to use the DOI (Digital Object Identifier) system or a similar persistent identifier system.

Keyword Optimization: Datasets will be labelled with relevant keywords linked to cancer prevention, individual and environmental factors, and targeted European regions.

4.2. Making Data Openly Accessible

Data Repository: We'll utilize trusted data repositories that align with the project's data requirements. For sensitive data, access controls will be implemented to protect participant privacy.

Data Access Mechanism: Data will be made available through a user-friendly portal, ensuring that researchers and the public can access the findings with ease. The project portal will facilitate intuitive ways of downloading filterable aggregated project data as well as exploring it visually through an online dashboard. Whenever necessary, data access requests will be streamlined through a straightforward request and approval mechanism.

Data Accessibility Constraints: While the aim is to make the data as open as possible, some datasets, especially those containing personal or sensitive information, may have restricted access. Justification for such restrictions will be clearly provided.

4.3. Making Data Interoperable

Interoperable Formats: Data will be stored in formats that ensure interoperability, facilitating easy data sharing and integration with other studies.

Vocabularies and Standards: Where relevant, standardized vocabularies will be employed to ensure the data is understood and can be utilized across different platforms and studies.

Data Integration: As the project seeks to understand the interplay between individuals and their environment, datasets will be designed in a way that they can be integrated, offering a holistic view.

Results, deliverables, publications, and reports will be stored in a secure access controlled and regularly backed up Microsoft Teams / SharePoint project repository, as well as an online encrypted structured database which will facilitate querying publicly accessible data through the online project portal. In addition to this, storing data in third party Open Access repositories, such as Zenodo, will also contribute to the long-term preservation and curation of the datasets.

4.4. Increase Data Reuse (Through Clarifying Licenses)

Licensing: All datasets will have clear licensing information, which will be chosen to maximize data reuse while respecting any necessary restrictions. Licenses like CC BY (Creative Commons Attribution) might be considered for datasets that can be openly shared.

Data Preservation: Data will be preserved for a minimum of 5 years. Preservation practices will ensure that the data remains usable beyond the lifespan of the 4P-CAN project.

Encouraging Reuse: By ensuring that the data is findable, accessible, and interoperable, and by providing clear licensing, the project will inherently encourage and facilitate data reuse by the wider scientific community and stakeholders. A dedicated section on the project's website is already in place and promoted as such to store the project's outcomes.

5. Allocation of resources

The 4P-CAN Data Management Plan current version is developed by the coordination team in close collaboration with the key project's partners (WP and task leaders). INOMED will see to its implementation during the 4-year duration of the project.

6. Data security

The 4P-CAN project is committed to ensuring the secure management of all collected or generated data. To achieve this, the project will implement appropriate measures to prevent any potential data breaches. The project's data repository is hosted on Microsoft Teams and Microsoft SharePoint, an independent platform accessible to all project partners. Each consortium member is responsible for handling data in compliance with data protection regulations, such as GDPR, by utilizing secure methods such as private servers or trusted cloud services.

To mitigate the impact of data loss, regular backups will be performed, taking into account data change frequency and importance. In addition to automatic platform backups, physical backups to external drives will occur every three months starting M6 of the project. These backups will be securely stored using various storage options, including external hard drives and trusted cloud services, ensuring data protection and accessibility.

Access to restricted data will be limited to authorized project partners. In the event of a personal data breach, the responsible project partner will promptly notify the relevant supervisory authority and affected individuals within 48 hours of becoming aware of the breach, if feasible. Detailed records of data breaches will be maintained, including breach details, impact, and remedial actions taken.

The project will implement robust identification and authentication access controls to protect the data gathered or produced during 4P-CAN, especially personal information. Microsoft Teams and SharePoint, the project's internal repository, enforces security and privacy mechanisms, including data encryption (in transit and at rest), multi-factor authentication (MFA), secure guest access, compliance with industry standards and regulations, information protection, threat protection, admin controls, privacy controls, secure file sharing, and regular security updates.

These measures collectively ensure the secure management and protection of project data, safeguarding it from potential breaches or unauthorized access.

Additionally, in order to safeguard users' privacy on the project's website, detailed privacy policies will be established to transparently explain how the platform collects, processes, and utilizes personal information. These policies will also cover the security measures in place, outline users' rights, and provide information on the adopted cookies policy.

7. Ethics

As 4P-CAN Project covers Ethics issues as described in the EU Grants Ethics self-assessment guideline, periodic Ethics evaluation reports will also be developed and submitted. For more information, see D8.2 and D8.3 – Ethics Reports.

8. Unified Data Management Approach: mapping strategies across the Prevention and Early Detection Cluster

All partner projects included in the "Prevention and early detection" Cluster of the Mission Cancer initiative recognize the relevance of adopting proper strategies for data management during the course of the research. The partner projects recognize that making research data "as open as possible, as closed as necessary" by sharing them with the research community according to the FAIR principles is a practice of great value in ensuring research quality, reproducibility and transparency. Each project will either reuse or generate new data for the purpose of its activities, as defined in the respective Data Management Plan.

Starting from the acknowledgement of the differences among the types of data that each project will manage, Cluster partners started to meet in order to recognize and to bring together all the possible similarities and points of contact in terms of management strategies and key attention points.

The partners are aware that working on personal data collected from enrolled subjects or derived from pre-existing datasets identifies privacy regulation (e.g. GDPR - Regulation EU 2016/679) compliance as the major constraint directing the selection and implementation of data management strategies. Any processing of personal data will respect privacy by design and by default principles. To ensure the secure management of sensitive data, during the active phases of research, data storage will be performed only in secured local solutions or in GDPR-compliant cloud infrastructures. Data access to unprocessed data, which may lead to subject identification, will be limited to the minimum number of researchers needed and will be subject to access control via identification, data encryption, and pseudonymization or anonymization will be considered according to the purpose of the research.

Moreover, according to the principle of making data "as open as possible, as closed as necessary", strategies for the long-term preservation of data will be evaluated, taking into account the possibility of full anonymization prior to deposit; for data that can't be fully anonymized, deposit in repositories equipped with controlled access systems (e.g. with the control of a dedicated Data Access Committee) and/or the deposit of dataset description and metadata in a standard format will be evaluated, while taking measures to prevent identification of individuals at all times. The partners will adopt proper strategies in order to manage their data in accordance with the FAIR principles, hence ensuring they are Findable, Accessible, Interoperable, and Reusable. To this extent, data will be organized in datasets which will be (if and when appropriate) deposited in trusted repositories where they will be assigned a Persistent Identifier (PID), such as Digital Object Identifiers (DOIs), and will be associated to rich descriptive Metadata to grant their findability. As previously stated, data Accessibility will be defined based on privacy protection considerations. Data interoperability will be addressed by collecting adequate documentation to facilitate interpretation and re-analysis, which will be shared alongside the data. When possible, standard vocabularies (e.g. SNOMED CT) and data formats (e.g. OMOP CDM, HL7 FHIR) will be adopted. Data Reusability will be granted by associating an open data license (CC 0 or CC-BY, as per Grant Agreement requirements) to each deposited dataset.

The following table summarizes the cross-cutting topics that were identified across the different projects.

Table 3 Main Data Management Areas for Cluster Projects

	CO-CAPTAIN	CPW	ONCODIR	PIECES	4P-CAN	PREVENT
Personal data management	X	X	X	X	X	X
Qualitative data collection	X	X	X	X	X	X
Quantitative data collection	X	X	X	X	X	X
Standardized methods for data collection	X	X	X	X	X	X
Data anonymization for open data deposit	X	X	X	X	X	X
Standardized formats for data collection and analysis	X	X	X	X	X	X
Standard vocabulary for metadata collection	X	X	X	X	X	X
Secure data storage in collaborative environments	X	X	X	X	X	X
Data quality monitoring strategies	X	X	X	X	X	X
Data repository for data deposit	X	X	X	X	X	X
PID attribution to data	X	X	X	X	X	X
License attribution to data	X	X	X	X	X	X

8.1. Coordination of data management activities across the Cluster

The CPW project will take responsibility for coordinating the data management activities and sharing good practices across the "Prevention and early detection" Cluster. This work will be carried out by a Data Management Working Group which will gather one or more reference people from each partner project, ideally identifying the reference person among those responsible for data management actions (e.g. DMP drafting). The Working Group will meet periodically (e.g. via video conferences, at the cluster meetings) and upon partner request to share and discuss updates on data management strategies and to collaborate on the design of common strategies as well as on solutions to any issue that may arise. The discussions of the Working Group are expected to influence and benefit data management within each single project.

Common elements of the DMP will be identified, as well as aspects specific to each project. Recommendations will be drafted to optimize data management with the purpose of increasing comparability and interoperability across the projects included in the Cluster. These recommendations will be shared with the PIs of the other projects in the Cluster and will be discussed and adopted during one of the early Cluster meetings.
The main objective of the working group and of this coordination work is to collect and compare strategies for data management in the Cluster projects in order to:

- **Identify differences and points of contact among the data managed within the projects and the management strategies adopted**: the dialogue to implement this action has already started, allowing the drafting of the common chapter to be inserted in the DMPs of the partner projects and the identification of possible collaboration points.
- **Elaborate a strategy to further implement the possibilities for collaboration on specific identified topics:** during the first online meeting, two topics of future collaboration were identified. The first proposal is to elaborate a strategy to map the possible overlapping of variables that are analyzed within the different projects and assess potential options for standardization of the information. The second one is to evaluate the possibility of creating a Cluster community in a trusted repository to be selected for the deposit of data from the different projects.

As coordinator of this activity, CPW will also act as a consultation point for partner projects. The support is envisioned to take different forms, mainly as email exchanges, and will regard specific data management topics that are carried out in the CPW projects, such as the defined eCRF questionnaires with the related variables or the strategies for the mapping of the variables on the OMOP standard. Research sharing will be taken into consideration when appropriate.

The activities related to the coordination of data management and sharing within the Cluster will be conducted at the University of Bologna, CPW Coordinator and leader of WP1, which includes data management, in collaboration with WeDo, leader of WP7, which includes dissemination.

The Unified Data Management Approach constitutes a living document and may be adapted during the course of the cluster.

9. Conclusion

The 4P-CAN Data Management Plan (DMP) is designed to ensure the responsible handling of data collected, processed, and/or generated during the project's lifecycle, all while aligning with the FAIR scheme. This DMP provides a comprehensive overview of the key data management processes within the 4P-CAN project, including data collection, processing, and generation, in full compliance with GDPR guidelines. It offers insights into the types of data involved in the project, the specific objectives guiding the collection, processing, and generation of each dataset, the allocation of resources, data management responsibilities, data security protocols, and ethical considerations.

It's important to note that the 4P-CAN DMP is a dynamic document that will undergo periodic updates throughout the project's duration. These updates will be triggered by new data-related activities, changes in consortium policies, shifts in consortium composition, or external factors that impact data management within the 4P-CAN project.

Within the 4P-CAN project, we are working to establish a top-down strategy for managing personal data, ensuring that all partners adhere to a common set of minimal requirements. While some partners may already have more detailed systems in place, this approach serves as a guideline for those with less experience, particularly in Eastern Europe, where data collection practices may vary.

As part of our comprehensive data management plan, we plan to include a specific agreement within the Consortium Agreement that also governs the use of shared data among partners. This agreement will also outline that data should not be used for commercial purposes outside the scope of the planned project implementation and sustainability. In cases of alternative data usage, we will ensure clear references to the project and partner involvement.

Additionally, the Consortium Agreement will encompass a clear minimum set of requirements for handling personal data and managing intellectual property and project data, providing a unified framework for data management within the project.

For the sake of transparency, we would like to declare that we have also utilized the AI tool ChatGPT by OpenAI for text proofing, aiding our non-native speakers in improving the clarity and precision in their communication.

Annex 1 Data Management in WP5 Living-lab Activities

1. Data Summary

This report describes the Data Management Plan linked to the establishment of the Romanian Living Lab. The living lab is part of the 4P-CAN Horizon Europe project (HORIZON-MISS-2022-CANCER-01, Research & Innovation), focused on personalized research in primary cancer prevention.

A cohort of individuals will be identified and selected through ethnographic fieldwork methods and link-tracing sampling. Variables such as sex, age, and education will be used in participant recruitment. These recruitment variables should provide access to heterogeneous subjects in terms of their experiences and personal networks. By using this mix of variables, we expect the participants to have different profiles within the cohort (e.g., young men with low education, older women with low education, adult women with higher education, etc.). The plan involves conducting observations in one pilot stage and three waves, meaning each panel participant will be interviewed four times maximum. In these waves, surveys will be conducted. These waves are spaced over a period of 18 to 24 months, the interval between them taking into the intervention or information events happening in between. Through these surveys, opinion and perception-based data about respondents' and their social contacts will be collected. The research is geared by the research design. Besides personal network data of panel research subjects, other data can include outputs from interviews / NetMapping exercises with public health experts and radon levels in the living lab.

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.

The team will re-use data publicly available from official institutions (e.g., Eurostat, Romanian National Institute of Statistics) for documentation, sampling, comparison of observed phenomena – e.g., available statistics about cancer risk factors, among others.

The research team will mainly use primary data collected from the Living lab study. For papers and other research outputs, primary data are supplemented with existing data (e.g., National Institute of Statistics, Eurostat, etc.) which will provide context information for our studies.

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

We will use data in CSV format. Other possible formats are PDF, for codebooks, and JPEG for graphs and plots resulting from the analysis. Other formats can be dependent on publication requirements (e.g., PNG files for plots).

During the data collection process, the research team will gather both qualitative (responses to open-ended questions) and quantitative data (responses to close-ended questions), and the study design is based on social and personal network analysis (observational or survey data). The main data will be collected through a questionnaire, and it will comprise information as follows:

- Social demographic characteristics (e.g., sex, age, last level of education, etc.) —:> Mostly quantitative. Qualitative information will be re-coded into quantitative data as response categories.
- Different lifestyles and different eating habits (e.g., salt intake, alcohol consumption, physical activities, etc) —:> Mostly quantitative. Qualitative information will be re-coded into quantitative data as response categories.
- Perceived health status (e.g., pre-existing health conditions) —:> Mostly qualitative re-coded into quantitative data as response categories.
- Information about individuals with whom the respondent frequently interacts or communicates (over 18 years old) and the social relationships between them. —:> Both qualitative and quantitative. Qualitative information will be re-coded into quantitative data as response categories.

Other data can include interviews with experts who, through NetMapping exercises, will create layered networks about the stakeholders involved in influencing policy about primary cancer prevention, and measurements of radon levels inside the living lab. These data will also be available in formats that are open access (e.g., CSV format).

The data will be shared as part of the open science framework, to ensure replication and the possibility of the scientific community to build upon the results obtained in the 4P-CAN Living Lab study. Consequently, only the data underlying the specific scientific publications will be shared, with respect to anonymization rules which ensure that study participants or other experts cannot be directly or indirectly identified.

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

The purpose of the study and, implicitly, of the data generation is to develop primary cancer prevention measures by analysing risk factors such as smoking, alcohol consumption, physical inactivity, excess body weight, and poor dietary habits, as well as infections with HPV, HBV, and pollution. The general research objectives are for instance:

a) Understanding the role of socio-demographic factors and human behaviour in the health habits and lifestyle of the study participants.
b) Using empirical data to inform primary prevention strategies and define effective communication methods.
c) Conducting interviews with study participants to enhance the understanding of available social resources within personal networks.
d) Analysing empirical data using statistical modeling solutions such as ERGM and multi-level network analysis to explain the impact of networks on health and understand the explanatory factors of social interactions.
e) Performing a comparative analysis of different classes of personal networks to generate individual profiles and identify similarities and differences among study participants.
f) Evaluation of statistical models for hypothesis testing regarding network contagion mechanisms and social network generation factors.
g) Mapping the relevant stakeholders and the relations between them at local, national, and European level employing the NetMapping methodology through the Net-Map toolbox.
h) Measurements of radon levels inside the living lab.

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

An estimation of the data size cannot be provided at this moment. The primary data format will be in a CSV format, with additional PDF and JPEG files (where needed). We expect that they will not exceed 50 GB. We take 50 GB as a reference, given that we intend to use Zenodo as the main repository for data and other material that accompany them, where 50 GB is the maximum storage limit of Zenodo for a record. Additionally, we don't expect that an individual record will exceed 100 files, which is the files limit for an individual Zenodo project. However, Zenodo gives users the possibility to store more than 100 files inside a single record, provided that they are stored as archived folders. If needed, different datasets will be stored on different projects on Zenodo, yet we expect that it will not be the case.

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

This study aims to increase the level of understanding regarding the social risk factors of cancer on three levels: macro (national level), meso (community level), and micro (individual level). As part of the project, **the Living lab study** will allow collecting data for the testing of disease prevention interventions. A cohort of individuals is being identified and selected through ethnographic fieldwork methods and link-tracing sampling. Variables such as sex, age, and education are utilized in the recruitment of panel participants. These recruitment variables are expected to provide access to participants with diverse experiences and personal networks. By employing this mix of variables, it is anticipated that the participants will exhibit distinct profiles within the cohort (e.g., young men with low education, older women with low education, adult women with higher education, etc.). The study is designed to encompass observations in one pilot stage and three waves, entailing interviews with each panel participant. These three waves are spaced over a period of 18 to 24 months, the interval between them taking into the intervention or information events happening in between. As such, all the data collected using the Personal Network Analysis framework is based on respondents' perceptions and opinions about their behaviour and the behaviour of their social contacts.

Other data related to the Living lab study can be produced with: a) public health experts, whose role is to map the network dependencies between various stakeholders involved in primary cancer prevention; b) electronic devices capturing the radon levels inside the living lab. Both data generation processes are non-intrusive.

To whom might your data be useful ('data utility'), outside your project?

The results of our studies and outputs can be valuable to various audiences:

1) Other researchers who employ Social or Personal Network Analysis for the study of health behaviours.
2) Citizens, providing them with access to a comprehensive diagnosis of their health status and that of those around them, as well as a deeper understanding of how their close network can influence their own health.
3) Policymakers, authorities, and healthcare professionals granting them access to insights into the barriers hindering the adoption of recommendations proposed by the European Code Against Cancer. Additionally, this information can aid in developing more effective cancer prevention strategies and educating patients about social risk factors and preventive measures.

4) Other researchers can benefit from the project's methodology, as it can be replicated on different populations, allowing for further research and knowledge advancement in the field of cancer prevention and social risk factors.
5) Partners from the "Prevention and Early Detection" cluster.

2. FAIR data

2.1. Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

Yes. Anyone included in the (co)authorship structure, whether they come from the medical field or the social sciences, can be identified through their Open Researcher and Contributor Identifier (ORCID) code. The ORCID code functions as a personal identification number that links the researcher to their professional achievements and sets them apart from any other researcher with the same name. The ORCID identifier(s) will be linked to the (co)authorship structure of the Zenodo project, and all the published works will be automatically incorporated into the ORCID record through Crossref. What is more, not only will the researchers receive a persistent identifier, but also their published work. In this case, all the articles as well as the data repositories will receive a Digital Object Identifier (DOI). A DOI name or code is a unique identifier used in order to better track any kind of object, whether it's a physical item, a digital file, or an abstract concept.

All the results of the Living Lab study will be deposited in an Open Access Repository. The main repository will be Zenodo, as they have a partnership with the European Union, facilitating the "EU Open Research Repository". As such, Zenono is one of the open data repositories that can generate DOIs (along other Unique Persistent Identifiers – e.g., DOI codes of preprints or published articles). However, there may be cases where the editors of a certain journal (depending on the type of access: gold or green) ask for other options for data repositories (e.g., Open Science Framework (OSF), figshare, Dataverse, etc.) as long as these options are compliant with EU regulations regarding data storage.

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.

We will use repositories that index metadata and allow indexing and harvesting through open access tools. Zenodo, the main repository that will be used, indexes rich project metadata using community-developed open standards and tools, offering information such as description, funding information, license, DOI or tags (keywords), among others. As it is Besides project metadata, Zenodo allows the upload of file metadata (i.e., metadata for individual files inside a project). Alongside the metadata format on Zenodo, we will use the appropriate metadata standards for the information inside the data files according to publication specificities (e.g., Data Documentation Initiative (DDI) standard, which is a metadata standard for data produced through surveys in social or health sciences, or the Observational Medical Outcomes Partnership (OMOP) standard for organizing data inside the Prevention and Early Detection cluster).

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

For the metadata, a range of keywords will be included. These keywords will be linked to diverse terminologies employed within the research, encompassing terminology from fields such as social / personal network analysis and the medical domain. Furthermore, they will be

associated with terminologies in the statistical realm, including examples like Exponential Family Random Graph Models (ERGMs) and Relational Hyper Event Models (RHEMs).

Will metadata be offered in such a way that it can be harvested and indexed?

Yes. Zenodo indexes project metadata using community-developed open standards and tools, offering information such as description, funding information, license, doi or tags, among others. Zenodo metadata is compliant with the DataCite Metadata Schema. For both individual projects and collections, the metadata can be harvested using the OAI-PMH and REST API protocols, that are open and free protocols for web information retrieval.

2.2. Making data accessible

Repository:

The main repository that will be used for data storage is the Zenodo repository. Depending on journals' requirements and the type of access, other options may be considered (e.g., Open Science Framework (OSF), figshare, Dataverse, etc.) as long as these options are compliant with EU regulations regarding data storage.

Will the data be deposited in a trusted repository?

Yes. Zenodo aligns with the principles of open science, and it offers a platform for each researcher to share their work (e.g., preprints, concepts, databases, etc.). It can integrate the associated information other online data repositories and gives control of the uploaded versions of a specific work (in this case, the data uploaded is easy to reproduce). It also gives control access to their projects and data, allowing for both public and private sharing. Moreover, Zenodo has been chosen to be the main "EU Open Research Repository", giving researchers the possibility to create both individual and community projects inside the repository.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Given that Zenodo provides free services, no institutional arrangements have been made between the platform and the project team members, as they are not necessary. 4P-CAN, as an EU Horizon project, will adhere to using Zenodo for individual and community projects, both associated with the "EU Open Research Repository" community.

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

All the repositories that will be used (irrespective of whether it is OSF, Zenodo, or figshare) will assign a unique identifier. The repository will resolve the identifier to the corresponding digital object, providing a reliable and persistent link to the data. Additionally, Zenodo offers unique DOI codes for each version of a repository and a unique DOI code per repository that always resolves to the latest version.

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.

The different types of data generated in the Living lab study are open by default, with few exceptions. These refer to any type of identification data of the respondents (e.g., name and surname of the respondents, date of birth, etc.). Human subject data allowing for direct identification will not be shared. The information contained in the informed consent will also not be provided, given the privacy, security, and ethical implications. Additionally, data linked with publications will respect anonymization rules regarding sharing human subject data to not allow for indirect identification. It will offer other researchers and interested parties the possibility to reproduce our results without disclosing the identity of research subjects.

Other data, such as the network mapping of stakeholder or radon level measurements, will also be publicly available, following the same regulations regarding anonymization (if necessary) and adherence to FAIR principles.

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.

One of the aims inside the 4P-CAN project is to promote open science. All research articles produced in this project will have an open-access regime (preferably Gold Open Access). Thus, there is no embargo on the published data, as they will accompany research articles when they are published or submitted for review with an already attached DOI given by an open access repository.

Will the data be accessible through a free and standardized access protocol?

Yes. The data retrieval will be carried out without restricting access. A free protocol will be used, such as HTTP(S) or FTTP, both of which correspond to the FAIR principles (findability, accessibility, interoperability, and reusability). Zenodo relies on OAI-PMH and REST API as available APIs, that are open and free, and that can make HTTP requests.

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

The restrictions regarding available datasets encompass the following: only those datasets attached to an article under review will be made available online (on specific repositories) and data shared inside Prevention and Early Detection cluster. Each study output (scientific articles) will be accompanied by an anonymized dataset, facilitating data replicability and reuse by the scientific community.

How will the identity of the person accessing the data be ascertained?

There is no control over individuals who can access the information available on the online platform of the repository. Zenodo allows the download of data to all registered users, providing statistics in terms of number of views and downloads. The only identity that can be

known is of persons or organizations who will cite the Zenodo repository following the Creative Commons Attribution Non Commercial International 4.0 licence (i.e., CC-BY-NC).

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

No, because personal and sensitive data will not be offered, as they are not necessary for replicability.

Metadata:

Will metadata be made openly available and licensed 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?

Yes. Zenodo offers the possibility to create projects with CC0 licensing. Projects with CC0 licensing will not enable users to access the data, as they will contain only the metadata. The data will be made available only through Zenodo projects with Creative Commons Attribution Non Commercial International 4.0 licence (i.e., CC-BY-NC licence), ensuring that the data authors get proper attribution and re-use will be made in non-commercial purposes. The Zenodo project containing the metadata under CC0 licensing can redirect to a Zenodo project containing the actual data under a CC-BY-NC licence.

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Data linked to scientific publications will remain available and findable indefinitely, to allow for replicability. In case the data will have to be taken down from the repository, the metadata will remain or will be moved to another repository which allows indexing and harvesting.

Will documentation or references 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)?

The data will be stored using formats that can be read by open access softwares, such as CSVs and text files. The Zenodo projects, where the data will be stored, will also include R files (open-source code and software) describing the data analysis to ensure replicability.

2.3. 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?

These file formats have been selected due to their widespread adoption as accepted standards:

- Microsoft Word (.docx), OpenOffice Document (.odt) & portable document format (.pdf) for text-based documents;
- Comma-separated Values (CSV) & R files for quantitative data;
- Joint Photographic Experts Group (.jpeg) for illustrations and graphic design;
- MPEG Audio Layer 3 (MP3)/ Waveform Audio File Format (WAV) for any video or audio format.

In case of specificities, other formats might be used, if they are explicitly required by publications (e.g., PNG image formats).

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?

The Living Lab study will use existing and open ontologies or vocabularies, such as the International Classification of Diseases 11th Revision (ICD-11) for the classification of declared diseases (if any) or the International Standard Classification of Occupations (ISCO-08) for declared occupations.

Will your data include qualified references[5] to other data (e.g. other data from your project, or datasets from previous research)?

All data collected in the Living Lab study involve primary data, with the published data also falling into this category. Moreover, the shared data will be only data which allows for the replicability of analyses associated with a scientific article. Consequently, the inclusion of qualified references to other datasets will not be necessary, as they are not required for replicability.

2.4. Increase data re-use

How will you provide 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.)?

The research data will be made easily comprehensible and reusable by associating them with a set of rich metadata. Interested parties will have access to a codebook (specifically for the quantitative data types resulting from the study) and code files (containing information on data processing methods). Working files will be distinctly labelled with a version suffix, such as v1, v2, to indicate the most recent database versions, facilitating user understanding.

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?

The data associated with research articles will be openly stored in a designated repository, like Zenodo, where the use of a Creative Commons Attribution Non Commercial International 4.0 (CC BY-NC) license is permitted. In this manner, we ensure the proper citation of data and its uses for non-commercial purposes. Also, the metadata of a Zenodo project can be duplicated into a secondary project that will contain only the metadata of the associated project under a CC0 licence.

[5] *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/)*

Will the data produced in the project be useable by third parties, in particular after the end of the project?

The datasets available for third-party use will be those uploaded to the Zenodo server or any server used by the journal in which the publication decision is made. Each article under review will be accompanied by a dataset specific to the article's content. The datasets will be retained for a minimum of five years. The usage of the data is encouraged in order to be in line with the FAIR principles related to reusability. Given that data will be licensed under the Creative Commons Attribution Non Commercial International 4.0 (CC BY-NC) standard, data will be available to anyone through the Zenodo repository as long as their uses comply with proper citation and non-commercial use.

Will the provenance of the data be thoroughly documented using the appropriate standards?

The data uploaded to the online repository will be curated and checked by the research team. It will also comprise the metadata file, easily readable and comprehended by all third parties interested in re-using the data. All information related to data captions, data processing or data analysis will be provided in order to ease the re-use and interpretation of the data.

Describe all relevant data quality assurance processes.

To facilitate informed decision-making, data must exhibit reliability and high quality. The initial requisite to achieve this is a smooth process of data cleaning. This preliminary step ensures the data's pertinence and its ability to be interpreted by different types of public (to be compliant with FAIR principles). Moreover, the data must be made available in easily accessible formats, enabling seamless access. Furthermore, a rigorous examination of the data's conformity with legal and ethical regulations is imperative.

The dissemination of the results from the Living Lab study will be made available through Open access journals. All researchers will make sure that the data used in their publications is available for checking after the publication, following the journal's internal regulations.

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.

No study output, regardless of its nature (research article, press release, etc.), will contain data that could directly or indirectly lead to the identification of research subjects. The files in which the data will be stored during data collection and processing will be password-protected. In cases where a file format does not allow direct password protection, it will be placed within a password protected folder. Passwords for files and folders will only be known to the project director and the research team members responsible for data collection, processing, and analysis.

In the case of other outputs, e.g., network mapping of stakeholders through exercises with experts, the data will follow all the FAIR principles of data sharing as well as compliance with ethical regulations regarding data anonymization, if necessary.

3. Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

The Living lab study is unique both at the national and international levels. There are several methodological aspects that need to be addressed. First and foremost, the research will be replicated in Bulgaria, specifically in the Plovdiv region. The project will offer a series of training sessions to facilitate the exchange of experiences and knowledge pertaining to field research and technical aspects. Furthermore, ethical, organizational, and planning aspects carried out in the field will be documented. Other research outputs related to the Living lab study will adhere to the same FAIR principles and ethical regulations as the PNA data, in terms of data collection and sharing, if necessary.

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

In consideration of the FAIR principles, an exchange of best practices will be conducted with colleagues in Bulgaria concerning repository types (e.g., Zenodo), the creation of metadata for datasets, and the format in which data will be made available for the purpose of facilitating their reuse. We will also coordinate with other members inside the Prevention and Early Detection Implementation Research Cluster to make ensure that all research outputs that will be shared inside the cluster will meet FAIR principles and accepted ethical regulations.

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.)?

With regard to the cost related to storage, archiving, or security, as the services provided by Zenodo are free of charge, there will be no additional costs. While there may be journals interested in collaborating with specific types of online repositories, the majority of these services are offered free of charge.

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)

In the event of incurring additional costs, which may include exceeding the storage limit for uploading datasets or metadata to certain online repositories (e.g., for the Zenodo repository, the storage limit is 50 GB), these expenses can be covered through the Horizon Europe Grant.

Who will be responsible for data management in your project?

With regard to data management within the Living lab study, the point of contact is Dr. Marius Geantă, and he can be reached at marius.geanta@ino-med.ro.

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 general policy of Zenodo states that data preservation (i.e., longevity) is assured through replicas in a distributed file system backed up nightly to tape, replication into multiple copies in the online system, MD5 checksum of the file content for fixity and authenticity, and integration of all content into suitable alternative institutional and/or subject based repositories as a succession plan in case of repository closure.

Regarding the type of data that will be made available online, the anonymized database will not contain any means of identifying the respondents or individuals mentioned in their social structure. The datasets that will be available online are those used in articles that are currently under review and data shared inside the Prevention and Early Detection Implementation Research Cluster. Each article will have an underlying dataset that will be accessible on the Zenodo repository.

In order to ensure the FAIR principles, data will be available in the repository for at least 5 years.

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)?

Internal storage: The data collected through questionnaires pertain to socio-demographic characteristics, behavioral information (e.g., smoking, physical exercise, etc.), and details regarding individuals within the subject's personal network. These data will be stored locally (offline) in electronic formats (e.g., Excel, CSV, etc.) within password-protected folders on dedicated data collection computers. These computers will have cloud storage synchronization turned off.

Contact information will be stored in password-protected Excel files, while questionnaire data will be saved in files with the specific export format used by the data collection program (e.g., CSV), which will also be password-protected. Access to the data will be restricted to those research team members responsible for contacting respondents, administering questionnaires, data processing, and analysis. It should be noted that the only individuals with access to all the data are those involved in contacting, processing, and analysis. Interviewers who are not engaged in other activities (e.g., data cleaning and processing) will only have access to the data of the subjects assigned to them during the study. The storage of password-protected files will be strictly local (offline). Transfers of raw data between computers will be made with a specially designated USB device.

External storage: Each member of the research team will ensure that enhanced data protection measures are in place. Given that OSF will serve as the primary data repository for the project, access will be secured through a username and password. Team members will ensure the use of a sufficiently strong password to prevent unauthorized access to the data. The datasets that will be available online (in online repositories) are those used in articles that are currently under review. Data stored on OSF will be anonymized, without data that can be considered sensitive, and will contain only those variables necessary for study replication.

The data obtained following the signing of the informed consent form will be recorded in physical format (on paper) and stored at the INOMED headquarters in a dedicated file, with access restricted to the project coordinator.

Will the data be safely stored in trusted repositories for long term preservation and curation?

The datasets will be available on the Zenodo platform, a recognized repository aligning with the guidelines for Horizon Europe projects. One of the attributes underscoring the Zenodo's classification as a trusted repository is its capacity for long-term data preservation, among other features. As their general policy, Zenodo stores items for the lifetime of the repository, which is also the lifetime of the host laboratory at CERN. Also, regarding preservation and curation, in the Zenodo general policy is mentioned that data files and their metadata are backed up nightly and that files are regularly checked against their MD5 checksum to assure that their content remains constant.

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).

The purpose of data sharing is to ensure that open science and open access principles are respected, providing the scientific community the possibility to replicate and build other studies upon the results of 4P-CAN.

Regarding the data collection and analysis process, there are specific laws regarding the protection and sharing of data, and all team members will fulfil the legal requirements in accordance with the EU legal framework. The project team is committed to adhering to the GDPR - Regulation (EU) 2016/679 on the rules concerning the protection of natural persons with regard to the processing of personal data, as well as Law No. 190/2018, which establishes the national implementation of the provisions of Article 6(2), Article 9(4), Articles 37-39, 42, 43, Article 83(7), Article 85, and Articles 87-89 of Regulation (EU) 2016/679. All research team members involved in data collection, processing, and analysis will be required to sign a confidentiality agreement regarding the processing of personal data. Secondly, no study output, regardless of its nature (research article, press release, etc.), will contain data that could directly or indirectly lead to the identification of research subjects. Thirdly, the files in which the data will be stored will be password-protected. In cases where a file format does not allow direct password protection, it will be placed within a password-protected folder. Passwords for files and folders will only be known to the project director and the research team members responsible for data collection, processing, and analysis.

Social and Personal Network Analysis take into account personal data (e.g., name, date of birth, phone number) for data curation procedures (i.e., ensuring data accuracy) and for fieldwork organization (i.e., contacting research subjects for their participation in upcoming PNA data collection waves or 4P-CAN events. These data are known only to the project director and fieldwork researchers. However, beyond the data cleaning and preparation process, and fieldwork organization, the personal data is not necessary for data analysis and reproducibility of results. Thus, all shared data will be anonymized. The anonymization process will take into account not only the removal of information that directly identifies a person (e.g., name) but also that information that can indirectly identify a person.

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

Each participant will be receiving an information sheet that includes details about:

- The purpose and objectives of the study;
- The type of questions included in the questionnaire;
- The option to withdraw at any time without providing explanations;
- Data anonymization procedures;
- Contact details of the project director and other research team members responsible for data processing and analysis;

After the participant has read the information sheet and had any questions clarified, they will be presented with a dual copy document titled 'Participation Consent in Study' for signing. A copy of the information sheet and the study participation consent will be provided to the respondent.

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)?

At present, there are no additional data management procedures beyond those outlined in this document. However, it is possible that other project partners may have specific data management requirements. Nonetheless, all partners will ensure adherence to the FAIR principles.