PREDIMED PLUS STUDY

Centre for Biomedical Research as a Network – Obesity and Nutrition (CIBER-OBN in Spanish)

(with the collaboration of the Centre for Biomedical Research as a Network – Epidemiology and Public Health (CIBER-ESP) and the Centre for Biomedical Research as a Network – Diabetes and Metabolic Illnesses (CIBER-DEM))

Data Management Plan and Data Sharing Policy

Version: 1.6.

16-04-2018

1. INTRODUCTION

1.1. Conceptualisation and current context

A data management plan, or DMP, is a document that describes the way in which data will be managed, both during the active research period and after a project has been completed. This plan contains detailed information on the generation, storage, access, preservation, organisation, sharing, and reuse of research data, as well as aspects related to personal data protection, security, privacy, ethical implications, and intellectual property. DMPs first came into use in with three basic objectives: to ensure that data are not lost, especially at the end of a research project; to facilitate proper safeguarding of data, beginning at the time of production; and to allow for preservation of data, via documentation of their entire technological environment.

The basic elements to be included in a DMP are:

1. The type of data that will be generated during the research.
2. The standards that will be used.
3. The policies on data access and reuse.
4. The measures used to protect privacy, security, confidentiality, and intellectual property.
5. The data file and preservation of data.

At the present time, it is not compulsory to produce a DMP for research projects funded by Spanish public research institutions, and up until now DMPs have not been required for projects funded by the European Commission either. However, the creation of a DMP is considered to be a way to add value in relation to best practices for data management. It has therefore been decided to draft one for the PREDIMED PLUS study. Furthermore, the production of a DMP will be required in the near future for new projects seeking funding within the context of Europe's H2020 Programme. In preparation for new funding requests that various projects derived from PREDIMED PLUS may be submitting under those calls for proposals, it has become necessary to establish a general, framework DMP for the PREDIMED PLUS project as a whole.

Also, if international funding is sought outside the European Union, other agencies may require the submission of a DMP, and it is therefore appropriate to establish some general guidelines that can be easily adapted to the specific requirements of each of those potential funding bodies.

In the United States, research projects were initially asked to submit what is known as a "data sharing policy". Later, those requests began to include not just that policy, but also a more general perspective, with the data sharing policy integrated into the DMP. Then, beginning on 18 January 2011, all project proposals submitted to the National Science Foundation (NSF) had to include a DMP as a supplementary document. In February of 2013, the White House Office of Science and Technology Policy (OSTP) issued a memorandum asking funding institutions to require a plan for the results generated by research financed with public federal funds. As a result, on 26 February 2003 the National Institutes of Health (NIH) adopted a data sharing policy for projects with budgets of more than $500k per year. In response to the 2013 memorandum from the OSTP, the NIH will require *all* researchers to produce DMPs, not just those whose projects receive more than $500k in funding. Since PREDIMED PLUS will be

applying to the NIH for international funding in the near future, in this case it will also be necessary to start with a general basic DMP for the study.

Given this general set of circumstances, the drafting of such a document is now being proposed for the PREDIMED PLUS study, and we have given that document the English title of “Data Management Plan and Data Sharing Policy” (abbreviated as DMP or DMP/DSP below).

The DMP is a living document, and it may be updated based on changes to the general framework for the project’s activities or other external requirements.

As part of the general framework of the DMP for the PREDIMED PLUS study, specific sub-DMPs will have to be drafted for new projects applying for funding from the various agencies in the future, while always adhering to the guidelines from the general DMP.

There are several guides available for drafting a DMP, and the formats vary depending on the funding agencies involved. For example, the Scholarly Publishing and Academic Resources Coalition (SPARC) has compiled an excellent series of resources on the DMP/DSP requirements applied by various federal funding agencies in the United States http://datasharing.sparcopen.org/data).

Part of the reason behind the increase in these DMP requirements is the movement promoting open access to scientific information. That movement has a significant presence in relation to increasing public access to scientific publications appearing in journals. Because of this, many research funding agencies and institutions already have policies in place to guarantee open access to publications derived from scientific research conducted using public funds.

This movement for open access and for the creation of e-infrastructures that will support the use of scientific information by the scientific community has also addressed the importance of research data, which are acknowledged to be a unique source of knowledge independent from the publications they appear in, and which can be used to validate the results of research published as articles, or to generate new knowledge by use in interdisciplinary works. The claim is that in order to ensure that data can be taken advantage of in this way, it must be available and accessible on the Internet, in the same way that the publications are. However, ensuring access to data involves technical and legal requirements that are more complex than those related to publications, and with biomedical research there are ethical restrictions that prevent the production of generalised guidelines for general open access, which instead must be handled on an individual, case-by-case basis.

Various scientific journals (*PLOS*, *Nature*, etc.) have implemented publication policies requiring open access to data, including requiring the provision of the raw data used to generate the results appearing in published articles. Those data can be provided directly, as downloadable files included as supplementary materials, or can be deposited in one of the several available public repositories. If provided as downloadable files, any person can freely access the data and freely make use of them for other purposes. Since biomedical research often makes use of personal data that require special protection, those journals also provide the option of indicating that the data will be provided by the researchers upon request, after specific forms are completed that have been designed specifically for each study according to its particular characteristics and ethical restrictions.

In parallel, the International Committee of Medical Journal Editors (ICMJE) has issued guidelines on providing and granting access to raw data, as used to produce results in relation to clinical trials published in leading medical journals (Taichman et al., 2016). These guidelines will have to be considered in relation to the future publication of results from the PREDIMED PLUS clinical trial. However, since the start date for the PREDIMED PLUS study occurred prior to the publication of those ICMJE guidelines in 2016, **we would not be obligated** to provide open access to the published data, since the editors clearly specify that the requirement for open access will only enter into force for clinical trials that start enrolling participants at least one year after the ICMJE adopts its data-sharing requirements (which the ICMJE is planning to adopt after it considers the feedback received on the proposal published in Taichman et al., 2016).

These recent initiatives are still stirring up intense debate and are subject to improvement in order to protect all of the interests involved (Warren, 2016). One of the most heavily debated points is how to acknowledge the contribution of the researchers who have generated the data, since open access gives great freedom to researchers who re-analyse data, without any obligation to acknowledge the authorship of the researchers who published the study. Because of this, there are now calls to acknowledge this need for balance, so that open access to data can continue to make due progress (Kalager et al., 2016). In addition, as requirements continue to be developed on the open publication of clinical trial results, general criticism has also emerged with regard to the potential for improper use of data and other drawbacks that must be resolved in order to allow the goals of open access to be achieved while avoiding risks and preventing improper data use. There are also proposals to extend the time period allowed for data analysis by a study's authors, between the time when the data are generated and the time when they are made available for analysis by other researchers (International Consortium of Investigators for Fairness in Trial Data Sharing et al., 2016). On this subject, a study was recently published on the percentage of articles appearing in the BMJ that offered access to raw data, and the conclusion reached was that the percentage is very low. The causes that underlie this situation include, notably, the lack of incentives for researchers being asked to share their data (Rowhani-Farid et al., 2016), and this limitation will have to be taken into account in order to improve future policies on this subject.

A consortium has also been created under the name of "Academic Research Organization Consortium for Continuing Evaluation of Scientific Studies – Cardiovascular", or ACCESS CV (Academic Research Organization Consortium for Continuing Evaluation of Scientific Studies – Cardiovascular et al., 2016). The purpose of this group is to provide guidelines for sharing data related to cardiovascular clinical trials in a manner that takes into account the ethical restrictions inherent to studies of that type. That consortium has produced a preliminary proposal, and we will be monitoring its evolution in view of the need for data sharing with future publications from the final PREDIMED PLUS cardiovascular clinical trial.

In relation to privacy, confidentiality, and patient protection, the US Federal Policy for the Protection of Human Subjects ("Common Rule") was recently updated. This policy was initially published in 1991, and it applies consent requirements for studies involving human subjects and the use of biological samples. This update will enter into force in 2018, and it will give rise to changes with respect to the previously existing situation (Menikoff et al., 2017). We will have to take these changes into account when applying for research project funding from the various agencies in the United States.

In summary, we are facing a changing international context in which we still need to fill in many details and address some concerns, needs, and insecurities regarding the various phases of data sharing. More resources must also be provided along with adequate technology, in order adequately adapt to the requirements while preventing any negative impacts on the integrity, security, and confidentiality of the data, or on the efforts and intellectual property of the researchers involved.

1.2. Guidelines for producing a DMP under Horizon 2020

In Spain, the guidelines being used as a reference for the drafting of DMPs are those established as a pilot by the European Commission. The use of DMPs in Horizon 2020 is a new development. The Commission is therefore conducting a flexible pilot study within the context of Horizon 2020, which is known as the Open Research Data (ORD) pilot. The objective of the ORD study is to improve and maximise access to, and reuse of, research data generated by the Horizon 2020 projects. At the same time, this pilot programme especially takes into account the need for a balance between openness and protection for scientific information and between commercialisation and intellectual property rights, while also addressing issues related to data management and preservation. In the work programmes for 2014-16, the ORD pilot only included some selected areas of Horizon 2020. Recently however, in the revised version of the work programme for 2017, the ORD pilot has been expanded to cover all thematic areas of Horizon 2020.

To assist with this management, the Commission has published a Data Management Guidelines document, with the first version released in 2013 and subsequently updated in June of 2016 (European Commission, Directorate General for Research & Innovation. H2020, 2016). This document is directed at funding applicants and beneficiaries for projects included in this pilot ORD programme. Furthermore, the European Commission has already given notice that in 2017 the H2020 data project will no longer be a pilot, and that all projects funded under H2020 (except for certain justified exceptions) will have to guarantee open access to research data. Although PREDIMED PLUS is currently not subject to this open data requirement and is not required to produce a DMP, it is nevertheless opportune to adapt the DMP to these guidelines in anticipation of future funding applications for projects linked to PREDIMED PLUS. The purpose of these guidelines is to offer instructions on how to comply with the requirements related to research data quality, sharing, and security.

In May 2017, the European Commission (European Commission, Directorate General for Research & Innovation. H2020, 2017) published an updated version of the Guide containing the guidelines for the DMS document, entitled “H2020 Programme. Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Version 3.2. 21 March 2017”. Although this document maintains the same model for the general guidelines as contained in the document previously published in 2016, it also adds three additional points: A link to the specific guides for projects funded by the European Research Council (v.3.1, 25/08/16); a clarification of the objectives on page 3 of the document; and an appendix with the specific “agreements” that must be established with the journals to which articles funded by the European Union are submitted, since all of these must provide open access to the publication. This new updated version establishes that it is not compulsory to provide open access to the data generated during the funded project along with the publication. What it does clarify is that readers must at least be given open access to the articles. It differentiates between journals that initially offer open access (gold access) and

those that do not initially have an open access policy, and with whom an agreement must therefore be signed (with the stipulated amount paid), so that once a certain embargo period has passed, readers will have open access to the article, or as an alternative, the article can be deposited in an open-access repository (green access).

In both cases (gold access and green access), it is recommended that a copy of the article should be made available in another open-access repository in addition to the journal's location. Although there are various repositories existing, depositing a copy of the article in the Openaire repository is recommended (https://www.openaire.eu/).

Although open access to research data applies by default to Horizon 2020, the Commission also acknowledges that there are good reasons to maintain closed access for some of the research data generated during a project, or even for all of them. The Commission therefore offers possibilities for excluding any project from the open data obligation, during any phase (in other words, during the application phase, the funding agreement preparation phase, and after signing the funding agreement).

It is also important to remember that the ORD pilot is being applied mainly to the data required in order to validate the results presented in scientific publications. Other project data can also be provided voluntarily by the researchers, and this can be specified in the respective DMPs.

The guidelines proposed by the Commission are based on the acronym FAIR, which stands for "Findable, Accessible, Interoperable and Reusable". (The terms used in Spanish are: *descubribles*, *accesibles*, *interoperables*, and *reutilizables*.) The document produced by the Commission helps to improve searchability, accessibility, interoperability, and reusability of data from the research it funds, in order to ensure proper management. This is based on the idea that good management of research data is not a goal in itself, but rather a key type of conduct that leads to the discovery of knowledge and innovation, and then to the integration and reuse of data and knowledge.

In relation to the guidelines used, it is also important to consider the OpenAIRE projects (https://www.openaire.eu/). OpenAIREplus is a continuation of the OpenAIRE project, and it is focused on the publication of research datasets and their incorporation into the scientific articles funded by the Horizon 2020 programme.

The data for all projects that have European funding and that are subject to the DMP obligation must be deposited in a repository in order to facilitate their open reuse, with the following exceptions: incompatibility with the obligation to protect the results if they could be commercially or industrially exploited; incompatibility in terms of confidentiality or security; incompatibility with regulations on personal data; if such access could put the project's objective at risk; no data will be generated or collected; or some other legitimate reason, which must be justified during the project proposal phase.

In order to comply with the open-data obligation, participants in a project must complete two steps: deposit the data in a repository, and allow open reuse by means of licencing.

The European Commission will be able to cover the technical and professional costs associated with handling and distribution under open access.

In Spain, the Spanish Science and Technology Foundation (FECYT in Spanish), in collaboration with the University Library Network (REBIUN) of the Conference of Spanish

University Rectors (CRUE), manages and coordinates RECOLECTA, a project for the creation of a network of interoperable institutional repositories. This can be considered as Spain's first nationwide initiative for the creation of an infrastructure to facilitate open science. Part of this objective is also to improve the visibility of research results and provide services for scientific output in Spain. The working group for the RECOLETA project has drafted a report on the current situation in Spain in relation to the main aspects of open access to research data (Working Group from the RECOLETA project, "Depositing and Management of Data under Open Access" (Depósito y Gestión de datos en Acceso Abierto, 2012)).

1.3. Definition of data

The term data can be defined in a variety of ways, but it seems to be generally accepted that research data are the facts, observations, or experiences on which theories or experimentation are based. Data can be numerical, descriptive, or visual. Data can also be either raw or analysed, and they can be experimental or observational. Sources of data include: laboratory notebooks, field notebooks, primary research data (either printed or stored on digital media), questionnaires, audio recordings, videos, models developed, photographs, films, and evidence derived from testing and verification. The data generated by a research project may also include PowerPoint presentations, designs, and samples. Information about the sources of data can also include aspects related to collection: how, when, where, and with what (for example, the instruments used). The software code used to generate, comment on, or analyse data can also be considered as data itself. In summary, there are several types of data generated during a research project such as the PREDIMED PLUS study, and many of these types are not subject to any ethical restrictions and can be easily shared in the various existing public repositories.

1.4. Tools to assist with the creation of a DMP

There are variety of tools available online to help create a DMP, based on the requirements of the various funding agencies. At the European level, and specifically in Spain, the most commonly used tools are the following:

- **DMP Online**: developed by the **Digital Curation Centre** (https://dmponline.dcc.ac.uk). This can be used as a template to produce a Data Management Plan following the European Commission's model. The tool is accessed by creating an account and then answering the questions on the form. It is a good idea to have the responses to the questions prepared in advance, so the "Checklist for a DMP" found there is very useful.

- **PAGODA**. This is a translation of the DMP Online tool into Spanish, produced by Consorcio Madroño. This consortium is made up of the Universities of the Community of Madrid and the UNED Foundation for Library Cooperation. This tool can be accessed just by creating a free account.

-**DMP CSUC**: This is a translation of the DMP Online tool into Catalan, performed by the Consorci de Serveis de les Universitats de Catalunya (Service Consortium for the Universities of Catalonia) and the Universitat Jaume I de Castellón (http://www.csuc.cat/es). This tool can be used just by creating an account. The CUSC has also created manuals to assist with filling in the DMP Online.

2. Development of DMP headings for the PREDIMED PLUS study

2.1. Summary information

2.1.A Description of the study

The name of the study is abbreviated as PREDIMED PLUS, and its full title is: "*Effect of an intensive lifestyle intervention based on a reduced-calorie traditional Mediterranean diet, physical activity, and behaviour conducive to the prevention of cardiovascular disease*". This is a multi-centre, randomised and controlled intervention trial. The objective is to evaluate the effect on primary cardiovascular prevention of an intensive intervention based on promotion of a reduced-calorie Mediterranean diet and specific weight loss goals, physical activity, and behavioural intervention versus a low-intensity intervention (normal care) with the recommendation of a Mediterranean diet without calorie restrictions (control group). The main objectives established are the differences observed between the two interventions in relation to:

(i) the incidence of cardiovascular events (non-fatal myocardial infarction, non-fatal stroke, or cardiovascular fatality);

(ii) weight loss and its long-term maintenance.

There are multiple secondary objectives, which are detailed in the Study Protocol.

There are currently 6,874 participants who have been recruited and randomised. This is not a single project funded under a single call for proposals, but rather a multi-centre study being initiated at different times at each of the participating centres, and currently relying on funding from a variety of entities. It is expected that funding will continue to be sought under later calls for proposals. The PREDIMED PLUS study, which is being coordinated by Jordi Salas-Salvadó, is being conducted by 23 recruiting nodes and 8 scientific-technical support nodes. The majority of the funding for these nodes comes from the government of Spain by way of the Carlos III Health Institute of the Ministry of the Economy and Competitiveness, through the Centre for Biomedical Research as a Network – Obesity and Nutrition (CIBER-OBN in Spanish), and with the collaboration of the Centre for Biomedical Research as a Network – Epidemiology and Public Health (CIBER-ESP) and the Centre for Biomedical Research as a Network – Diabetes and Metabolic Illnesses (CIBER-DEM). FIS grants (Strategic Health Action) were also received for coordinated projects in 2014 (coordinator Jordi Salas-Salvadó), 2015 (coordinator Pep Vidal), 2017 (coordinator Jordi Salas-Salvadó), and 2018 (coordinator Pep Vidal), and additional PIs received FIS grants under other calls for proposals. In addition, an ERC Advanced Research Grant was awarded to Miguel A. Martínez-González.

Nodes: The Principal Investigators (PIs) for the 23 recruiting nodes are, in order of the numbering for the nodes: 01 Julia Warnberg, 02 José Lapetra, 03 Alfredo Martínez, 04 Dora Romaguera-Miquel Fiol, 05 Ramón Estruch, 06 Montse Fitó, 07 Jordi Salas-Salvadó, 08 Aurora Bueno, 09 Clotilde Vázquez, 10 Miguel A. Martínez-González, 11 Fernando Arós, 12 Dolores Corella, 13 Lluis Serra-Majem, 14 Xavier Pintó, 15 José López-Miranda, 16 José M. Ordovás, 17 Pilar Matía-Miguel A. Rubio, 18 Francisco Tinahones, 19 José A. Tur, 20 Josep Vidal, 21 Jesús Vioque, 22 Miguel Delgado, and 23 Vicente Martín. The PIs for the support nodes are: Emilio Ros, Fernando Fernández-Aranda, Cristina Botella, María Puy-Portillo, Rosa M. Lamuela-Raventós, Ascensión Marcos, Guillermo Sáez, and Enrique Gómez-Gracia.

A Steering Committee that is currently made up of 7 PIs (Dolores Corella, Montse Fitó, Ramón Estruch, Miguel Ángel Martínez-González (replaced by Miguel Ruiz Canela while on leave), Emilio Ros, Francisco Tinahones, and Jordi Salas-Salvadó) manages the study.

2.1.B Relationship between data generation and project objectives

The project began in 2013 at one of the centres, and the other recruiting centres were gradually added afterwards. Finalisation of the project is planned for 2022. Longitudinal tracking is required in order to monitor cardiovascular events, since these incidents are the project's main subject of quantification. However, right now it cannot be guaranteed that the funding available will last until the end of the project. Nevertheless, this preliminary DMP will be developed under the hypothesis that enough funding will be available, so that the final results can be planned as a database containing the final data for the monitored events after the scheduled longitudinal tracking has been completed. In parallel, there are also plans to generate partial databases containing baseline results and the results of 6-month and annual monitoring, which will allow the primary objective of weight loss to be evaluated as well. The secondary objectives can also be assessed at various levels by using the various monitoring intervals and sample sizes.

It is important to point out that the PREDIMED PLUS study consists mainly of what is known as the **General Project**, on which the initial efforts and funding received have been focused. This is the project developed in the Study Protocol and being carried out jointly by the 23 recruiting groups, with application of the same questionnaires and other methods for collecting data and biological samples. However, a series of established specific projects are being conducted in parallel which include the participation of various groups and nodes, either individually or in coordination. The list of **Specific Projects** and their coordinators can be found at the website for the PREDIMED PLUS study (http://predimedplus.com/).

2.1.C. Specify the types and formats of data generated

Both the baseline data and the data from monitoring will be generated in a variety of types and formats. As part of the general project, observational data are generated in real time, such as measurements of weight, height, and blood pressure. Data are also generated based on questionnaires and from the analyses of images from various instruments, accelerometers, electrocardiographs, etc. After blood samples have been drawn, laboratory data are generated for certain biochemical determinations, either as part of the basic analytics or using other more specific analyses. Data on primary and secondary events will also be obtained from clinical records and records on fatalities, which will be validated by the Events Committee with the periodicity established.

In the specific projects, experimental data may be generated from more complex measurements using biomarkers, microbiota, omics data, ultrasound, DEXA, etc.

These data may also be contained on paper questionnaires or spreadsheets, etc.

The formats generated essentially include text, numerical, and images.

In addition to these data, original research articles will be produced, and an effort will be made to publish them with open access.

Review articles will also be written on various subjects related to PREDIMED PLUS, again with an effort made to publish these with open access whenever possible.

2.1.D. Specify whether existing data are being reused

Some data already collected, taken from the patients' clinical records, may be used to annotate the analytics and as quality control for data on related pathologies. Later, after monitoring, data from the clinical records will be used to document primary and secondary events.

2.1.E. Specify the data source and the data flow

2.1.E.1. Data source

The data will be fundamentally generated within the project. Based on the general protocol for the project, information will be obtained from each participant by direct measurement of weight, height, waistline, blood pressure, heart rate, etc. Analytical data related to the variables specified in the general project will also be obtained from blood and urine samples, and questionnaires of various types will be administered in order to obtain information about foods consumed, adherence to a Mediterranean diet, physical activity, quality of life, cognitive data, etc. Biological samples are also being collected (serum, plasma, urine, buffy coat, and nails), and a protocol on collection, storage, and distribution of samples has been established specifically for the study. Additional types of data will be obtained in the specific projects, such as data from questionnaires, images, and data from laboratory determinations, and other samples will also be collected (stool samples, blood cell types, etc.). The specific projects tend to be carried out using subsamples of participants from the PREDIMED PLUS study. In order to perform their analyses, these projects depend on the availability of additional external funding, or else funding from the node itself. For this reason, it can be anticipated that the data from many of the specific projects will be generated a little later than the data from the general project.

2.1.E.2. Data flow

The recruiting centres generate data on the trial participants they have recruited, and they also collect biological samples from them. The support groups may generate additional data from the samples analysed. The recruiting nodes primarily collect data from their corresponding participants using what are known as data collection notebooks, which may be on paper or in digital format. After being collected by the nodes, these data are sent on a daily or weekly basis to the data coordination centre, which is the Municipal Medical Research Institute of Barcelona (Instituto Municipal de Investigación Médica, abbreviated as the IMIM in Spanish and hereinafter), and to other centres that are receiving data. These data are derived from printed questionnaires with a format that allows electronic reading (on food consumption frequencies, quality of life, neurocognitive aspects, etc.), or the data may be accelerometry data or images. The PREDIMED PLUS website includes protocols that provide details on how these data must be collected and sent. For the most part, the general data are submitted by filling in electronic forms that have been designed specifically for the study (using the TELEFORM format), and then transmitted in real time to the centralised database at the IMIM. These forms can be accessed and sent via the website, but only after the required password has been entered into order to access the private area for PREDIMED PLUS researchers. The online transmission of data online has also been established using the private-access area at the PREDIMED PLUS website, in conformity with the corresponding manual.

Some questionnaires are not processed electronically at the recruiting node but are instead submitted by post, then later scanned at the corresponding data processing centre (University of Navarre and Bellvitge Hospital). Once the questionnaires on paper have been processed at the centres that receive them for electronic reading, the data generated for the variables are

sent to the central database at the IMIM to be included with the rest of the general data for the project. For this purpose, patient identifiers (ID) are always included on every data collection form. However, only the recruiting node has the ability to associate this patient ID with the patient's name. Considering the importance of data privacy, the recruiting node is responsible for maintaining the confidentiality of these assignments and for implementing the mechanisms required in order to ensure that the relationship between the patient's full name, national ID document (DNI), and patient ID cannot be publicly revealed. It is recommended that these files used for matching up identification information should be housed on computers or data storage devices that are not connected to the Internet.

Data are sent by the recruiting nodes at baseline, at six months, and annually, according to the contents of the Study Protocol. The data are entered into the central database at the IMIM and also sent to the various additional data processing centres (University of Navarre, Son Espases-Illes Balears Hospital, Bellvitge Hospital, Universidad Rovira i Virgili, Universidad de Málaga), either electronically (electros and accelerometry to the Balearic Islands and Universidad de Málaga, respectively) or on paper. The IMIM receives these data for the General Project from the recruiting nodes and from the other data processing centres, and saves them in a single database for the General Project. This database is subject to the respective quality controls, including periodic consultation with the nodes regarding any data that are anomalous, missing, illegible, etc. After the nodes respond to these queries the rest of the quality control process for the data is completed, which takes place in accordance with applicable international standards.

Periodically, the IMIM sends a file to each recruiting node with all of the data for the variables, as generated for that node's patients. This is done so that the node can safeguard the file and also perform an additional quality control, with the IMIM being notified of the results quickly so that the appropriate corrections can be made for the variables involved.

Databases will be generated with the baseline data, and also at 6 months and annually according to the general data collection protocol, and these databases will also be subject to longitudinal application of the pertinent quality controls.

The characteristics of the current data management system being used for the PREDIMED PLUS study do not allow researchers to gain direct access to the IMIM database. The ideal situation would be to have a **virtual data enclave** system available (Bao et al., 2016), where the data could be securely deposited and managed. This type of system gives researchers remote access, so they can view data the variables and perform data analyses using the most suitable software, which is also installed within the enclave, and with a record maintained for the syntaxes and statistics generated. This type of system guarantees that the data for the variables will not be altered, that the quality control is very high, and that the security, confidentiality, and privacy of the data can be better protected. At its in-person meeting held in Barcelona on 19 January 2017, the DSMB recommended a system with these characteristics as a good strategy for improving the data management taking place for the PREDIMED PLUS study.

Since we do not have this type of data enclave system available at this time, the storage, updating, and quality control for the data is being carried out at the IMIM in a controlled manner by those responsible for the system, in order to guarantee the integrity of the data and following the protocols used to ensure its security (with periodic backups), privacy, and confidentiality.

That database is used to generate data files with various characteristics, which are the files sent to the nodes and made available to the researchers.

Since PREDIMED PLUS is a longitudinal study, data are continually being stored, updated, expanded, and reviewed. To allow for this, a periodicity must be established for the generation of the data files, so that these can be made available to the researchers with the guarantee that all of the cases available have been incorporated for each variable, and that the corresponding quality control has been carried out on the data based on the queries and feedback from the various nodes. For this purpose, realistic dates must be established as deadlines for the periodicities indicated, with the nodes required to submit all data they have available for the general project by those dates.

With this periodicity, or in a series as indicated, a **basic data file** must be generated (corresponding to the baseline data plus the S1, S2, S3 visits), containing the **data from all questionnaires and measurements stipulated in the General Project** and including the **randomised participants from all the nodes**.

Although an effort is being made to ensure that this basic data file has all of the up-to-date data for the participants, for some variables or nodes this will not be fully possible by the stipulated dates, which means that the reception of data will be closed in order to avoid further delays in creating the data file. The corresponding data files will be assigned a version number and a date of generation. This version number and date of generation must be cited in the data analyses and publications derived from those data files, also indicating the number of participants from which each data element was obtained. This will allow missing data to be added to the database after the deadline for the generation of the database at any of the specified times, instead of allowing those data to be discarded. In the past, any data not submitted on time were excluded, in order to maintain consistency in the variables across the various analyses and to prevent discrepancies in the results. However, currently the preference is to incorporate data that become available over time, up until the final closure of the database at the end of the study. Always mentioning the version and date of the data file analysed is a good practice, and one that helps ensure the overall quality of the analyses.

After facilitating access to the baseline data file, the data and the quality control of those data will continue to be updated, generating data files on a yearly basis (also including the 6-month variables), on a 3-year basis (including the 2-year data) and on an odd-numbered basis, until the final database is completed with all the years. This periodicity is being established in order to ensure that the provisional data files are as complete and up-to-date as possible, and that the quality is controlled. If an analysis requires the use of variables from even-numbered years prior to the generation of the data file for an odd-numbered year, access to that data file would be provided, with the quality control being focused on the specific subset of variables required.

The data files for the **general project**, with the periodicity indicated, will be made accessible to the principal investigators from each node by means of a personalised password, along with the completion of a form that states the researcher’s commitment to maintaining the security, integrity, confidentiality, and privacy of the file. The principal investigators will also be responsible for ensuring the application of best practices in the use of data by the researchers from their node, and for ensuring that the data are not distributed to unauthorised personnel from other research groups outside of their node.

Generation of data by the **specific projects** will depend on when the funding for the projects becomes available. Since an **additional effort** is made to generate new results during the specific projects, this must be acknowledged. Because of this, and without establishing a formal embargo period, it is considered appropriate to establish some general considerations for assessment of the additional work performed, and to give priority to the node(s) generating specific data in terms of the publications derived from those data, especially when the results involved are very costly in terms of time and resources. As such, in the case of very costly analytical results (metabolomic, genomic, proteomic, metagenomic, epigenomic, expensive biochemical analyses, etc.) that have used samples from multiple nodes, and where the expenses have been covered using funds primarily obtained by the node producing those results, and even though that node must send its results to each node that provided samples when the article is written and sent out for review, no other proposals involving the analysis of those data will be accepted from other nodes until the initial paper written by the first node is accepted.

If another PI has an interest in those data before the paper is accepted, that PI must contact the person responsible for the study under consideration in order to establish a collaboration, including the production of a joint proposal for an article or articles according to the publication policy. This contact can be established beginning at the time when the study of the data is initiated. In all cases, both before and after publication of the initial paper, the node that initially studied the data will have priority in terms of proposing new papers related to their determinations, and the proposal of publications with co-lead authors is recommended in cases where another node has a substantial interest in the same subject.

These considerations, which are applied to the data generated during specific projects, will also be applicable in cases in which external funding is obtained to study samples from multiple nodes. If the terms and conditions of external funding require the data to be deposited in a particular repository at the end of the project, it will be a priority to ensure that PREDIMED PLUS researchers will have prioritised access to those data, before any external researchers that may request such access.

In general, all data derived from information or biological samples taken from the majority of the centres participating in the PREDIMED PLUS study (more than 50% of the centres), or data that have been determined in cases of primary events, will have to be integrated into the general database maintained at the IMIM, which will be available to all PIs from the recruiting nodes. Periodically, the Steering Committee will define the variables that are susceptible to being integrated into the data files for periodic access by the PIs. In order to publish manuscripts using these data, which are not part of the general project, a PI must contact the person(s) responsible for the generation of that data in order to draft a proposal for an article and submit it to the Steering Committee according to the publication policy. Independent of this periodicity, PIs may contact the coordinators of specific projects to request collaboration on a shared work. Further details on this general and specific access are found in section 2.2.2 (Making the data accessible).

2.1.E.3. Ethical considerations regarding the sources of data

In addition to collecting health-related data from the participants using questionnaires and various other instruments, data on health incidents will also be collected for the participants (with the periodicity agreed on by the DSMB/Steering Committee). These data on events will

be documented based on the participants' clinical records. Since these are data related to personal health and subject to the legislation on personal data protection, the informed consent of the participant must be obtained, not only to allow the questionnaires and conclusions to be used, but also to give express consent for the participant's clinical records to be included. This point is also specified in the informed consent form available for the general study, which must be completed by all participants from each node. Completion of an additional informed consent form must be requested for the genetic study. The patient must sign the informed consent form(s) in order to participate in the study, but can withdraw his or her consent at any time.

Each centre is responsible for obtaining authorisation from the Ethics Committee for the participants it recruits, and for safeguarding the signed informed consent forms under appropriate conditions of security and confidentiality.

Since various health questionnaires are also being administered during the study, including the collection of analytical data, data on related illnesses and risk factors, and data on lifestyle variables, it is not possible to make the data generated accessible to the general public. The patient is informed that his or her data will be treated in a confidential manner and that it will be deposited in a database registered with the Spanish Data Protection Agency (Agencia Española de Protección de Datos), with an obligation to comply with the legislation on that subject. Each PI from a participant recruitment centre is responsible for overseeing protection of the data contained in any database or data file registered under his or her name (for more details see sections **2.4** and **2.5**).

According to the Spanish Data Protection Act, the participant must be informed regarding the use of his or her data, and may then either grant or deny consent for such use. The general informed consent form for the PREDIMED PLUS study informs the participant that a portion of the data generated using his or her information may be shared with other researchers, while also stating that a collaboration must be established in order to do this. Any such data sharing can only take place under conditions where access to the data is controlled, after a researcher has been identified and after a collaboration agreement has been established.

For purposes of quality control, there are also plans to perform periodic audits of the participant recruitment centres, in order to verify the application of best practices in relation to ethical matters.

If any participant from the PREDIMED study withdraws his or her consent and expressly requests the destruction of his or her data or samples, the Principal Investigator from the node involved must notify the IMIM about this in writing, as well as the biological sample banks and specific centres that are processing samples (accelerometry, DEXA, etc.), so that the data can be deleted from the specific databases and sample banks.

2.1.F. Specify the size of the data generated

The intention is to generate multiple data files of various sizes periodically and longitudinally throughout the study. In the end there will be a final database available with all of the baseline data and monitoring data from the general project, into which the data generated by the specific projects, which are in the order of multiple GB, can be incorporated.

In addition to data derived from questionnaires there are data generated on the questionnaires, manuals, images, and videos, which in turn represent hundreds of GB in size. The data

generated in the specific projects may be larger in size than the database from the general project (some bulk omics data), and this characteristic should be taken into account in order to opt for the availability of a single, federated database structure housed in various locations for purposes of improved data management.

The total volume is estimated as multiple TB.

2.1.F. Describe the usefulness of the data generated

The data generated during the production of the questionnaires, study protocol, procedure manuals, dietary intervention materials, physical activity intervention materials, behavioural intervention materials, recording of events, presentations at conferences, etc. may be of interest to other groups involved with research or teaching, or even to the general public.

The data generated in the database may be of great interest for other researchers in terms of collaboration on comparative studies, replication of results, meta-analyses, and other types of additional studies.

The data generated by means of statistical analyses will be useful in terms of the generation of new knowledge on each of the study's objectives. The data will be published in the form of original articles, and an effort will be made to ensure that there is open access to the publications.

2.2. FAIR Data (Findable, Accessible, Interoperable, and Reusable)

2.2.1 Making the data Findable

2.2.1.A. Making the data findable: the provision of metadata

The metadata that are standard in the discipline will be used to describe the databases generated. Specifically, the standards used for metadata will be those described in the Dublin Core Schema (http://www.dcc.ac.uk/resources/metadata-standards/protocol-data-element-definitions), which is frequently used as a reference source. The metadata described there include, among other types: Specification: http://prsinfo.clinicaltrials.gov/definitions.html. Standard's website: http://clinicaltrials.gov/ct2/manage-recs/resources.

We will create metadata to describe each database we generate, including at least the following metadata:

- Title: Name of project, based on the data set or research produced
- Names of the creators and addresses of the organisation or individuals that created the data
- Identification code for the data, including internal reference codes
- Words or phrases that describe the subject or contents of the data
- Sponsors: organisations or agencies funding the research
- Rights: any type of intellectual property rights associated with the data
- Access to the information: where and how the data can be accessed by other researchers
- Language of the content
- Key dates associated with the data, including the start date and end date for the project
- Launching, period of time covered by the data, and other dates related to the life cycle of the data (for example, maintenance cycle, updating of the programme)

- Location to which the data refer (e.g. a physical location, spatial coverage, etc.)
- Methodology: how the data were generated
- Data processing: all information about how the data have been processed
- List of filenames for the list of all data files associated with the project, with their filenames and file extensions
- File formats for the data, as required in order to read the data
- File organisation: structure of the data file(s) and arrangement of the variables
- List of variables in the data files
- Explanation of any codes or abbreviations used
- Versions with date / date and time for each file, and using a different ID for each version
- Verification operations to confirm that the files have not changed over time (algorithms such as checksum or hash, MD5, SHA-1, etc.), to protect the integrity of the data

2.2.1.B. Describe the identifiability of the data, with reference to standard identification mechanisms. Will Digital Object Identifiers (DOI) be used?

The data corresponding to questionnaires, user manuals, recipes, protocols for dietary intervention or physical activity intervention, etc. will be deposited in public repositories, such as those managed by our universities. These include the RODERIC repository (http://roderic.uv.es/) at the University of Valencia as well as other similar ones where this depositing could take place. Specifically, RODERIC reflects the University of Valencia's commitment to adhere to the Berlin Declaration (30 September 2008). It uses standardised international protocols to ensure that documents appear in the results displayed by Internet search engines such as Google, etc.

These institutional repositories provide a unique URL for accessing the corresponding documents, in the format https://repository/record/1234.

The data deposited in certain repositories such as Zenodo, Dryad, Pangaea, or FigShare receive a DOI, and in such cases we would use the DOI as the unique identifier for the data deposited there.

The ideal situation is to have a DOI for the documents, metadata, and aggregate data deposited, and use of DOI Citation Formatter gives us a simple interface for first extracting metadata automatically based on a DOI, then building a complete citation. It is compatible with more than 500 different citation styles in 45 languages.

2.2.1.C. Explain how the folders and files are named and structured

The folders and files for the general project will have an initial identifier with the name the study, then the specific name of the file, the creation date, the version, and the specification (XX) (e.g., "PREDPLUS_filename_date_version_XX"). The last two digits (XX) will be used if the file created corresponds to the entire general project at all of the PREDIMED PLUS recruiting nodes. When specific files containing data from just one node are involved, the last two digits will correspond to that node, based on the standardised node numbering established for the PREDIMED PLUS study (node 01, node 02, node 03, etc.). Also, specific thematic folders will contain files corresponding to a particular subject. The folders and files for the specific projects will have an initial identifier that indicates that the study named is a specific study (ES for

específico in Spanish), followed by the filename, date, and version: e.g., "ESPREDPLUS_filename_date_version".

2.2.1.D. Indicate how different versions of the same dataset will be identified

Control of the various versions of a dataset will take place using a suffix that consists of the date (in yyyymmdd format), followed by the numbers of the version and sub-version (for minor modifications), e.g., "V_1_S_0", "V_1_S_1", etc.

2.2.1.E. Describe how the metadata are captured/created

Metadata will be captured/created in a mixed manner, based on the type of metadata being generated and on whether or not standards exist.

When standards exist they will be followed, using either the Dublin Core or the ISO/IEC 11179 metadata registry standard, accordingly.

In other cases, a manual process will be used, so that each specific file will be accompanied by its own metadata in order to facilitate and clarify its use.

2.2.2 Making the data accessible

2.2.2.A. Explain how the data generated will be made available internally to researchers from the multi-centre study, to external researchers, and to the general public

The PREDIMED PLUS study is a multi-centre study, in which each recruiting node generates all of the general project data corresponding to the patients recruited at that centre. These data are then sent to the data processing centres, either electronically or on paper, then finally integrated into a general database housed at the IMIM. The research involved is a long-term longitudinal study, which means that the final database for the project would be available over the long term (more than 10 years after initiation of the project), once collection and validation of primary and secondary events has been completed as well as quality control for the data. Therefore, even though in most cases data access policies for research studies refer to the final database for a project, with the PREDIMED PLUS study we are also considering access to the series of partial databases that will be generated periodically over time. The detailed periodicity for generation of the data files has been specified in section 2.1.E.2 (Data flow).

In parallel, there are specific projects that will be progressively generating data as funding becomes available, and the results from these projects will be gradually integrated into the central database (genomic analyses; DEXA measurements; identification of microbiota; identification of various types of biomarkers in biological samples (plasma, serum, urine); bone densitometry; ultrasound; other imaging tests; etc.). These data are generated by specific nodes, sometimes in a coordinated manner, in order to contribute to the specific project.

The data can be shared internally or externally, and in both cases best practices must be reliably applied in relation to use of the data. It must also be remembered that there are personal data involved that are subject to privacy and confidentiality issues, and that those data have been provided for the purposes expressed on the informed consent form. The procedures to be used for internal data sharing are described in detail below, as are those for sharing such data with external researchers and with the general public.

2.2.2.A1. Procedure for internal data access/sharing

The **Principal Investigators from the recruiting nodes** for PREDIMED PLUS will have access to the complete data files for all randomised patients, as generated during the general PREDIMED PLUS study. This includes access to both the final file for the study after it has been completed, and the various files generated periodically as the questionnaires are given and the quality control is performed on the data: basic data file, one-year data, three-year data, etc. (see section 2.1.E.2: Data flow, where this periodicity is explained in more detail).

The PIs from the recruiting nodes will also be able to access general project data corresponding to a periodicity other than the standardised ones. Such access must be expressly requested from the Steering Committee so that the viability of the request can be evaluated, and the quality control can be performed for the corresponding variables before the file is provided. In the same way, PIs from the recruiting nodes will also have access to data for the variables generated during the specific projects (genetic, biomarkers, other types of biochemical determinations, DEXA, microbiota, ultrasound, other imaging tests, etc.), at the time when these data become available as indicated in point **2.1.E.2.**). In all cases, before any requests to produce papers using those data can be considered, the first papers submitted for publication by the nodes that generated the specific, costly determinations must be accepted, or else a collaboration request must be put in. The nodes that are responsible for specific, costly determinations will have priority for proposing new papers derived from generation of those specific data. Recognition of this time period is important, because there is evidence that one of the main problems leading to the rejection of data sharing among investigators is a lack of recognition for those who generated the data (Longo et al., 2016; Kalager et al., 2016). It is therefore important to take the need for such acknowledgment into consideration during the PREDIMED PLUS study. Once the data become available for other researchers, active collaboration with the groups that generated the pertinent data is recommended, with the ideal situation being one where the lead authorship positions for the articles are shared (Longo et al., 2016). In the same way, if a research group has a special interest in conducting research using the data generated during a particular specific project, a collaboration agreement can be established between the groups during the earliest stages of the analytical determinations and, in general, submission of a joint paper proposal to the Steering Committee is recommended.

If a researcher performs any new determination of a measured variable for PREDIMED PLUS participants that are from some other node, all of the other researchers from the recruiting nodes must be notified. Specifically, the PI from the node performing that determination must provide notification within a period of no more than two months.

Each data file is accessed by means of a personalised username and password, created to gain access to a platform on the IMIM website established for this purpose. The PIs from the recruiting nodes will have access by downloading the data file after completing a request form, indicating the file required (basic file from the general project, one-year, three-year, or other files for other periodicities or with specific data), and they must make a commitment to applying best practices when handling and managing those data. The PIs from each node will be responsible for ensuring the security, integrity, confidentiality, and privacy of the data received, since in addition to data related to the participants from their own node, data on participants

from all of the other nodes are received as well. Appendix 1 contains the data request and commitment form for PIs at the recruiting nodes.

Similarly, the PIs from other non-recruiting PREDIMED PLUS nodes will be able to access the data they need for their research by submitting a specific request for data to the Steering Committee, which will evaluate the viability of the request. The form attached as Appendix 2 must be filled in for this purpose. Essentially, these documents confirm the commitment to respect the following points:

- To maintain the confidentiality and privacy of the data

- To ensure data quality and follow best practices for analysis and processing

- To refrain from distributing data to third parties not authorised to use the data

- To acknowledge the efforts of the researchers that generated the data, by including them in the authorship of publications derived from access to those data

- To submit a proposal for paper publication using the standardised form, which must receive the pertinent authorisation from the Steering Committee

- To adhere to the PREDIMED PLUS publication policy

2.2.2.A2. Procedure for data access or data sharing requested by external researchers
Raw data from the PREDIMED PLUS study will not be deposited in repositories with fully open access. PREDIMED PLUS was initiated prior to the issuance of the requirements on universal access to data by the ICMJE (2016) and by the European Commission, and therefore the PIs are not obligated to provide the entire scientific community with open access to the data used in publications. In addition, there is an ethical restriction on providing open access to the data, since the informed consent form signed by the patients does not authorise such access.

If an outside group wishes to conduct research using PREDIMED PLUS data, an agreement can be established. However, this will first require a detailed joint collaboration proposal (with at least one PI from PREDIMED PLUS included). The Steering Committee for the PREDIMED PLUS trial must make a decision on whether or not the collaboration agreement can be signed based on a detailed study and review of the ethical implications. This decision may also include the possibility of charging a fee.

For collaboration with **external researchers who are only requesting data**, there will essentially be two modalities possible.

2.2.2.A:2.1: Collaboration in which the external researchers are only requesting aggregate data for their variables of interest, without requesting access to the raw data for variables

This procedure is often used by research consortiums to conduct meta-analyses following uniform protocols and a detailed, specific analysis plan. With this modality, the only collaboration being requested is the provision of aggregate data. The researchers from the PREDIMED PLUS study will perform the statistical analyses using the raw data, and they will then provide the external researchers with the descriptive statistics and measures of association required in each case. This procedure will prevent any additional problems related

to the potential for the improper use of the raw data for the variables by external researchers, since the statistical analyses and quality control will be performed internally.

In order to formalise this type of collaboration, a request letter from the external researcher(s) must be sent to the Steering Committee of the PREDIMED PLUS study, along with a completed data request form (**see Appendix 3**). This form will include information about the researcher(s) making the request: the reason for requesting data from the PREDIMED PLUS study; details about other groups that will be included in the meta-analysis; objectives of the proposed study; sample size required; initial hypothesis; detailed methodology; initial variables; and statistical analyses proposed, with detailed specifications on the models and coefficients and the aggregate data to be provided after the statistical analyses have been performed. Information will also be requested about the intended number of articles to be published based on the collaboration, the possible journals to which the articles will be submitted, and the number of PREDIMED PLUS researchers expected to be included among the authors of the articles potentially published. A Guide to Data Access Requests is attached as **Appendix 4**.

The Steering Committee will review the request letter by applying the criteria for excellence in research, and will then decide whether or not to approve the collaboration based on the information provided on the request form. The Committee will also take into account whether the proposal overlaps or interferes with the objectives and proposals of internal researchers from the PREDIMED PLUS study. If the collaboration with external researchers is approved, a decision will also be made regarding which group(s) should perform the statistical analyses proposed, based on the subject matter of the collaboration and also on their more direct involvement with the variables being requested. The authorship for PREDIMED PLUS researchers must be established according to the publication policy, with an effort made to ensure that the maximum possible number of PREDIMED PLUS researchers appear as authors. If the request involves variables from the general project collected from all of the participants or from a majority of them, the researchers from each node who will be included as authors on the various publications generated will be established using a rotation system. If the data involves only one node or a small number of nodes, the authors from those nodes will be preferentially included. If the collaboration involves variables generated during a specific project, the authors from the nodes that generated those variables will also be given preference.

For articles written for publication using such data, proposals and draft versions must be submitted for review by the Steering Committee, which will manage the information according to the procedures for the study.

It is possible that a fee could be charged to researchers requesting aggregate data, based on the amount of work that would be required in relation to the analyses proposed. Payment of that fee would be made into the bank account created for that purpose and managed by the CIBER-OBN, with direct allocation to the needs of the PREDIMED PLUS study (fundamentally dedicated to contributing to the payment of publication expenses for the various PREDIMED PLUS articles).

2.2.2.A:2.2: Collaboration in which the external researchers are requesting individual raw data for participants in the PREDIMED PLUS study

Collaborative projects in which the data requested are broken down at the individual level for participants in the PREDIMED PLUS study could have greater repercussions in terms of confidentiality and use of the data, and therefore the researchers making such requests will be subject to additional requirements in order to ensure compliance in relation to these issues. The current management system being used by the PREDIMED PLUS study to handle data does not provide data access security of the **virtual data enclave** type (Vie et al., 2013) for external researchers. This makes it a non-optimal system in terms of data sharing (since external researchers must be given the data file with the variables requested, which they could then copy, alter, or distribute in an uncontrolled manner), and one of our objectives for the future is therefore to improve the data management system. As mentioned in the previous section, the optimal data management system that we are taking as a reference is the one implemented at Harvard University for the "Nurses' Health Study" (Bao et al., 2016). That particular system has implementation and maintenance costs, but similar, smaller-scale alternatives can also be developed. Implementation of that type of data management system would provide controlled access to the variables requested by researchers working on collaborative projects, without giving them an entire data file. Instead, they perform the statistical analyses directly, using a personalised login and following the pertinent syntaxes, on a central computer that has the specific software used for their analyses installed. However, until that system is implemented, access to data will be subject to security restrictions, although with a variety of strategies now being applied in order to minimise these.

The data request document is attached as Appendix 3, and the Guide to Requests is specified in Appendix 4. **The general procedure that external researchers requesting raw data must use** will be as follows:

Since the numerous groups that are part of the PREDIMED PLUS study already produce a high level of internal demand for data analysis, proposals submitted by external researchers must be clearly justified, and must not overlap with similar proposals already in progress and being performed by internal researchers, or those that appear as objectives of the various research projects.

Collaboration proposals may be submitted for both general data from the PREDIMED PLUS study and for specific projects. In the case of specific projects, the coordinator of the corresponding specific project must be the one maintaining the most direct contact with the external researcher submitting the collaboration request.

I. Submission of a collaboration proposal

The following steps must be followed in order to submit a proposal for collaboration and access to data:

I.1. Collaboration request letter:

This letter is addressed to the Steering Committee for the PREDIMED PLUS study, and in it the investigator must provide a general description of the collaboration request, along with a brief CV summarising his or her research activities. The benefits of the collaboration must also be described (see further details on the form attached as APPENDIX 4. Collaboration Request Form for External researchers).

I.2. Completion of the collaboration request form for external researchers *for the PREDIMED PLUS study, in order to access data generated during the PREDIMED PLUS study.*

This form is also used to submit further details about the study: hypotheses, objectives, methodology, variables requested, sample size, duration of the study, etc.

In cases involving ***meta-analyses of aggregate data****, where the researcher is not requesting access to raw data for the variables but only to coefficients and P-values corresponding to specific models of association, the variables that will be included in the analysis must be indicated in this section, along with the specific analyses to be performed, and including specification of the control variables and the syntaxes corresponding to the models.*

I.3. Submission and evaluation of the proposal

The proposal must be submitted electronically to the Steering Committee for the study, during the first week of each month. Each proposal will be assessed and assigned a status: accepted, corrections needed, or rejected.

The fundamental criteria used to evaluate the proposals will be the following:

- *The scientific excellence of the proposal*
- *The strategic priority of the proposal, with respect to the PREDIMED PLUS studies in progress, in order to avoid unnecessary duplication of work*
- *The scientific research background of the person or group submitting the request*
- *The scientific quality and capacity to administer the project*
- *The proposed study's compliance with the contents of the informed consent form signed by the patients whose data are being requested*

II. Conditions for accessing data requested from the PREDIMED PLUS study

If the proposal is accepted, a ***Collaboration Agreement*** *(Appendix 9) will first need to be signed. This document contains details on the terms and conditions under which access to the requested data is being granted, and these terms and conditions must be accepted by the collaborating researcher. New authorisation from the PREDIMED PLUS study's Ethics Committee may also be required, depending on the type of study proposed (the centralised committee or those from the various recruiting nodes from which data are being requested).*

Since the PREDIMED PLUS study does not currently have a data management system based on data enclaves, which would give external researchers the ability to access data for the variables securely but without the ability to modify or copy them, some additional requirements must be specified in order to ensure the security, integrity, and confidentiality of the data.

External researchers will be given a password which is used to access a data file that contains the variables requested. In order to better protect the identities of the participants, this data file will be anonymised, without any variables that could serve as identifiers. The corresponding metadata will also be provided. External researchers must make a commitment to:

- refrain from attempting to identify the individual patients, while maintaining confidentiality;

- maintain the integrity of the data;

- maintain secure data storage environments;

- refrain from distributing any data to unauthorised third parties;

- use the data provided only in relation to the specific objectives for the collaboration;

- destroy the data provided at the end of the active established period for the Collaboration Agreement.

*In the case of **collaboration proposals for performing meta-analysis of aggregate data**, the external researcher will not be requesting raw data for the variables, but instead only performance of statistical analyses with provision of the coefficients, and therefore the situation related to the security and integrity of the data for the variables remains internal to the PREDIMED PLUS study itself.*

III. Data analysis and publications

Like internal researchers, external researchers must ensure that statistical analyses are performed with attention to quality and best practices. Details must be recorded on the software used, and the syntaxes corresponding to each article must be saved and submitted, along with the data used.

*In the case of proposals for **meta-analyses**, the statistical analysis will be performed by an internal researcher from the PREDIMED PLUS study assigned based on the experience required for the collaboration, and in this case the aggregate results requested will be provided in the form of tables.*

The publication proposal form must be completed, including the names of the authors of the publication(s). In the case of original studies, the publication policy will be followed when establishing authorship, taking into account whether the collaboration involves data from the general study and for the total sample or if specific projects are involved.

In the case of collaborative studies or meta-analyses in which a large number of authors from other groups are participating and the number of authors that can be included is limited, the general rule to be followed is to include the maximum number of authors from the PREDIMED PLUS study, also establishing an ongoing rotation system for authorship of the various meta-analyses.

The publication proposal must be approved by the Steering Committee, and the results in the form of tables and the full article must be submitted before they are sent in for publication, so that the appropriate quality controls can be performed.

IV. Cost

The procedures carried out by the researchers from the PREDIMED PLUS study incur a cost in terms of their time investment, including their review of the data access request, preparation of the data, provision of access to the data and metadata, statistical analyses in some cases, and review of the publication proposal and results. Therefore, like other large international groups have done, a fee is being established in relation to data access, which will be collected after a data request has been approved. A bank transfer will be used to deposit this amount into the PREDIMED PLUS (CIBER OBN) account opened for this purpose. Depending on the type of collaboration involved, this fee would vary from situations with full exemption up to a

cost of €5,000 (see the document on "Fees for data and samples based on type of collaboration requested and volume of data required"; Appendix 10).

2.2.2.A3. Procedure for access/sharing with external researchers who are only requesting samples

External researchers requesting samples only must complete the request form attached as Appendix 5, and must follow the corresponding Guide found in Appendix 6. They will also need to establish a specific collaboration agreement for requesting samples, as detailed in Appendix 9.

The process would essentially be as follows:

In this case, the external researcher is only requesting samples to perform certain types of analyses, which involves no additional costs for the PREDIMED PLUS study. These external researchers also receive a commitment from researchers from the PREDIMED PLUS study, who will later establish the link between the results from the samples analysed and the corresponding data from the PREDIMED PLUS study. The internal researcher will in turn have to request authorisation to formalise the collaboration from the centres that recruited the patients whose data and samples will be analysed.

Special value will be given to collaborations with relevant and novel objectives, and to those in which the collaborating researcher could provide other data for comparative purposes or to increase the sample size. Proposals to evaluate highly speculative hypotheses or those with ethical restrictions are not considered appropriate.

Since the numerous groups that are part of the PREDIMED PLUS study already produce a high level of internal demand for sample analysis, and given the finite nature of the samples, proposals submitted by external researchers must be clearly justified, and must not overlap with similar proposals already in progress and being performed by internal researchers, or those that appear as objectives of the various research projects.

I. Submission of a collaboration proposal

The following steps must be followed in order to submit a proposal for collaboration and access to samples:

I.1. Production of a collaboration request letter:

This letter is addressed to the Steering Committee for the PREDIMED PLUS study, and in it the external researcher must provide a general description of the collaboration request, along with a brief CV summarising his or her research activities. The benefits of the collaboration must also be described (see further details on the form from APPENDIX 5. Collaboration Request Form for External Researchers for Access to Samples).

***I.2. Completion of the collaboration request form for external researchers** for the PREDIMED PLUS study, in order to access samples generated during the PREDIMED PLUS study.*

This form is used to provide further details about the study: hypothesis, objectives, methodology to be used, PREDIMED PLUS researcher who will perform the data study, variables to be analysed, sample size, duration of the study, type of samples being requested,

quantity, concentration, conditions for transport, shipping location, determinations to be performed, etc.

I.3. Submission and evaluation of the proposal

The proposal must be submitted electronically to the Steering Committee for the study during the first week of each month. Each proposal will be assessed and assigned a status: accepted, corrections needed, or rejected.

The basic criteria used to evaluate the proposals will be the following:

- *The availability of samples*
- *The scientific excellence of the proposal*
- *The strategic priority of the proposal, with respect to the PREDIMED PLUS studies in progress, in order to avoid unnecessary duplication of work*
- *The scientific research background of the person or group submitting the request*
- *The quality of the group that will be performing determinations using the samples*
- *The availability of the PREDIMED PLUS researcher(s) supporting the sample sharing proposal, and their commitment to carry out data analysis and/or collaboration with other internal PREDIMED researchers*
- *The type and availability of other data that would be combined with the results from the sample analysis*
- *The proposed study's with the contents of the informed consent form signed by the patients whose data is being requested*

II. Conditions for access to data requested from the PREDIMED PLUS study

If the proposal is accepted, a ***Collaboration Agreement for Access to Samples*** *(Appendix 9) will first need to be signed. This document contains details on the terms and conditions under which samples are being transferred to the external researcher, and these terms and conditions must be accepted by the collaborating researcher. An agreement must also be established with the internal PREDIMED PLUS researchers, and their permission must be requested to use the data that the analyses performed on the samples will require.*

New authorisation from the PREDIMED PLUS study's Ethics Committee may also be required, depending on the type of study proposed (the centralised committee or those from the various recruiting nodes from which data are being requested).

Since in this case the internal PREDIMED PLUS researchers are the ones who will perform the data analysis, the data file for the variables does not need to be given to the external researcher. This means that the internal researchers from PREDIMED will be the ones responsible for ensuring compliance with the principles of security, confidentiality, privacy, data integrity, and non-distribution to third parties, and they will fill in the data access commitment form.

The data requested may be either from the General Project or from Sub-Projects, and the internal researchers must be the ones to request those data. This will therefore require the collaboration of the researchers who generated those data (general and/or specific). Access to the data will be provided by means of transfer of the corresponding data file, after these agreements have been established.

External researchers gaining access to samples (which will be sent to them at the address specified) are committed to:

- maintaining the security and integrity of the samples, ensuring that transport takes place under appropriate conditions of freezing/refrigeration and storage;

- using the samples exclusively for the determinations approved by the PREDIMED PLUS study, with no permission to perform additional determinations or to transfer samples to other researchers;

- returning any remaining samples under optimal conditions of preservation following the determinations;

- submitting a copy of the results from the determinations performed to the central database for the PREDIMED PLUS study, or to the federated database that will be designated for this purpose if there are technical restrictions on the formats, in order to allow those data to be incorporated after the processing period has ended, which is established as 2 years after reception. In exceptional cases it will be possible to request an extension. Additionally, whenever a paper is produced using determinations performed on samples from multiple nodes, each node must be delivered a copy of the data corresponding to the samples it provided.

- establishing the pertinent collaboration agreements with internal PREDIMED PLUS researchers, who will carry out data analyses using the determinations produced;

- adhering to the publication policy and data management and sharing plan for the PREDIMED PLUS study.

III. Data analysis and publications

External researchers must ensure the quality of the determinations performed on the samples. Each node that has provided samples used for the determinations must be provided with, in addition to the data, details on the methodology used, the type of data generated, the quality control performed, and the metadata corresponding to each variable measured.

The internal researchers will perform the statistical analyses after establishing linkage with the pertinent PREDIMED PLUS variables, and they will provide the corresponding syntaxes and data for each article submitted for publication.

The internal researchers, together with the external ones, must complete the publication proposal form, including the names of the authors of the publication(s). In the case of original studies, the publication policy will be followed when establishing authorship, taking into account whether the collaboration involves data from the general study and for the total sample, or if specific projects are involved, and also considering the contributions made by the external researchers and internal researchers involved.

In the case of collaborative studies or meta-analyses where a large number of authors from other groups are participating and the number of authors that can be included is limited, the general rule to be followed is to include the maximum number of authors from the PREDIMED

PLUS study, also establishing an ongoing system of rotation for authorship of the various meta-analyses.

The publication proposal must be approved by the Steering Committee, and the results in the form of tables the and full article must be submitted and approved before they are sent out for publication, so that the appropriate quality controls can be performed.

IV. Cost

In addition to an economic cost, the procedures that must be carried out in order to locate samples, prepare them, ship them using dry ice, etc. incur a cost in terms of the time investment for the researchers from the PREDIMED PLUS study. Therefore, like other large international groups have done, a fee is being established in relation to data access, which will be collected after a data request has been approved. A bank transfer will be used to deposit this amount into the PREDIMED PLUS (CIBER OBN) account opened for this purpose. Depending on the type of collaboration involved, this fee would vary from situations with full exemption up to a cost of €5,000 (see the document on "Fees for data and samples based on type of collaboration requested and volume of data required").

2.2.2.A4. Procedure for access/sharing with external researchers requesting samples and data

This case normally involves an additional request for a research project between PREDIMED PLUS researchers and external researchers. The request document attached as Appendix 7 must be completed, and the Guide appearing in Appendix 8 must be followed. This request has characteristics similar to a combination of the two previous situations described for data and samples, plus the request for data from the research project. In this case a **collaboration agreement** for a request of both data and samples will have to be established, as detailed in Appendix 9. The combined requirements from the documents on access to data and on access to samples will be applied.

2.2.2.A5. Procedure for data sharing with the general public

Given the ethical restrictions on the data generated during the PREDIMED PLUS study, no sharing of raw data with the general public is being considered. This means that raw data from the PREDIMED PLUS study will not be deposited in repositories where access is completely open. If an article is submitted to a journal that requires this type of deposit, the journal must be notified that the data analysed are subject to restrictions based on ethical considerations, but that it will be possible to provide data to qualified researchers on receipt of a formal request (the data request form must be completed).

Data shared with the general public can include metadata, original articles on the study's results, documents, videos, educational materials, and similar.

As an extension of the considerations above, in some special cases it may be possible to consider providing, in parallel with the publication of the primary final results of the project in a high-impact journal, a small database containing the specific variables from the article, in a manner available to the general public, if the journal so requires and if there is significant value added to justify such a decision. In order to do this, specific authorisation would have to be requested from the Ethics Committee from each recruitment node, with specific details on the variables that would be released, and the procedures used to ensure their anonymisation. If

each Ethics Committee decides in favour, then the data could be released in this way. However, if any node decides otherwise, the data corresponding to that node could not be included.

2.2.2.B. Specify the methods used or software required in order to access the data

The current data management system used for the PREDIMED PLUS study only allows the distribution of data files. These data files can be provided in text format or in the most common formats used by the various statistics packages (SPSS, STATA, etc). A personalised password is used to gain online access to these files.

If a data enclave system is available in the future, access to the data will take place via that system, which will also include the software needed for accessing and analysing the data.

2.2.2.C. Specify where (in which repositories, enclaves, etc.) the data and corresponding metadata will be deposited

The data are deposited at the various nodes where they are generated, as well as in the central database at the IMIM. It is possible to establish a more dynamic federated database and so required. The data files that are generated periodically and those that are accessed by means of authorised requests are also distributed to the various nodes, which become responsible for their security.

A repository dedicated only to the PREDIMED PLUS study will be created, to be used for depositing data files generated for publications that include a commitment to providing controlled access to data if requested by researchers. This will be a closed-access system, requiring those researchers to first be identified and authorised.

The metadata and documents will be deposited in open-access public repositories (those from the corresponding universities, Zenodo, etc.). If project funding is obtained from the NIH, the data involved will have to be deposited in the repository specified by that funding body, but while always maintaining controlled access because of the type of data and the associated ethical restrictions. The dbGaP is one such repository (further details on repositories appear in the sections below).

In order to ensure maximum security, confidentiality, privacy, and best practices in relation to access to and use of the data, there are plans to create a virtual data enclave system for the PREDIMED PLUS study in the future (see further details in the sections below).

2.2.3. Making the data interoperable

2.2.3.A. Establishing interoperability for the data: specify the data, metadata, vocabularies, standards, or methodologies that will be followed in order to facilitate interoperability

To facilitate maximum interoperability, the data will be stored in platform-agnostic formats. The data on spreadsheets will be stored as .cvs files, the data in text files will be stored as .txt, and images will be stored as .txt. In cases of data that must be stored in some other format, whether this may be their own proprietary format or any format other than those expressed above, the corresponding software will also be associated with the data (including version number), along with the related INFO.txt file.

Other data models will be adjusted to the exchange structures broadly defined in the standards, including: a) semantic files in XSD format (XML Schema Definition), classified based on their contents; b) explanatory documents in portable document format (PDF files).

For metadata, the standard codes will be followed whenever possible.

2.2.3.B. Will standard vocabularies be used for all data types generated?

Standard vocabularies will be used for the data generated whenever possible. The set of standards provided by Health Level Seven (HL7) on electronic sharing of clinical information will also be applied whenever applicable. In the same way, specific vocabularies will be used for clinical concepts when appropriate, taking SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) as a reference. Internal codes will be used for other types of data, which will be clearly specified in the corresponding file.

2.2.4. Making the data reusable

2.2.4. A. Explain how it will be possible to reuse the data and what licenses will be used

Dissemination and reuse of the data will take place in accordance with the ethical and legal context in which the PREDIMED PLUS study is taking place.

Privacy and confidentiality: PREDIMED PLUS uses data of a personal nature (identity data and health-related data), so the contents of the legislation on data protection must be complied with (regulated in Spain by the Personal Data Protection Act – Ley de Protección de Datos de Carácter Personal). That legislation applies to personal data recorded on all types of media. Data processing refers to all activities carried out for collecting, recording, storing, recovering, querying, using, and disseminating data. This, along with the content of the informed consent form signed by the participants, places restrictions on reusing the data in an open-access manner, with the need to implement restrictions instead.

Taking these factors into account, the data (in the broad sense, including documents and data summaries) from the PREDIMED PLUS study may belong to one of the **three general categories of data** recognised:

a. Public data: can be made available to any user, with open access and with no restrictions. Given the characteristics of the data from the PREDIMED PLUS study, there will be few occasions when this option can be used.

b. Restricted data: can only be viewed by certain users. This is the option that will be preferentially used during the PREDIMED PLUS study. Therefore, instead of making the data publicly available in open-access repositories, or as supplemental files to the articles published, it must always be stated that access to the data will be restricted, but that data may be provided to qualified researchers on request.

c. Private data: cannot be made public. These data are confidential. Some data from the PREDIMED PLUS study will always have to remain in this category.

In addition to confidentiality, users that are reusing data must comply with the terms and conditions from the licence and use permits, while also acknowledging the **intellectual property** rights of the researchers who produced the data. According to the legislation in Spain, intellectual property is the set of rights corresponding to authors or to other holders with

regard to works and materials that have been created. In Spain the primary legislation that regulates intellectual property rights is the Intellectual Property Act (Royal Legislative Decree 1/1996 of 12 April, which approves the consolidated text of the Intellectual Property Act), which has undergone various modifications. These include the changes introduced by Law 23/2006 of 7 July in order to adapt the Spanish legislation to the new circumstances created by the information society, and another subsequent modification in 2014.

The rights established as intellectual property rights include a distinction between moral rights and economic rights. Compared to systems in the English-speaking world, Spanish legislation provides a stronger defence of what are known as moral rights, which are recognised for authors of artistic works and for other artists performing or executing such works. These rights remain associated with the author for life and cannot be renounced or revoked. They include the right to be acknowledged as the author of a work and the right to demand respect for the integrity of a work or performance, including non-alteration. Rights that are economic in nature include the rights related to the use of a protected work or performance, which are in turn subdivided into exclusive rights and compensated rights.

However, in order to better summarise these rights, it may be useful to clarify the types of data that are considered a “work”. According to Article 12 of the cited consolidated text of the Intellectual Property Act, since collections of data and databases represent intellectual creations they are protected as intellectual property, in this case by what are known as *sui generis* rights. ‘The protection refers only to its structure for selection or availability of contents, in any form of expression,’ not to the actual data. Authorship rights belong to the creators, as long as the work is original.

Exploitation rights or **copyrights** are transferable. The holder of these rights has the exclusive ability to exercise them, which therefore cannot otherwise take place without the holder’s authorisation, except within the limits established by law. Exploitation rights cover a series of acts such as playback, distribution, public communication, and transformation.

Depositing **datasets** into a repository implies the exercise of exploitation rights, and it therefore requires explicit permission from the holder of those rights, which is granted by means of a non-exclusive assignment agreement for the rights required.

Other licenses as alternatives to copyrights: There are standard and open licenses that an author can apply to his or her research data, which provide the terms and conditions under which that data can be shared and reused via the Internet. An example of such licenses are those known as **Creative Commons**. Creative Commons is a not-for-profit corporation that was established based on the idea that some people may not want to exercise all of the intellectual property rights they are granted by law. Creative Commons licenses therefore allow works to be shared and reused under more flexible legal terms. Creative Commons offers six basic license types that allow for the copying, distribution, downloading, and transformation of digital documents. The following link can be used to download high-resolution Creative Commons logos: https://creativecommons.org/about/downloads/. Creative Commons (CC) licenses can be applied to any type of work, including educational resources, photographs, databases, and many other types of creative content. The four concepts applied are: ATTRIBUTION (BY): Any exploitation of the work authorised by the licence must give credit to the author (BY). NON-COMMERCIAL (NC): Exploitation of the work is limited to non-commercial uses. NO DERIVATIVE WORKS (ND): Authorisation to exploit the work does not

include transformation to create a derivative work. And SHARE ALIKE (SA): The exploitation authorised includes creation of derivative works, as long as the same licence is maintained when those works are distributed. The combination of these four concepts gives rise to the six types of licences. Attribution (**CC BY**); Attribution – Share alike (**CC BY-SA**); Attribution – No derivative works (CC BY-ND); Attribution – Non-commercial use (**CC BY-NC**); Attribution – Non-commercial use – Share alike (**CC BY-NC-SA**); and Attribution – Non-commercial use – No derivative works (**CC BY-NC-ND**).

CC0: Access is completely open. This license is used in cases where no rights are desired, whether for authorship or other similar rights, and these files tend to be referred to as fully public.

Normally the description of the data and metadata released or submitted to repositories takes place under CC0.

Depending on the type of data considered appropriate to share in this way (metadata, text documents for manuals, recommendations, aggregate data from statistical analyses, or certain variables that do not require confidentiality), PREDIMED PLUS will opt for one type of licence or another when depositing data into the corresponding repository, and taking into account the considerations expressed above on confidentiality, privacy, and intellectual property.

However, the general policy on the reuse of data during PREDIMED PLUS will be to use licences and repositories that allow only restricted access to the data deposited. In general, authorisation from PREDIMED PLUS for controlled access to the data is preferable.

The request for authorisation to reuse data must be submitted to the Steering Committee of PREDIMED PLUS, using the forms designed for this purpose. That request will be considered, and a decision will then be made on whether or not to provide access to the data for the researchers making the request.

2.2.4. B. Specify when the data will be available for reuse; specify the length of the embargo period required

In general, data will not be available for use immediately after they are generated. An embargo period will be specified, which will depend on each situation and on whether or not a research project is funded by international sources that specify compliance with a certain embargo period in their calls for proposals. In the same way, when the data involved are the raw data for an article published in a journal that specifies a certain time of availability for reuse, that specified time period will be given preference, as long as it is compatible with the PREDIMED PLUS policy on sharing data from published articles, which allows access to such data only after a request has been received in advance.

2.2.4. C. Describe how the data generated by the project could be used by third parties, especially after the project ends; specify any data that cannot be reused

In section 2.2.4.A. a general explanation was provided on how it will be possible to reuse our data and which licences will be used. This reuse of data can occur both while the project is in progress as well as after it has ended. Access to data while the project is still in progress will be more restricted, because of the purpose of the project itself and because of possible interference with the final results of the clinical trial. After the project has ended, access to the data may become more appropriate.

As a general principle for the PREDIMED PLUS study, all raw data generated during the project in relation to the participants via the various questionnaires and measurements will be used by the researchers from the PREDIMED PLUS study, and will not be publicly released for unrestricted use by third parties. Since the data involved is related to health, the principles of confidentiality and security recognised by Spanish legislation must be respected, along with the data sharing policies of the European Commission, and therefore this **raw data will not** be deposited **in public repositories with fully open access.**

However, it will be possible to provide restricted access to certain data for **qualified researchers** who **submit a specific request** for data, by filling out the data request form designed for that purpose, and after the corresponding authorisation has been received from the Steering Committee of the PREDIMED PLUS study. These data will be fully anonymised, with removal of any information that could reveal the patient's identity, even if only numerical, in order to minimise the possibility that the patient could be identified.

Principal investigators are never authorised to distribute the general or specific data files given to them, as mentioned in 1.2, to unauthorised personnel not belonging to their node. The PIs will have to ensure that the data files are not distributed to anyone outside of their research group.

Provision and **use of aggregate data** derived from multiple original variables is also being considered during the PREDIMED PLUS study, for any **researchers requesting such collaboration** by means of the corresponding form. The use of aggregate data presents fewer problems related to confidentiality and improper use of the data for the variables, and it is therefore considered as a preferred option for data sharing with other researchers internationally. These collaborations will also require approval in advance from the Steering Committee of the PREDIMED PLUS study.

Because of the type of health-related data being generated, the recommendation made by various agencies in relation to DMPs is to use a web-based portal created and maintained by the project's researchers, which in this case is an improved version of the official website for the PREDIMED PLUS study. After a request has been submitted by filling in the official form and obtaining the pertinent approval, the external researcher will receive a unique password that can be used to access the authorised, anonymised data for the variables. The current system of databases and file extraction means that outside users will have access to raw data for the variables, which they would be able to copy, modify, and analyse using their own software. Proper access and usage require a commitment to apply best practices, to refrain from altering the data, and to respect the time periods established by destroying the data for the variables once the time period stipulated for their use has expired.

However, as an ideal alternative it is recommended that the project have what is known as a "**data enclave**" system (Vie et al., 2013; Bao et al., 2016). Although this may require a larger financial investment, it will guarantee secure access to the data, eliminating the possibility of unauthorised copying, alteration, or later distribution of the data for the authorised variables. A data enclave is a secure data environment where authorised individuals can perform statistical analyses of the data, but under restricted access and without the ability to modify the data (https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#enclave).
There are multiple examples of data enclave systems that can be taken as references for the PREDIMED PLUS study, such as the one from the Harvard School of Public Health which includes what is known as the "Channing cohorts computer system" (Bao et al., 2016), or the

systems used at the University of Michigan (https://micda.psc.isr.umich.edu/enclave/) or University of Chicago (http://www.norc.org/Research/Capabilities/Pages/data-enclave.aspx), among others.

A data enclave systems allows outside users to make use of a desktop space on a central computer, where the various types of statistical analysis software are also installed. This allows external researchers to execute their instructions for analysis by using specific programming, obtaining the corresponding results as the output. This **access may take place in person** or **remotely via online access**. Although the PREDIMED PLUS study does not currently have this type of system, the groundwork is now being laid so that one can be implemented in the future.

Some of the data generated may be of interest to other research groups, and it will be possible to provide open access to this information and make it available to the entire scientific community via the project's website (predimedplus.com), after the intellectual property has been registered. These types include data generated in the form of public presentations of results and published articles, as well as the questionnaires, study protocol, dietary intervention materials, physical activity intervention materials, behavioural intervention materials, events being compiled, etc.

Although the first-choice option is to deposit the raw data at a portal or enclave created and administered by the PREDIMED PLUS researchers, depositing subsets of data in the **Zenodo** repository sponsored by the European Union is also being considered for a portion of the data for certain sets of variables, and as a long-term option once the study has ended.

This repository has been selected (https://zenodo.org) because it allows data to be deposited with closed access as well as with restricted access, under embargo, and open access (https://zenodo.org/policies). Restricted files can be deposited at Zenodo with the possibility of shared access only for those who meet certain requirements. These files will not be made available to the public, and sharing will only be possible if approved by the depositor of the original file. This repository is housed at the European Organization for Nuclear Research (known as CERN), and it uses Invenio software for the technical support. Access to metadata and data files is provided via standard protocols such as HTTP and OAI-PMH. Zenodo is the closed alternative to OpenAirPlus, since the European Commission is aware that certain data generated during medical research with human subjects cannot be openly shared, but instead requires restricted or closed access. Zenodo initially offers a long-term storage guarantee (20 years), as well as 50 GB of capacity per data set.

Although Zenodo is undeniably still in its infancy, and there may be some doubts about its long-term maintenance, the European Union assures us that Zenodo can be trusted, since the relevant common infrastructure funding has already been committed from the beginning, and CERN is also familiar with the preservation of large research data sets that may be multiple petabytes in size. In the unlikely event that Zenodo has to shut down, migration of all of its contents to other suitable repositories is guaranteed, and since all of the uploads have a DOI, none of the citations and links to the Zenodo resources would be affected. Zenodo is being developed by CERN as part of the European Union's FP7 project OpenAIREplus (funding agreement no. 283595) and OpenAIRE2020. In this repository, preservation of the files is guaranteed by means of a process in which the data files and metadata are backed up every night, with multiple copies housed on the online system. Furthermore, all data files are stored

along with an MD5 checksum for the file's contents. The files are regularly checked against their checksums to ensure that their contents have remained constant.

In addition to the data generated during the PREDIMED PLUS general project, there are other sub-projects that can generate new data, using additional funding that they will have to apply for. The new data obtained from these funded projects may be genomic, such as data from large-scale GWA studies. If the projects are funded by the NIH, the genomic data from these GWA studies, as well as the data for the main phenotypic variables, must be deposited in the Database of Genotypes and Phenotypes (known as **dbGaP**). This (Tryka et al., 2014) is a repository sponsored by the NIH, which is responsible for archiving, curating, and distributing information generated in studies investigating interactions between genotype and phenotype. The repository was launched in 2006 as a response to the development of the NIH's policy on GWA studies, and it provides access to genetic studies funded by the NIH and by other agencies around the world.

Access to **dbGaP** is **public**, which means that the general information submitted about the studies can be freely accessed, along with the data at the summary level and documents related to the studies (http://www.ncbi.nlm.nih.gov/gap). Depositing of aggregate data, metadata, or summary data from PREDIMED PLUS into dbGaP does not present any problems in terms of the open access that would be applied to them.

It is not possible to access data for individuals in the public access portion of dbGaP. That data is accessed via **restricted access to dbGaP** (https://dbgap.ncbi.nlm.nih.gov/aa). This type of access is only granted to qualified researchers upon request, after they have completed a form describing the objectives of the research and demonstrating the capacity to adequately protect the data. Currently, PREDIMED PLUS is taking the first steps towards forming a consortium sponsored by the NIH to perform randomised clinical trials with lifestyle intervention for weight loss, and it is likely that some funding can be obtained from the NIH for GWA studies. This therefore puts us in a situation where we must deposit part of the data from the PREDIMED PLUS study (genetics for the full genome and the corresponding phenotypes for anthropometric variables) into dbGaP, both in its public part and in its restricted access part. Depositing data in the restricted access repository requires institutional authorisation, after a favourable report has been obtained by the corresponding Ethics Committee. There is also an embargo period for this repository. In order to make use of the authorised access system, users that are not with the NIH must have an NIH eRA Commons account.

There are other European repositories operating in a manner similar to dbGaP. One of the more notable examples is the "European Genome-phenome Archive (EGA)" (Lappalainen et al., 2015). The EGA was launched in 2008 by the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) to support voluntary archiving of data that require secure storage. This archive's capacities were recently expanded through a collaboration with the Centre for Genome Regulation (CGR).

Finally, it is important to keep in mind the existence of the recently created **Vivli** platform (www.vivli.org). Although new requirements have been proposed for the sharing of data from clinical trials, there has not yet been an organised effort to coordinate the existing platforms and servers or to provide a basic platform that would give most data generators a simple way to share their trial data. However, something of this type is currently being created, sponsored by the Brigham and Women's Hospital Center for Multi-Regional Clinical Trials at Harvard

University (the MRCT Center). This platform is called Vivli, from the Greek word *vivlithíki* meaning library and Latin *viv* meaning life. The objective of Vivli is to connect existing data exchange platforms and communities, while at the same time accepting data from researchers who want to share data but lack the necessary resources (Bierer et al., 2016).

An **important consideration** in terms of depositing raw data at various repositories or on various devices is that the **data on the variables must be deposited without any type of identity information**. For security, it is recommended that the variable "identifier" should be eliminated (Tucker et al., 2016), even though it is a numerical identifier, since there is always a risk that the file connecting the PREDIMED PLUS IDs with the full names of the participants could be located.

2.2.4.D. Describe the quality control process used for the data generated

Internal quality control has been implemented for the data generated during the PREDIMED PLUS study. This includes all stages, from production of the data followed by detailed, standardised data collection protocols and manuals, through to the sending, storage, control of outliers and missing data, and double checking between those responsible for the central database and those at each node generating the data, based on periodic queries. In parallel, other control procedures are applied to the various nodes and processes, such as selective auditing. Essentially, each node is responsible for verifying the quality of its data, and it must establish quality controls to ensure that the data submitted to the database corresponds with the data obtained. Each node will also be responsible for reviewing its outliers, notifying the central database regarding any corrections that should be made in the case of errors, or else confirming the value reviewed. In parallel, in order to verify the validity of the data, a randomised quality control system will be established for the data at all nodes, carried out during the quality visits performed. This quality control follows the guidelines from the Guide to Best Practices published by the Inter-University Consortium for Political and Social Research (ICPSR, 2012) on the preparation and archiving of data throughout the entire data lifecycle.

One of the important aspects that must be decided on in order to allow consistent use of the data generated is whether or not to **assign values to missing data**. Files will initially be generated with no assignment of values for missing data. However, since many analyses will require assignment of data that supplements the real data for sensitivity analysis or for other types of analyses, for certain variables it may be considered appropriate to generate and assign data for missing values. Assignment of these data will take place in a centralised manner by staff members who are experts on the subject, and it will also be possible to make two or more groups responsible for this assignment, so they can compare methods and strategies for the assignment and agree on the best ones to use. The resulting assignment file will be made accessible to all PREDIMED PLUS researchers, so that their resulting publications can be produced using the same assignment values when applicable.

To ensure quality control for data processing and the generation of results, a syntax must be generated and saved for the statistical processing and data analysis corresponding to each article, which along with the corresponding data, must be made available to the authors of the article for consultation upon request. The lead authors for the article will be responsible for ensuring the quality of the data and analysis.

Another measure being planned for purposes of quality control and security is that once an article has been published, the main researcher for the article must provide the syntaxes used in relation to the statistical analysis of the data from the article, as well as the data, to the PI for the study, who will be responsible for making a repository available with the IT security guarantees required in order to ensure that access remains restricted.

2.2.4. D. Preservation of data

During the PREDIMED PLUS study, the data management system will continue to be improved using additional resources to ensure that the data will be preserved and remain usable for future research for as many years as possible. Since the study involves a pilot, the data from the general project will be generated gradually over a multi-year period (until 2020). After that, data will continue to be generated through specific projects, and by compilation of a larger number of events during a potential "extended follow-up". We foresee at least 15 years of active data preservation. The data management system will be designed to facilitate execution of the long-term data preservation plan, in conformity with international standards. In principle, all data generated will be preserved. Backup copies will be made regularly, so that they can be used if the need to restore the original files ever arises. As a minimum standard to be applied, an incremental backup will be performed for the files on a daily basis, and a full backup on a weekly basis. The integrity of the files will also be confirmed by verifying the MD5 checksum, the size of the file, and the file date. We will also address the issue of potential obsolescence of the hardware and software in the data preservation policy.

2.3. Storage process for data and resources

The PREDIMED PLUS study recently launched a new website to provide access to users and participants in the project. It has a public-access area as well as a private-access area for the research staff. The private content is protected by use of a password, and is only available in Spanish. Right now, a common password is used for each centre and member of the research staff. This represents a significant limitation in terms of quality, integrity, and security, and is something that must be improved immediately by providing a different password for each node that enters data.

The central Teleform platform is housed at the data coordination centre in Barcelona (the IMIM). This platform allows for management of electronic forms as well as printed forms and documents. The electronic forms are filled in by staff from the recruiting centre, which can be done using any common Internet browser with the Acrobat Reader plug-in, then sent to the IMIM electronically. Forms on paper are downloaded and processed by the staff from the recruiting centre, then sent to the IMIM by post when necessary. The entire set of electronic forms is being developed and managed using the Teleform Enterprise platform for automatic capture of forms and documents. There are also printed questionnaires that use optical reading, which are sent by post to the general data processing nodes, which in turn transfer the data to the IMIM to be stored in the central database. Similarly, the accelerometer and electro data are sent electronically to the respective processing centres and to the general IMIM database. At the IMIM, all of the data are integrated into a central database for storage, with quality control performed and backup copies made. A team of database administrators and managers are in direct contact with the coordinator of the PREDIMED PLUS study. Over the course of the study, general project data will continue to be stored in an ongoing manner,

and quality control will be performed on a daily basis by logging the daily incidents and by sending queries to the various nodes about any data that are missing or incomplete, outliers, etc.

Data files are periodically extracted for each node and sent to that node for local quality control, and once this has been completed, the IMIM is notified regarding any incidents or corrections made. The data reviewed by the nodes are then sent to the IMIM and the correction is saved in the database. This results in a more accurate version, which will be used to extract the full data files for the general project, which contain all of the information for all variables and participants in each group.

The data generated by the specific projects are initially stored at the node generating the data, in databases designed for this purpose and according to the characteristics of each data element generated. After an embargo period, which is specific for each type of project and data, the data are sent to the central IMIM database for storage. This also allows data files containing those data to be generated when appropriate. Depending on the nature of the data generated during the specific project, for example, if there are omics data for multiple high-density arrays, the data may require dozens of GB of storage, or the structure of some data may make it impossible to store them in conventional databases. One option in this case is to save the metadata for the results in the tables, with the results stored in their own unstructured repository separate from the databases. The repository or repositories would be interconnected with the databases under the federated database scheme. There are currently 2-5 MB, 3.5” hard disks being sold on the market with a 6 GB/sec SATA interface, rotation speed of 7200 rpm, 128 MB buffering, and NativeCommandQueuing technology. An initial capacity of less than 10 TB, which could be expanded as storage needs increase, can be achieved by mounting multiple hard disks with those characteristics in racks or on hot-swap hard disk backplanes with multiple bays and redundancy management. It is likely that solid state disk (SSD) technology will achieve notable progress in the next few years, with capacities similar to those of conventional hard disks becoming available with better speed specs and a significantly lower cost. If this occurs, SSD disks can be used to replace the conventional rotating hard disks.

2.3.1 Improvements to be introduced and cost estimation for making the data FAIR

The data management system currently available for the PREDIMED PLUS study has room for improvement and many limitations in relation to data sharing. This is because it can only distribute data files that provide direct access to all data from the study, which can put the security, integrity, confidentiality, and privacy of those data at risk, while at the same time allowing improper use of the variables. This is particularly true in cases in which external researchers are provided with data along with a commitment to destroy it after the active period of the agreement, which is a situation that in reality cannot be verified.

Therefore, the best option for making the data FAIR would be to implement a data management system based on the data enclave model (with in-person/virtual access), as described in section 2.2.4. C. This **data enclave** would be managed by using various privileges to provide access to certain variables, based on a profile assigned to both internal and external users. A data enclave system with real-time operation is very costly, and although costs for infrastructure and human resources can be minimised by operating a system only during specified periods, the system used for the PREDIMED study would have to operate every day

of the week in order to allow the database to be queried or analysed at any time. The cost for this cannot be estimated yet, but we are working on calculating a figure for the future. This system must guarantee:

- the security of the information stored in the databases.
- controlled access to the data.
- that the activities of the researchers accessing the data are supervised at all times.
- that, whenever possible, access is provided internally to data software to prevent the uncontrolled distribution of the data or unauthorised access to programs.
- that the exact scope of the data access is determined based on the privileges associated with each user's profile.
- the integrity of the data, preventing its dispersal and ensuring control over updating.
- efficiency in terms of the management and presentation of and access to the data by standardising the design for the database and applications on a single platform.

2.4. Data security

The purpose of a security policy is to guarantee the integrity, availability, and confidentiality of the data. Integrity means that no unauthorised modifications can been made to the data. This aspect is very important for PREDIMED PLUS, since the data management system currently being used for the study does not guarantee the integrity of data when distributed to various researchers (whether internal or external), since it provides direct access to the data for the variables.

Those responsible must therefore assign individual passwords to identify and authenticate the individuals authorised to view or process data; establish which persons can access the physical location where the files, etc. are kept; include a timeout feature for the system that closes a work session if nothing is being done; and ensure that computers cannot be manipulated via the Internet.

Availability means that the data are always available to the authorised individuals, and also that they can be recovered if any type of event, physical or otherwise, affects the normal operations.

Confidentiality means that the data can only be viewed and accessed by authorised users of the information system. Any security problems existing in this area could affect the duty to secrecy. Based on all of the above, the best option would be to use what is known as a data enclave.

This section is complemented by section 2.5 on ethical aspects, which contains details on the security requirements that must be applied to the files for participants in the PREDIMED PLUS study because they include health-related data and are therefore subject to the Spanish Data Protection Act (Organic Law 15/1999 of 13 December, on personal data protection, and its implementing Regulations, which were approved by Royal Decree 1720/2007 of 21 December). That Act has been supplemented recently by **REGULATION (EU) 2016/679** OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April **2016** on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

The new General Data Protection Regulation (GDPR) entered into force across Europe in May 2016, and will be applicable beginning in May 2018. During this transitional period, and even though the provisions from Directive 95/46 remain in force along with the corresponding national implementing legislation, the data controllers and processors must adopt the necessary measures in order to be prepared for compliance with the provisions from the GDPR at the time when they become applicable. The GDPR is directly applicable legislation, which means that it does not require transposition into national legislation, or require regulations for implementation or application in the majority of cases. Data controllers must therefore assume above all that the legislation of reference is the GDPR rather than the national legislation, as was the case up until now with Directive 95/46. Nevertheless, the legislation that will replace the current Organic Law on Data Protection (known as the LOPD in Spanish) will be able to include some clarifications or further developments, in relation to subjects allowed by the GDPR. In November of 2017 the new Data Protection Act was passed in Spain, but procedures are still being carried out in preparation for its applicability beginning in May 2018. The novel aspects of the GDPR are discussed in detail in the Guide to the General Data Protection Regulation for Data Controllers (Guía Reglamento General de Protección de Datos para responsables de tratamiento), which is a digital document produced by Spain's Data Protection Agency.

From the perspective of both ethics and security it is crucial to maintain the confidentiality of the data by means of a variety of strategies to eliminate any possibility of the patient's identity being revealed. In addition to direct identifiers, which include numerical IDs that could allow a person to be identified directly, there are other types of **identifiers, known as indirect identifiers**, which can be combined in various ways to allow a person to be identified. In addition to clinical data and lifestyle data, genetic data could also allow an individual to be identified (Check-Hayden, 2013), although such data are essential in studies where direct sequencing or high-density genome analysis is involved. This makes it advisable to keep high-density genotyping data in separate files from those used for other clinical or lifestyle data.

Anonymised files will be used as a general rule, without microdata on first names, surnames, national ID documents (DNI in Spanish), social security numbers, etc. Internally, however, the numerical ID generated during the PREDIMED study will be used for each patient. This ID is susceptible to being used for the unauthorised identification of the patient, as it includes the number of the node and a sequential patient number. In addition, if the files containing the link between the patient's name and the PREDIMED PLUS ID are not handled with maximum security, this information could easily be leaked. Therefore, when data is being shared with other researchers, the PREDIMED PLUS ID must be removed from the data file, and if there is a need to include some sort of ID, various anonymisation algorithms must be used, such as hash algorithms like SHA-1 or MD5. These allow the ID to be replaced by what is known as a randomised digital password or fingerprint, which acts as an unequivocal, inviolable identifier for the person.

According to Spain's Personal Data Protection Agency in its Guide for Anonymising Data (Guía para la anonimización de datos, 2016), the process of anonymisation cannot guarantee, in absolute terms, that re-identification of individuals will be impossible, which is why the legally established requirements that are necessary for preserving the rights of the data subjects must be taken into account.

In any database management system, there are three fundamental aspects of security that must be guaranteed: (a) protection of the system against external attacks; (b) protection of the system against software crashes and hardware malfunctions; (c) protection against manipulation by unauthorised users. The security of the communications and information sent between the central database and the nodes must also be ensured.

Securing the system against software crashes and hardware malfunctions is based on the concepts of redundancy and backup. The operational value of the database management system is not sufficiently critical to require availability of a hot site that could take over in a few hours in the event of a service failure. However, there is a cold site for the IT system that supports the central database and public data repository that will be generated during the PREDIMED PLUS study, and there is a certain degree of redundancy for the databases under the responsibility of the nodes. A protocol for the production of security copies is applied at all of the centres, which in general consists of making incremental copies at night, full copies every week, a cycle of six incremental copies (one business week plus one day), and five full copies (one month plus one week), all done automatically. A duplicate of each copy in the cycle is kept in a fire-resistant cabinet, in a different office that also has restricted access.

For protection against manipulation, a compartmentalised database access strategy is being implemented at the user level. The technology applied to encrypt and protect the information being transmitted online using the HTTPS protocol between the various databases included in the federated database management system is Transport Layer Security (TLS) Protocol Version 1.24 for public key, with cipher suites TLS_RSA_WITH_AES_128_CBC_SHA or AES_256_CBC_SHA256, 128 bytes in length. This guarantees the security of the encryption, interoperability, expandability, and efficiency in relation to data transmission.

2.5. Ethical aspects

The rules that apply to personal data go beyond imposing duties in relation to obtaining such data and obligations for ensuring that the data controller will handle the data appropriately. This is why the quality, security, and secrecy of the data must be guaranteed.

As explained in the previous sections, the PREDIMED PLUS study is a multi-centre clinical trial using human subjects, where personal data and health-related data are collected both directly from patients and by means of clinical records, after their consent has been obtained. In addition, each PI from a node who has registered a data file at the Spanish Data Protection Agency becomes the data controller for that file, taking on the legal obligation to comply with the specifications on the subject contained in the Data Protection Act. Organic Law 15/1999 of 13 December, on Personal Data Protection, and its implementing regulations which were approved by Royal Decree 1720/2007 of 21 December, define the data controller for the file or processing as ‘a natural person or legal entity, public or private, or administrative body, that alone or jointly with others decides on the purpose, content and use of the processing, although he does not effectively do it. Entities without legal personality acting as separate parties in the operation may also be data controllers.’ A filing system is understood to mean ‘any structured set of personal data which are accessible according to specific criteria, whatever the form or method of its creation, storage, organisation and access’ (article 5.1.k). Furthermore, processing is defined as ‘any operation or technical process, whether automated or not, that allows the collection, recording, storage, creation, amendment, consultation, use, rectification,

erasure, blocking or deletion, as well as the disclosure of data arising from communications, consultations, interconnections and transfers' (article 5.1.t).

As such, the principle of data security established in article 9 of Organic Law 15/1999 requires the data controller for a file to adopt the necessary technical and organisational measures to ensure the security of the personal data in that file and prevent any alteration, loss, or unauthorised processing or access. These measures had to be developed under Title VIII – of the implementing regulations for the LOPD, which were approved by Royal Decree 1720/2007 of 21 December. To assist data controllers in their adoption of the provisions established by the regulations, a Data Security Guide (Guía de Seguridad de Datos) summarises the security measures and verifications to be performed, and provides a model security document, all of which must be taken into account by the PIs from each PREDIMED PLUS recruiting node, as well as by the Steering Committee when maintaining custody of the general comprehensive file for the various nodes. In accordance with Article 44.3.h) of Organic Law 15/1999, it is a serious infringement to 'maintain files, premises, programs or hardware containing personal data without the security required by regulations.'

The type of population analysed and the data obtained make it compulsory to also comply with national and international ethical principles, as well as with all applicable legislation on the subject. Each PI must obtain the corresponding permission from the Ethics Committee from his or her centre and/or region, including a favourable report.

According to article 6, processing of personal data will require the unequivocal consent of the data subject, except in situations where the law has established otherwise. The participants in the study have been able to read the patient information sheet on the details of the study, and they have signed the general informed consent form for the PREDIMED PLUS study. That form contains two sections: a general section on consent for the study and a section on consent for genetic analysis. The option signed by the patient on the informed consent form will determine the scope allowed for subsequent use of the data. If it has not been specified on the informed consent form that open sharing of the patient's data is going to occur, then the data cannot be deposited in public repositories. This ethical restriction in relation to the PREDIMED PLUS study is currently limiting the sharing of data. This means that it will only be possible to consider the option of data sharing after a request for data has been received from a qualified researcher, with the access remaining restricted and only if the reason for the data sharing is related to the project's initial objectives. This aspect is important, since in many cases personal data that was collected for a specific purpose is later used for a purpose other than the one for which it was initially collected (also known as "secondary use" or "additional processing"). As a general principle enunciated in article 5 of the GDPR, processing of personal data for purposes other than those for which they were initially collected is only permissible when the new objective of the processing is compatible with the purposes for which the data were initially collected. In practical terms, the new draft version of the Data Protection Act restricts those aspects even further, along with the uses of the informed consent: it does not allow for use of a general informed consent form, but states instead that the informed consent document must be more specific, with details on the specific uses for which consent is being given.

For all of these reasons, until alternatives can be implemented for obtaining informed consent in better agreement with the current data sharing proposal, such as the dynamic informed consent proposed by Budin-Ljøsne et al. (2017), the data from the PREDIMED PLUS study must not be subject to storage under an open-access system.

During the PREDIMED PLUS study we will therefore follow the general recommendations for data sharing when patient health data is involved (Tucker at al., 2016), which are based on:

a) Anonymisation/de-identification of data: Those possessing the data are responsible for generating de-identified data sets, which are created in order to offer better protection for the patient's privacy by masking or generalising other identifiers;

b) Controlled access to the data, including the use of data sharing agreements with qualified researchers. A legally binding data sharing agreement should be established, including agreements to refrain from downloading or sharing additional data or attempting to identify the patients. Adequate levels of security must be used when transferring data or providing data access. One solution is use of a secure "data enclave" system, which provides additional safeguards.

2.6. Other aspects

2.6. A. Indicate whether other procedures have been followed for the DMP, as promoted by other financial entities or national or international guidelines

The general guidelines from the European Commission have been followed for production of the DMP for the PREDIMED PLUS study. These general guidelines are also promoted at the national level by the Spanish Foundation for Science and Technology (FECYT in Spanish) and by Spain's Science and Innovation Act.

None of the financial entities currently funding the PREDIMED study require our project to have a compulsory DMP or open data sharing. Instead, the effort to produce the DMP has been based on expectations for the future.

2.7. Additional support for development of the DMP

In order to create this DMP, an extensive review of the literature and legislation was performed in relation to its principle aspects, at both the national and international level, fundamentally within the context of the European Union and the USA. Advice has also been received from experts on each subject, and from the PREDIMED PLUS Data and Safety Monitoring Board (DSMB) (meeting of 20/01/2017), and via collaboration with an IT engineer who is a certified information systems auditor (CISA). We have made queries regarding the current limitations of the IT system and the possibilities for improving the data management system for the PREDIMED PLUS study, including advice on possible alternatives and the viability of the proposals for providing the DMP. This plan has also been reviewed by the Steering Committee and by the PREDIMED PLUS researchers, who have contributed substantial improvements.

The present DMP for PREDIMED PLUS has been created based on the model proposed by the European Commission, published in 2016 (European Commission, Directorate General for Research & Innovation. H2020, 2016), using the software provided by the CSUC consortium (Catalonia).

PREDIMED-PLUS Data Management & Sharing

SUMMARY OF KEY POINTS

In Barcelona on 27 April 2017

(Revised in Barcelona on 15/01/2018, revised in Madrid on 22/01/18 and 16/04/18)

1. Internal access to data

1.1 In order to ensure the privacy of the data, the general database must never include first names, surnames, or national ID document numbers (DNI in Spanish). Only a numerical code (ID) will be included to identify the participant. Only the recruiting node has the ability to associate this patient ID with the patient's name. Mechanisms will be implemented to prevent public disclosure of the relationship between a patient's full name and DNI number and the ID assigned. The correspondence between the ID and the respective full name and DNI will be stored on secure devices without connection to the Internet or any possibility for simple external distribution.

1.2 On request, following the procedure indicated in the "data management and sharing" document, the IMIM will provide all of the principal investigators from PREDIMED PLUS with a complete data file for the general variables from the study. This file will have a version number and a date of generation. It is expected that multiple files will be generated as the data continues to be processed. This will begin with generation of the basic data file, and will continue with the generation of data at 1, 3, 5, and 7 years. This version number and date of generation must be cited in the analyses of the data and in the publications. It will be possible to make a specific request for another type of data, or data for periods other than those mentioned, again by following the procedure indicated in the "data management and sharing" document.

1.3 As soon as possible, this process of periodic sending of data files will be replaced by a virtual data enclave system, where the data can be deposited and managed securely, with remote access for online viewing and analysis but without allowing downloads to any personal computer. This will guarantee the integrity and security of the data, which are limited under the current system.

1.4 All data that have been collected from 50% or more of the participants from the PREDIMED PLUS study, or that have been determined in cases of primary events, will be included in the general database, which will be made available to all PIs at the recruiting nodes. In order to publish manuscripts using these data, which are not part of the general project, a PI must contact the person(s) responsible for generation of that data in order to produce a proposal for an article and submit it to the Steering Committee according to the publication policy.

1.5 In the case of very costly analytical results (metabolomic, genomic, proteomic, metagenomic, epigenomic, etc.) that have used samples from multiple nodes, and where the expenses have been covered using funds primarily obtained by the node producing such results, and even though that node must send the results of the determinations to each node that provided samples at the time when the article corresponding to such determinations is being finished and sent out for review, the period of time that passes until acceptance of the paper for the node that performed the determinations must be respected, prior to accepting any proposals from other nodes to analyse those data. If another PI has an interest in those data before the paper is accepted, that PI must contact the person responsible for the determinations performed in order to establish a collaboration, including production of a joint proposal for an article or articles according to the publication policy. This contact can even be

established beginning at the time when the determinations are being initiated. In all cases, both before and after publication of the initial paper, the node that has performed the determinations will have priority in terms of proposing new papers related to those determinations, and the proposal of publications with co-lead authors is recommended in cases where another node has a substantial interest in the same subject.

1.6 If a researcher performs any new determination of a measured variable for PREDIMED PLUS participants that are from some other node, all of the other researchers from the recruiting nodes must be notified. Specifically, the PI from the node performing that determination must provide notification within a period of no more than two months.

2. External access to data

2.1 No raw data from the PREDIMED PLUS study will be deposited into repositories with access that is completely open. PREDIMED PLUS was initiated prior to issuance of the requirements on universal access to data by the ICMJE (2016) and by the European Commission, and therefore the PIs are not obligated to provide the entire scientific community with open access to the data used in publications. In addition, there is an ethical restriction on providing open access to the data, since the informed consent form signed by the patients does not authorise such access.

2.2 Principal investigators are never authorised to distribute the general or specific data files they are given, as mentioned in 1.2, to unauthorised personnel not belonging to their node. The PIs will have to ensure that the data files are not distributed to anyone outside of their research group.

2.3 Some of the data generated may be of interest to other research groups, and it will be possible to provide open access to this information and make it available to the entire scientific community via the project's website (predimedplus.com), after the intellectual property has been registered. These types include data generated in the form of public presentations of results and published articles, as well as the questionnaires, study protocol, dietary intervention materials, physical activity intervention materials, behavioural intervention materials, events being compiled, etc.

2.4 If an outside group wants to carry out research using data from PREDIMED PLUS, an agreement can be established. However, this will first require a detailed joint collaboration proposal (with at least one PI from PREDIMED PLUS participating). The Steering Committee for the PREDIMED PLUS trial must make a decision on whether or not the collaboration agreement can be signed, based on a detailed study and review of the ethical implications. This decision may also include the possibility of charging a fee.

REFERENCES

- Academic Research Organization Consortium for Continuing Evaluation of Scientific Studies--Cardiovascular (ACCESS CV). Patel MR, Armstrong PW, Bhatt DL, Braunwald E, Camm AJ, Fox KA, Harrington RA, Hiatt WR, James SK, Kirtane AJ, Leon MB, Lincoff AM, Mahaffey KW, Mauri L, Mehran R, Mehta SR, Montalescot G, Nicholls SJ, Perkovic V, Peterson ED, Pocock SJ, Roe MT, Sabatine MS, Sekeres M, Solomon SD, Steg G, Stone GW, Van de Werf F, Wallentin L, White HD, Gibson M. Sharing Data from Cardiovascular Clinical Trials--A Proposal. N Engl J Med. 2016 Aug 4; 375(5):407-9.
- Spanish Data Protection Agency (*Agencia Española de Protección de Datos*). Orientations and guarantees in Procedures for anonymizing personal data (*Orientaciones y garantías en los procedimientos de anonimización de datos personales*). 2016. Digital document.
- Spanish Data Protection Agency (*Agencia Española de Protección de Datos*). Guide to the General Data Protection Regulation for data controllers (*Guía del Reglamento General de Protección de Datos para responsables de tratamiento*). 2016. Digital document [http://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/index-ides-idphp.php].
- Amarnath Gupta, William Bug, Luis Marenco, Xufei Qian, Christopher Condit, Arun Rangarajan, Hans Michael, Müller Perry L., Miller Brian, Sanders Jeffrey, S. Grethe, Vadim Astakhov, Gordon Shepherd, Paul W. Sternberg, Maryann E. Martone. Federated Access to Heterogeneous Information Resources in the Neuroscience Information Framework (NIF). Neuroinformatics. 2008; 6(3):205–217.
- AREA FOR PROFESSIONALS [http://predimedplus.com/acceso-equipo-medico/]
- Bao Y, Bertoia ML, Lenart EB, Stampfer MJ, Willett WC, Speizer FE, Chavarro JE. Origin, Methods, and Evolution of the Three Nurses' Health Studies. Am J Public Health. 2016; 106(9):1573-81.
- Bauchner H, Golub RM, Fontanarosa PB. Data Sharing: An Ethical and Scientific Imperative. JAMA. 2016; 22-29;315:1237-9.
- Bierer BE, Li R, Barnes M, Sim I. A Global, Neutral Platform for Sharing Trial Data. N Engl J Med. 2016; 374(25):2411-3.
- Bruland P, Breil B, Fritz F, Dugas M. Interoperability in clinical research: from metadata registries to semantically annotated CDISC ODM. Stud Health Technol Inform. 2012;180:564-8
- Budin-Ljøsne I, Teare HJ, Kaye J, Beck S, Bentzen HB, Caenazzo L, Collett C, D'Abramo F, Felzmann H, Finlay T, Javaid MK, Jones E, Katić V, Simpson A, Mascalzoni D. Dynamic Consent: a potential solution to some of the challenges of modern biomedical research. BMC Med Ethics. 2017; 18(1):4.
- Callaghan, S., Tedds, J., Kunze, J., Khodiyar, V., Mayernick, M. Lawrence, R., Murphy, F., Roberts, T. and Whyte, A. (2014) 'Guidelines on recommending data repositories as partners in publishing research data' Proceedings IDCC14, San Francisco, Feb. 25, 2014
- Chassang G. The impact of the EU general data protection regulation on scientific research. Ecancermedicalscience. 2017; 11:709.
- Check-Hayden E. Privacy protections: the genome hacker. Nature. 2013; 497:172–174.
- Consilium Teleform webpage [http://www.teleform.nl/index.php?menu=290]
- Dankar FK, Badji R. A Risk-Based Framework for Biomedical Data Sharing. J Biomed Inform. 2017 Jan 23. pii: S1532-0464(17)30016-3.
- DCC (2014) Five steps to decide what data to keep: a checklist for appraising research data v.1 Edinburgh: Digital Curation Centre. Available online: www.dcc.ac.uk/resources/how-guides
- Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, Minion J, Boyd AW, Newby CJ, Nuotio ML, Wilson R, Butters O, Murtagh B, Demir I, Doiron D, Giepmans L, Wallace SE, Budin-Ljøsne I, Oliver Schmidt C, Boffetta P, Boniol M, Bota M, Carter KW, deKlerk N, Dibben

C, Francis RW, Hiekkalinna T, Hveem K, Kvaløy K, Millar S, Perry IJ, Peters A, Phillips CM, Popham F, Raab G, Reischl E, Sheehan N, Waldenberger M, Perola M, van den Heuvel E, Macleod J, Knoppers BM, Stolk RP, Fortier I, Harris JR, Woffenbuttel BH, Murtagh MJ, Ferretti V, Burton PR. DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol. 2014; 43(6):1929-44.

- Dyke SOM, Dove ES, Knoppers BM. Sharing health-related data: a privacy test? NPJ Genom Med. 2016; 1:160241-160246.
- EUROPEAN COMMISSION, Directorate-General for Research & Innovation. H2020 Programme Guidelines on FAIR Data Management in Horizon 2020. Version 3.0. 26 July 2016.
- EUROPEAN COMMISSION, Directorate-General for Research & Innovation. H2020 Programme. Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Version 3.2. 21 March, 2017.
- FORCE11. THE FAIR DATA PRINCIPLES. 2016 [https://www.force11.org/group/fairgroup/fairprinciples].
- *Grup de Treball de Suport a la Recerca del CSUC* (Working Group to Support the Research of the CSUC). Data Management Plans. Version 2, December 2016. [Doc.16/61: B6SR\GT SR\GDR\PGD_v2Publica_desembre16-EN.docx, 22.12.16].
- *Grup de Treball de Suport a la Recerca del CSUC* (Working Group to Support the Research of the CSUC). Recommendations for selecting a repository for depositing research data (In Catalan: *Recomanacions per seleccionar un repositori per al dipòsit de dades de recerca*). Version 2, November 2016. [Doc. 16/57: B6SR\GDR\1611RecomanacionsSeleccionarRepositoriDades.docx, 15.11.16].
- Working Group on "Depositing and Management of Data under Open Access" for the RECOLECTA project. Conservation and reuse of scientific data in Spain (*La conservación y reutilización de los datos científicos en España*). Report from the working group on best practices. Madrid: Spanish Foundation for Science and Technology (FECYT) (2012) [http://eprints.rclis.org/21007/1/informe_datos_cientificos_en_esp.pdf].
- Hsiao, D.K. Federated databases and systems: Part I – A tutorial on their data sharing. VLDB Journal. 1992; 1:127. doi:10.1007/BF01228709.
- International Consortium of Investigators for Fairness in Trial Data Sharing. Devereaux PJ, Guyatt G, Gerstein H, Connolly S, Yusuf S. Toward Fairness in Data Sharing. N Engl J Med. 2016 Aug 4; 375(5):405-7.
- Internet Engineering Task Force (IETF). Request for Comments (RFC) 5246, Aug. 2008. [https://www.rfc-editor.org/info/rfc5246].
- Inter-university Consortium for Political and Social Research (ICPSR). (2012). Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (5th ed.). Ann Arbor, MI. ISBN 978-0-89138-800-5.
- Kalager M, Adami HO, Bretthauer M. Recognizing Data Generation. N Engl J Med. 2016 May 12;374(19):1898
- Klann JG, Abend A, Raghavan VA, Mandl KD, Murphy SN. Data interchange using i2b2. J Am Med Inform Assoc. 2016; 23(5):909-15.
- Landau S. Control use of data to protect privacy. Science. 2015; 347(6221):504-6.
- Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R, Vaughan B, Laurent T, Rowland F, Marin-Garcia P, Barker J, Jokinen P, Torres AC, de Argila JR, Llobet OM, Medina I, Puy MS, Alberich M, de la Torre S, Navarro A, Paschall J, Flicek P. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet. 2015; 47(7):692-5.
- Law 14/2011 of 1 June, on Science, Technology and Innovation (LCTI in Spanish). Official State Gazette (BOE in Spanish) 02/06/2011; 131. Updating of 10/09/2015. [https://www.boe.es/buscar/pdf/2011/BOE-A-2011-9617-consolidado.pdf].

- Law 21/2014 of 4 November, which modifies the consolidated text of the Intellectual Property Act, approved by Royal Legislative Decree 1/1996 of 12 April; and Law 1/2000 of 7 January, on Civil Procedure: https://www.boe.es/diario_boe/txt.php?id=BOE-A-2014 11404
- Organic Law 15/1999 of 13 December on Personal Data Protection (LOPD in Spanish). Official State Gazette (BOE in Spanish) 14/12/1999; 298. Updating of 05/03/2011. [https://www.boe.es/buscar/pdf/1999/BOE-A-1999-23750-consolidado.pdf].
- Longo DL, Drazen JM. Data Sharing. N Engl J Med. 2016; 374:276-7.
- Menikoff J, Kaneshiro J, Pritchard I. The Common Rule, Updated. N Engl J Med. 2017 Jan 19. doi: 10.1056 (in press).
- MESA. Deidentified data distribution policy statement. https://www.mesanhlbi.org/PublicDocs/MESA_DeidentifiedDataDistribution_PolicyStatement_04122016.pdf. 2016.
- Mohammed EA, Slack JC, Naugler CT. Generating unique IDs from patient identification data using security models. J Pathol Inform. 2016; 7:55
- Musick BS, Robb SL, Burns DS, Stegenga K, Yan M, McCorkle KJ, Haase JE. Development and use of a web-based data management system for a randomised clinical trial of adolescents and young adults. Comput Inform Nurs. 2011; 29:337-43.
- NHLBI. NHLBI Research materials distribution agreement. https://biolincc.nhlbi.nih.gov/static/RMDA.pdf?link_time=2017-02-14_10:45:37.334691. Feb. 2017
- NIH. Data and Safety Monitoring Plan Writing Guidance. Guidance for Developing a Data and Safety Monitoring Plan for Clinical Trials Sponsored by NIMH. April 16, 2015. [https://www.nimh.nih.gov/funding/clinical-research/data-and-safety-monitoring-plan-writing-guidance.shtml].
- Nurses' Health Study. Guidelines for external collaborators for access to archived data. Website: http://www.nurseshealthstudy.org/researchers. Feb. 2017. Pdf document (http://www.nurseshealthstudy.org/sites/default/files/pdfs/Guidelines_Archived%20Data.pdf).
- PROYECTO PREDIMED-PLUS home page [http://predimedplus.com/]
- Rowhani-Farid A, Barnett AG. Has open data arrived at the British Medical Journal (BMJ)? An observational study. BMJ Open. 2016 Oct 13; 6(10):e011784
- Stefan Berger, Michael Schref. From Federated Databases to a Federated Data Warehouse System. Proceedings of the 41st Hawaii International Conference on System Sciences. 2008. DOI: 10.1109/HICSS.2008.178.
- Taichman DB, Backus J, Baethge C, Bauchner H, de Leeuw PW, Drazen JM, Fletcher J, Frizelle FA, Groves T, Haileamlak A, James A, Laine C, Peiperl L, Pinborg A, Sahni P, Wu S. Sharing Clinical Trial Data – A Proposal from the International Committee of Medical Journal Editors. N Engl J Med. 2016 Jan 28; 374(4):384-6.
- TELEform - Tecnomedia Sistemas SL website [http://www.oficinasinpapel.com/docs/TFSpanish.pdf]
- The European Prospective Investigation into Cancer and Nutrition (EPIC) study . The EPIC Access Policy. 2014. [https://epic.iarc.fr/docs/EPIC_Access_Policy_and_Guidelines.pdf].
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal L. 2016; 119(1).
- Tucker K, Branson J, Dilleen M, Hollis S, Loughlin P, Nixon MJ, Williams Z. Protecting patient privacy when sharing patient-level data from clinical trials. BMC Med Res Methodol. 2016; 16 Suppl 1:77.

- Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, Feolo M. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 2014 Jan; 42(Database issue):D975-9.
- Vie LL, Griffith KN, Scheier LM, Lester PB, Seligman ME. The Person-Event Data Environment: leveraging big data for studies of psychological strengths in soldiers. Front Psychol. 2013 Dec 13; 4:934.
- Visual Paradigm for Windows Community Edition, version 14.0. 2017. [https://www.visual-paradigm.com/].
- Warren E. Strengthening Research through Data Sharing. N Engl J Med. 2016 Aug 4; 375(5):401-3.
- Welcome To UML Web Site! 2017. [http://www.uml.org/].
- Whyte, A. (2015). 'Where to keep research data: DCC checklist for evaluating data repositories', v.1.1. Edinburgh: Digital Curation Centre. Available online: www.dcc.ac.uk/resources/how-guides
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016; 3:160018.