



# **ESTUDIO PREDIMED PLUS**

## **CIBER OBN**

**(con la colaboración de CIBERESP y CIBERDEM)**

## **Data Management Plan and Data Sharing Policy**

**Versión: 1.6.**

**16-04-2018**

## 1. INTRODUCCIÓN

### 1.1. Conceptualización y marco actual

Un plan de gestión de datos, en inglés Data Management Plan (DMP), es un documento que describe cómo se gestionarán los datos durante el período de investigación activa y después de completar el proyecto. En éste se detalla la información sobre la generación, almacenamiento, acceso, preservación, organización, intercambio y reuso de los datos de la investigación, así como aspectos relativos a protección de datos personales, seguridad, privacidad, implicaciones éticas y propiedad intelectual. Inicialmente los DMPs surgieron con tres objetivos básicos: garantizar que no se perdieran los datos, sobre todo al finalizar un proyecto de investigación; facilitar su correcta custodia desde su producción y permitir su preservación mediante documentación de todo su entorno tecnológico.

Básicamente los elementos a incluir en un DMP son:

1. El tipo de datos que se producirá durante la investigación.
2. Las normas que se utilizarán.
3. Las políticas de acceso y reutilización de datos.
4. Las medidas para proteger la privacidad, la seguridad, la confidencialidad y la propiedad intelectual.
5. El archivo y la preservación de los datos.

Actualmente, para los proyectos de investigación financiados por los organismos públicos de investigación españoles no es requisito obligatorio elaborar un DMP. Tampoco lo es para los proyectos financiados hasta ahora por la Comisión Europea. Sin embargo, se considera que es un valor añadido en las buenas prácticas de gestión de datos y por ello se ha planteado en este momento en el Estudio PREDIMED PLUS. Además en un futuro próximo, la elaboración de un DMP será un requisito para los nuevos proyectos que se soliciten en el marco del Programa Europeo H2020, y en previsión de nuevas solicitudes de distintos proyectos derivados del PREDIMED PLUS que se puedan presentar a estas convocatorias, es necesario establecer un marco general de DMP para el proyecto conjunto PREDIMED PLUS.

Del mismo modo, para optar la financiación internacional fuera de la Unión Europea, otras agencias pueden requerir la presentación de un DMP, y por ello resulta oportuno elaborar unas directrices generales que se puedan adaptar fácilmente a los requerimientos específicos de cada uno de los organismos potencialmente financiadores.

En Estados Unidos, inicialmente en los proyectos de investigación se comenzó solicitando la denominada “data sharing policy” o política de intercambio de datos. Posteriormente se ha ampliado este requisito solicitando no solamente dicha política sino una visión más general de la misma integrada en el DMP. Así, a partir del 18 de enero de 2011, las propuestas de proyectos presentadas a la National Science Foundation (NSF) debían incluir como documento suplementario un DMP. En febrero de 2013, la Oficina de Política Científica y Tecnológica de la Casa Blanca (OSTP) emitió un memorando que requería que los organismos otorgantes solicitaran la elaboración de un plan para los resultados generados en las investigaciones financiadas con fondos federales públicos. A consecuencia de ello, desde el 26 de febrero de 2003, los National Health Institutes (NIH) han tenido una política de “data sharing” para proyectos con presupuestos superiores a \$ 500K/año. En respuesta

al Memorandum OSTP 2013, los NIH aplicarán los requisitos para la elaboración de DMPs a todos los investigadores, y no sólo a aquellos con dotaciones en los proyectos superiores a los \$ 500K. Teniendo en cuenta que en un futuro próximo, desde el PREDIMED PLUS se solicitará financiación internacional a los NIH, es necesario también partir de una base general de DMP para el estudio.

Con estos antecedentes generales se plantea la elaboración del documento que hemos denominado en inglés “Data Management Plan and Data Sharing Policy” (Plan de Gestión de datos y Política de intercambio de los mismos) para el estudio PREDIMED PLUS. De manera abreviada nos referiremos a este documento de ahora en adelante como DMP/DPS.

El DMP es un documento vivo y se puede actualizar en virtud de cambios en los marcos generales de actuación del proyecto y los requisitos externos.

Dentro del marco general del DMP del PREDIMED PLUS se tendrán que ir adaptando sub DMPs específicos para los nuevos proyectos que se soliciten a las distintas agencias financiadoras en un futuro, todo ello teniendo en cuenta las directrices del DMP global.

Existen varias guías para elaborar DMPs cuyo formato varía en función de las agencias financiadoras. A modo de ejemplo, la SPARC (the Scholarly Publishing and Academic Resources Coalition) ha recopilado un excelente recursos sobre los requisitos de DMP/DPS de diferentes agencias financiadoras federales de Estados Unidos (<http://datasharing.sparcopen.org/data>).

Estos DMP surgen potenciados por el denominado “movimiento de Acceso Abierto a la información científica”. Este movimiento tiene una presencia importante en el acceso abierto a publicaciones científicas publicadas en revistas. En este sentido, múltiples agencias de financiación e instituciones que realizan investigación ya disponen de políticas para garantizar el acceso abierto a publicaciones científicas financiadas con fondos públicos.

Este movimiento de acceso abierto y de creación de e-infraestructuras que den apoyo al uso de información científica por la comunidad científica ha incluido también en sus debates la importancia de los datos de investigación, que se reconocen como una fuente de conocimiento propia e independiente de las publicaciones que pueden emplearse en la validación de los resultados de investigación publicados en artículos, para generar nuevo conocimiento y ser explotados de manera interdisciplinar. Se insiste en que para asegurar esta explotación de los datos, es necesario que estén disponibles y accesibles en la red, de la misma manera que lo están las publicaciones. Sin embargo, los requisitos técnicos y legales para garantizar el acceso son más complejos que los de las publicaciones y existen limitaciones éticas en investigación biomédica que impiden generalizar unas directrices para acceso abierto general por lo que hay que particularizar caso por caso.

Del mismo modo varias revistas científicas (PLOS, Nature, etc) han implantado políticas de publicación de datos en acceso abierto, requiriendo los datos crudos que se han utilizado para generar los resultados de los artículos publicados. Estos datos se pueden aportar como archivos descargables directamente como material suplementario del artículo o depositar en distintos repertorios públicos. En el primer caso, cualquier persona puede acceder libremente a los datos y utilizarlos libremente para distintos fines. Dado que la investigación biomédica utiliza muchas veces datos personales que requieren una protección especial, también existe la opción en estas revistas de indicar que los datos estarán disponibles tras

solicitud a los investigadores cumplimentado unos formularios específicos diseñados al respecto por cada estudio de acuerdo con sus particularidades y sus limitaciones éticas.

Paralelamente, el International Committee of Medical Journal Editors (ICMJE) ha lanzado unas directrices para la provisión y el acceso a datos crudos que sustenten los resultados que se publiquen en las principales revistas médicas derivados de ensayos clínicos (Taichman et al, 2016). Esta propuesta tendrá que ser tenida en cuenta para la publicación de futuros resultados del ensayo clínico PREDIMED PLUS. Sin embargo, dada la fecha de inicio del estudio PREDIMED PLUS, anterior a la publicación de estas directrices del ICMJE en 2016, **no estaríamos obligados** a facilitar el acceso abierto a los datos publicados ya que los editores especifican claramente que dicho requisito de acceso abierto entrará en vigor para los ensayos clínicos que comiencen a inscribir a los participantes a partir de un año después de que el ICMJE adopte sus requisitos para compartir datos (el ICMJE planea adoptar los requisitos de intercambio de datos después de considerar la retroalimentación recibida a las propuestas realizadas en la publicación de Taichman et al en el año 2016).

Estas iniciativas recientes todavía están despertando un intenso debate y todavía son susceptibles de mejora para salvaguardar todos los intereses implicados (Warren, 2016). Uno de los puntos más debatidos es el de cómo reconocer la contribución de los investigadores que han generado los datos, ya que el acceso abierto proporciona una gran libertad a los investigadores que reanalizan los datos, sin obligación de reconocer la autoría de los investigadores que han publicado el estudio. Por lo que existen llamadas a reconocer este equilibrio para la buena marcha del acceso abierto a los datos (Kalager et al, 2016). Además de ello, también existen críticas generales a la aproximación de requerimiento de publicación abierta de los resultados de los ensayos clínicos en términos de posible mal uso de los datos y otras debilidades que hay que subsanar para mejorar los objetivos que pretende el acceso abierto sin incurrir en riesgos y mal uso, al tiempo que se propugna una extensión del plazo del tiempo de análisis de los datos tras su generación por los autores del estudio antes de su puesta a disposición para los análisis de otros investigadores (International Consortium of Investigators for Fairness in Trial Data Sharing et al, 2016). En línea con estos comentarios, una reciente publicación sobre el porcentaje de artículos publicados en el BMJ que ofrecían acceso a los datos crudos, ha concluido que este porcentaje es muy bajo. Entre las causas que subyacen destacan la falta de incentivos que tienen los investigadores a quienes se les requiere compartir los datos (Rowhani-Farid et al, 2016). Esta limitación es necesario tenerla en cuenta para mejorar las futuras políticas en ese sentido.

Se ha creado también el consorcio denominado “Academic Research Organization Consortium for Continuing Evaluation of Scientific Studies — Cardiovascular (ACCESS CV) “ (Academic Research Organization Consortium for Continuing Evaluation of Scientific Studies—Cardiovascular et al, 2016) para proporcionar directrices para el intercambio de datos en ensayos clínicos cardiovasculares de manera que se tengan en cuenta las limitaciones éticas inherentes en estos tipos de estudios. Este Consorcio ha elaborado una propuesta preliminar, cuya evolución tendremos en cuenta, para el posterior intercambio de datos derivado de las publicaciones del ensayo clínico final cardiovascular PREDIMED PLUS.

En términos de privacidad, confidencialidad y protección al paciente, recientemente en Estados Unidos se ha actualizado el denominado “common rule” [Federal Policy for the

Protection of Human Subjects ('Common Rule')], publicado inicialmente en 1991, y que se aplicaba a los requerimientos de los consentimientos en los estudios humanos y la utilización de muestras biológicas. La actualización que entrará en vigor en 2018, ha supuesto algunos cambios sobre la situación anterior (Menikoff et al, 2017). Estos cambios los tendremos que tener en cuenta cuando vayamos a solicitar proyectos de investigación a las distintas agencias en Estados Unidos.

En definitiva, nos encontramos ante un marco internacional cambiante en el que todavía es necesario perfilar muchos detalles y atender preocupaciones, necesidades e inseguridades en las distintas etapas del intercambio de datos, así como dotar de más recursos y de la tecnología adecuada para poder adaptarse bien a los requerimientos sin menoscabar la integridad, seguridad y confidencialidad de los datos a la vez que el esfuerzo y propiedad intelectual de los investigadores implicados.

## **1.2. Directrices para la elaboración del DMP en Horizonte 2020**

En España se están tomando como referencia las directrices para elaboración de los distintos DMP las establecidas por la Comisión Europea de manera piloto. La utilización de los DMP constituye una novedad en Horizonte 2020. Por ello, la Comisión está llevando a cabo un estudio piloto flexible en el marco del Horizonte 2020 denominado estudio piloto de datos abiertos de investigación ("Open Research Data Pilot"; piloto de ORD). El objetivo de ORD es mejorar y maximizar el acceso y la reutilización de los datos de investigación generados por los proyectos Horizonte 2020. Sin embargo, este programa piloto tiene especialmente en cuenta la necesidad de equilibrar la apertura y la protección de la información científica, la comercialización y los derechos de propiedad intelectual, así como cuestiones de gestión de datos y preservación. En los programas de trabajo para 2014-16, el piloto del ORD sólo incluyó algunas áreas seleccionadas de Horizonte 2020. Recientemente, en la versión revisada del programa de trabajo de 2017, el piloto ORC se ha ampliado para abarcar todas las áreas temáticas de Horizonte 2020.

Para ayudar a esta gestión, la Comisión ha publicado un Documento de Directrices para la gestión de datos, lanzando la primera versión en 2013 y su actualización en junio de 2016 (European Commission, Directorate General for Research & Innovation. H2020, 2016). Este documento está dirigido a los solicitantes y beneficiarios de proyectos incluidos en este programa piloto ORC. Tal y como ha comunicado la Comisión Europea, en 2017 el proyecto piloto de datos de H2020 dejará de ser un piloto, y todos los proyectos financiados con H2020 (salvo excepciones justificadas) deberán garantizar el acceso abierto a los datos de investigación. Actualmente PREDIMED PLUS no tiene este requisito de datos abierto ni de elaboración de un DMP como obligatorio, sin embargo, en previsión de futuras solicitudes de proyectos ligados a PREDIMED PLUS, es conveniente adaptar el DMP a estas directrices. Dichas directrices tienen por objeto ofrecer indicaciones sobre cómo cumplir con los requisitos exigidos en cuanto a calidad, intercambio y seguridad de los datos de investigación.

Posteriormente, en mayo de 2017 la Comisión Europea (European Commission, Directorate General for Research & Innovation. H2020, 2017), publica una actualización de la Guía de las directrices del documento de DMS titulándolo "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". En dicho documento mantiene las directrices

generales en el mismo modelo que el documento anterior publicado en el año 2016 y añade tres puntos adicionales: Un link a las guías específicas de los proyectos financiados por el European Research Council (v.3.1, 25-8-16), una clarificación de los objetivos en la página 3 del documento y un anexo con los “agreements” específicos que hay que cumplimentar con las revistas a las que se remiten los artículos financiados por la Unión Europea ya que todos tienen que ser de Acceso Abierto a la publicación. En esta nueva actualización se mantiene que no es obligatorio que los datos generados en los proyectos financiados sean puestos en acceso abierto en la publicación. Lo que sí que se clarifica es que tanto los artículos tienen que ser de acceso abierto para los lectores. Distingue entre las revistas que inicialmente sean de acceso abierto (gold) y las revistas que inicialmente por su política no son de acceso abierto, pero con las cuales se tiene que firmar un acuerdo (pagando la cantidad estipulada) para que pasado un período de embargo el artículo sea accesible de manera abierta a los lectores o se pueda poner en un repositorio con acceso abierto (green Access).

Tanto en un caso como en otro (gold o green access), se recomienda, además del lugar de la revista, hacer disponible una copia del artículo en otro repositorio de acceso abierto. Entre los varios repositorios existentes se recomienda depositar una copia del artículo en el repositorio Openaire (<https://www.openaire.eu/>).

Si bien el acceso abierto a los datos de investigación se aplica por defecto en Horizonte 2020, la Comisión también reconoce que existen buenas razones para mantener cerrados algunos o incluso todos los datos de investigación generados en un proyecto. Por tanto, la Comisión ofrece posibilidades sólidas de exclusión del proyecto de la obligación de datos abiertos en cualquier fase, es decir: durante la fase de aplicación; durante la fase de preparación del acuerdo de subvención y después de la firma del convenio de subvención.

Es también importante tener en cuenta que el piloto ORD se aplica principalmente a los datos necesarios para validar los resultados presentados en publicaciones científicas. Otros datos también pueden ser proporcionados por los investigadores del proyecto de forma voluntaria, como se puede especificar en los DMP respectivos.

Las directrices propuestas por la Comisión están basadas en las siglas FAIR, correspondientes a “Findable, Accessible, Interoperable and Re-usable”. En castellano se han traducido estos conceptos como: Descubribles, Accesibles, Interoperables y Reutilizables. El documento elaborado por la Comisión ayuda a facilitar la búsqueda, la accesibilidad, la interoperabilidad y la reutilización de los datos de sus investigaciones a fin de garantizar su correcta gestión, ya que una buena gestión de los datos de investigación no es un objetivo en sí mismo, sino el conducto clave que conduce al descubrimiento e innovación del conocimiento y a la posterior integración y reutilización de datos y conocimientos.

En relación con estas directrices es importante conocer también los proyectos OpenAIRE (<https://www.openaire.eu/>). OpenAIREplus es el proyecto continuación de OpenAIRE dirigido a la publicación e incorporación de los conjuntos de datos (datasets) de investigación a los artículos científicos financiados por el programa Horizonte 2020.

Todos los proyectos con financiación europea sujetos al DMP deben depositar los datos en un repositorio a fin de facilitar su libre reutilización, salvo las siguientes excepciones: Incompatibilidad con la obligación de proteger los resultados si pueden ser explotados comercialmente o industrialmente; Incompatibilidad con temas de confidencialidad o de



seguridad; Incompatibilidad con regulaciones sobre datos personales; Puede poner en peligro el objetivo del proyecto; No se generarán o recogerán datos; Otra razón legítima que debe justificarse en la fase de propuesta del proyecto.

Para cumplir el mandato de datos en abierto los proyectos participantes deben cumplir con dos pasos: depositar los datos en un repositorio y facilitar su libre reutilización mediante una licencia.

La Comisión Europea podrá cubrir costes técnicos y profesionales asociados a la gestión y difusión en acceso abierto.

En España, la Fundación Española para la Ciencia y la Tecnología (FECYT), en colaboración con Red de Bibliotecas Universitarias (REBIUN) de la Conferencia de Rectores de las Universidades Españolas (CRUE), gestiona y coordina RECOLECTA, un proyecto para la creación de una red de repositorios institucionales interoperables y que puede ser considerado como la primera iniciativa nacional en la creación de una infraestructura que facilita la “open science” o ciencia en abierto. El objetivo es además dotar de mayor visibilidad y servicios a los resultados de la investigación y de la producción científica española. El grupo de trabajo del proyecto RECOLECTA, ha realizado un informe sobre la situación actual de los principales aspectos relacionados con el acceso abierto de los datos de investigación en España (Grupo de Trabajo de “Depósito y Gestión de datos en Acceso Abierto” del proyecto RECOLECTA, 2012).

### 1.3. Definición de dato

Existen varias definiciones de datos, pero en general se considera aceptado que los datos de la investigación son hechos, observaciones o experiencias en que se basa la teoría o la experimentación. Los datos pueden ser numéricos, descriptivos o visuales. Los datos pueden ser en estado bruto o analizado, pueden ser experimentales u observacionales. Los datos incluyen: cuadernos de laboratorio, cuadernos de campo, datos de investigación primaria (incluidos los datos en papel o en soporte informático), cuestionarios, audios, videos, desarrollo de modelos, fotografías, películas, y las comprobaciones y las respuestas de la prueba. Las colecciones de datos para la investigación pueden incluir presentaciones en power point; diseños y muestras. En la información sobre la procedencia de los datos también se podría incluir: el cómo, cuándo, donde se recogió y con qué (por ejemplo, instrumentos). El código de software utilizado para generar, comentar o analizar los datos también pueden ser considerados datos. Por lo tanto existe gran diversidad de datos que se generan en un proyecto de investigación con el estudio PREDIMED PLUS, muchos de los cuales no están sujetos a limitaciones éticas y que pueden intercambiarse fácilmente en distintos repertorios públicos.

### 1.4. Herramientas para la creación de un DMP

Existen distintas herramientas online para la creación de un DMP en función de los requerimientos de las distintas agencias financiadoras. A nivel europeo y concretamente en España, las herramientas más utilizadas son las siguientes:

**-DMP Online:** desarrollado por el **Digital Curation Center** (<https://dmponline.dcc.ac.uk>). Sirve como plantilla para elaborar el Plan de Gestión de Datos siguiendo el esquema de la Comisión Europea. Se accede mediante la creación de una cuenta y se van

cumplimentando las distintas preguntas que plantea el formulario. Anteriormente conviene tener resueltas las respuestas, y para ello es muy útil su “Check List for a DMP”

**-PAGODA.** Es la traducción al castellano de la herramienta DMP Online, realizada por Consorcio Madroño. Este consorcio está formado por las Universidades de la Comunidad de Madrid y de la UNED para la Cooperación Bibliotecaria. Para acceder a la herramienta sólo hay que crear una cuenta gratuita.

**-DMP CSUC:** Traducción al catalán de la herramienta DMP Online, realizada por el CSUC (Consorti de Serveis de les Universitats de Catalunya), incluyendo también a la Universitat Jaume I de Castellón (<http://www.csuc.cat/es>). Para utilizar esta herramienta sólo es necesario crear una cuenta. El CUSC ha creado también manuales para facilitar la cumplimentación del DMP Online.

## **2. Desarrollo de los epígrafes del DMP para el estudio PREDIMED PLUS**

### **2.1. Resumen de datos**

#### **2.1.A Descripción del Estudio**

El acrónimo del estudio es PREDIMED PLUS y se titula: “*Efecto de una intervención intensiva sobre el estilo de vida a base de una dieta mediterránea tradicional con restricción de energía, actividad física y tratamiento conductual sobre la prevención de enfermedad cardiovascular*”. Se trata de un ensayo de intervención multicéntrico, aleatorizado y controlado. El objetivo es valorar el efecto en prevención primaria cardiovascular de una intervención intensiva basada en promover una dieta Mediterránea con restricción calórica y metas específicas de pérdida de peso, actividad física, e intervención conductual frente a una intervención de baja intensidad (cuidados habituales) recomendando una dieta Mediterránea sin restricción calórica (grupo control) .Los objetivos principales que se plantean son las diferencias observadas entre ambas intervenciones en cuanto a:

- (i) La incidencia de eventos cardiovasculares (infarto de miocardio no mortal, ictus no mortal o muerte cardiovascular)
- (ii) La pérdida de peso y su mantenimiento a largo plazo.

Existen otros múltiples objetivos secundarios que se detallan en el Protocolo del Estudio.

Actualmente se cuenta con 6874 participantes reclutados y aleatorizados. No se trata de un proyecto único financiado en una única convocatoria, sino que se trata de un estudio multicéntrico iniciado con distinta temporalidad en cada uno de los centros participantes y sujeto a financiación por varios organismos en la actualidad, con la previsión de seguir solicitando financiación en siguientes convocatorias. El estudio PREDIMED-PLUS, coordinado por Jordi Salas-Salvadó, se está llevando a cabo por 23 nodos reclutadores y 8 nodos de soporte científico-técnico. Estos nodos están financiados en su mayor parte por el Instituto de Salud Carlos III (ISCIII) del Ministerio de Economía y Competitividad del Gobierno de España a través del CIBEROBN, y con la colaboración del CIBERESP y CIBERDEM. También se han recibido ayudas FIS para proyectos coordinados 2014 (coordinador Jordi Salas-Salvadó), 2015 (coordinador Pep Vidal), 2017 (coordinador Jordi Salas-Salvadó), 2018 (coordinador Pep Vidal), otras ayudas FIS a IPs que quedaron fuera



de las convocatorias anteriores, así como por un ERC Advanced Research Grant concedido a Miguel A. Martínez-González.

Nodos: Los Investigadores Principales (IP) de los 23 nodos reclutadores son, por orden de numeración: 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 Lluís 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 y 23 Vicente Martín.

Los IPs de nodos de apoyo son: Emilio Ros, Fernando Fernández-Aranda, Cristina Botella, María Puy-Portillo, Rosa M. Lamuela-Raventós, Ascensión Marcos, Guillermo Sáez y Enrique Gómez-Gracia.

Para la gestión del estudio existe un Comité de Dirección o Steering Committee, formado actualmente por 7 IPs (Dolores Corella, Montse Fitó, Ramón Estruch, Miguel Ángel Martínez-González (sustituido por Miguel Ruiz Canela durante su baja), Emilio Ros, Francisco Tinahones y Jordi Salas-Salvadó).

### 2.1.B Relación entre la generación de datos y los objetivos del proyecto

El proyecto comenzó en 2013 en uno de los centros, y paulatinamente se fueron sumando los demás centros reclutadores. Está prevista su finalización en el año 2022. Este seguimiento longitudinal es necesario para que se vayan produciendo eventos cardiovasculares incidentes que son el objetivo principal de cuantificación en el proyecto. Sin embargo, en la actualidad no se dispone de financiación garantizada para llegar al final del proyecto. Sin embargo, como DMP preliminar, se hipotetizará que se dispone de dicha financiación para poder plantear como resultado final una base de datos que contenga los datos finales de eventos previstos tras el seguimiento longitudinal programado. Paralelamente, también está previsto generar bases de datos parciales con resultados basales y de seguimientos a 6 meses y anuales que permitirán evaluar el también objetivo primario de pérdida de peso. Los objetivos secundarios se podrán valorar a distintos niveles con los diferentes seguimientos y tamaños de muestra conseguida.

Es importante señalar que el estudio PREDIMED PLUS consta del denominado **Proyecto General** (desarrollado en el Protocolo del Estudio) y llevado a cabo de forma común por los 23 grupos reclutadores, administrando los mismos cuestionarios y otros soportes de recogida de datos y muestras biológicas, en el que se han centrado inicialmente los esfuerzos y la financiación recibida. Paralelamente, existen una serie de proyectos específicos llevados a cabo de manera establecida con la participación de distintos grupos y nodos, bien de forma individual o a su vez coordinada. El listado de **Proyectos Específicos** y sus coordinadores se encuentra en la Web del estudio PREDIMED PLUS (<http://predimedplus.com/>).

### 2.1.C. Especificar los tipos y formatos de datos generados

Los datos se generarán de manera basal y tras el seguimiento y son de varios tipos y formatos. En el proyecto general se generan datos observacionales creados en tiempo real como puede ser la medida del peso, talla, presión arterial. También se generan datos derivados de cuestionarios y del análisis de imágenes de distintos instrumentos, acelerómetros, electrocardiogramas, etc. Tras extracción de sangre se generan datos en el

laboratorio sobre determinaciones bioquímicas en analíticas básicas o en otras analíticas más específicas. Se obtendrán también datos de los eventos primarios y secundarios a partir de las historias clínicas y registro de mortalidad que serán validados por el Comité de Eventos con la periodicidad establecida.

En proyectos específicos se pueden generar datos experimentales de medidas más complejas sobre biomarcadores, microbiota, datos ómicos, ecocardiogramas, DEXA, etc.

Estos datos se pueden también recoger en cuestionarios en soporte papel, en hojas de cálculo, etc.

Los formatos generados son fundamentalmente texto, numéricos e imagen.

Además de estos datos, se elaborarán artículos originales de investigación que se intentarán publicar en acceso abierto.

De igual modo, se elaborarán revisiones sobre distintos temas relacionados con el PREDIMED PLUS que se intentarán publicar en acceso abierto cuando sea posible.

#### **2.1.D. Especificar si se rehúsan datos existentes**

Algunos datos de la historia clínica del paciente ya recogidos pueden ser utilizados para anotar analíticas y como control de calidad de datos de patologías relacionadas. Posteriormente, tras el seguimiento se utilizarán datos de las historias clínicas para documentar los eventos primarios y secundarios.

#### **2.1.E. Especificar el origen y flujo de los datos**

##### **2.1.E.1. Origen de los datos**

Los datos se generarán fundamentalmente dentro del proyecto. De cada participante y de acuerdo con el protocolo general del proyecto se obtendrá información mediante medición directa del peso, talla, perímetro de la cintura, presión arterial, frecuencia cardiaca, etc. Se obtendrán también datos analíticos de sangre y orina de acuerdo con las variables especificadas en el proyecto general y se administrarán cuestionarios de distintos tipos para obtener información sobre consumo de alimentos, adherencia a dieta mediterránea, actividad física, calidad de vida, datos cognitivos, etc. También se recogen muestras biológicas de suero, plasma, orina, buffy-coat y uñas, para las cuales existe un Protocolo de recogida, almacenamiento y distribución de muestras elaborado específicamente para el estudio. En los proyectos específicos se obtienen también datos adicionales mediante cuestionarios, imágenes, determinaciones de laboratorio, así como muestras adicionales de heces, tipos de células sanguíneas, etc. Los proyectos específicos suelen realizarse en submuestras de participantes del estudio PREDIMED PLUS. Estos proyectos dependen de disponer de financiación adicional externa o del propio nodo para llevar a cabo los análisis, por lo que estos datos, para muchos proyectos específicos, está previsto que se generen un poco más tarde que los datos del proyecto general.

##### **2.1.E.2. Flujo de datos**

Los centros de reclutamiento de participantes producen datos y recogen muestras biológicas de los participantes del ensayo que han reclutado. Los grupos de apoyo pueden generar datos adicionales de muestras analizadas. Fundamentalmente, los nodos reclutadores

recogen los datos de sus participantes mediante los denominados cuadernos de recogida de datos en soporte papel o en soporte electrónico. Estos datos recogidos por los nodos son remitidos diaria o semanalmente al centro coordinador de datos que es el Instituto Municipal de Investigación Médica (IMIM) de Barcelona y a otros centros que reciben datos; bien de cuestionarios en papel con formato de lectura electrónica (de frecuencia de consumo de alimentos, calidad de vida, neurocognitivos, etc) o datos de acelerometría, imágenes. En la Web del PREDIMED PLUS existen protocolos que detallan cómo se tienen que recoger y remitir dichos datos. Fundamentalmente, los datos generales se remiten cumplimentando formularios electrónicos diseñados específicamente para el estudio (se está utilizando el formato TELEFORM), que en tiempo real se transmite en la base de datos centralizada en el IMIM. El acceso a los formularios y su envío se realiza a través del área privada de investigadores PREDIMED PLUS, mediante el requerimiento de contraseña. También se ha informatizado el envío de eventos online a través de la Web de acceso privado del PREDIMED PLUS, de acuerdo con el manual correspondiente.

Otros cuestionarios no se procesan electrónicamente desde el nodo reclutador, sino que se remiten por correo postal para su posterior lectura óptica en el centro de tratamiento de datos correspondiente (Universidad de Navarra y Hospital Bellvitge). Una vez procesados los cuestionarios en papel en los centros que reciben los cuestionarios en papel para su lectura óptica se remiten los datos de las variables generadas a la base central del IMIM para su integración con los demás datos generales del proyecto. Para ello existe un identificador de paciente (ID) que siempre se incluye en cada formulario de recogida de datos. La correspondencia entre el nombre del paciente y su ID queda restringida al nodo reclutador. Teniendo en cuenta la importancia de la privacidad de los datos, el nodo reclutador es el responsable de mantener la confidencialidad de estas asignaciones e implantar los mecanismos necesarios para que la relación entre nombre y apellidos del paciente, DNI y su ID no pueda ser desvelado de manera pública. Se aconseja mantener estos archivos de equivalencia de identificadores en ordenadores o soportes de almacenamiento de datos no conectados a internet.

Los datos son remitidos por los nodos reclutadores a tiempo basal, a los 6 meses y anualmente, de acuerdo con lo especificado en el Protocolo del estudio, a la base de datos central en el IMIM y a los distintos centros adicionales de procesamiento de datos (Universidad de Navarra, Hospital Son Espases-Illes Balears, Hospital Bellvitge, Universidad Rovira i Virgili, Universidad de Málaga) bien de manera electrónica (electros y acelerometría a les Illes Balears y a la Universidad de Málaga, respectivamente) o en papel. El IMIM recibe los datos de los nodos reclutadores y de los demás centros de procesamiento de estos datos del Proyecto general y los integra en una base de datos única del Proyecto general. Esta base es sometida a los respectivos controles de calidad, realizando consultas periódicas a los nodos sobre datos anómalos, datos faltantes, datos ilegibles, etc. Los nodos responden estas preguntas y se va completando el proceso de control de calidad de datos siguiendo los estándares internacionales para ello.

Periódicamente el IMIM remite a cada grupo reclutador un archivo con todas las variables que han generado de sus pacientes para que custodien dicho archivo y realicen también otro control de calidad adicional, cuyo resultado comunicarán rápidamente al IMIM para que haga las correcciones oportunas en las variables implicadas.

Se generarán bases de datos basales, a 6 meses y anualmente de acuerdo con los datos recogidos en el protocolo general que serán sometidas también longitudinalmente a los controles de calidad pertinentes.

Debido las características actuales del sistema de gestión de datos del estudio PREDIMED PLUS, los investigadores no acceden directamente a la base de datos del IMIM. La situación ideal sería disponer de un sistema de **enclave virtual de datos** (Bao et al; 2016), en el que los datos son depositados y gestionados de manera segura, mediante un acceso remoto de los investigadores a la visualización de las variables, al análisis de datos con el software correspondiente, alojado también en el enclave, y a un registro de las sintaxis y estadísticas generadas. De esta manera, se garantiza que las variables no serán alteradas, el control de calidad es muy alto y se protege mejor la seguridad, la confidencialidad y la privacidad de los datos. El DSMB, en su reunión presencial del 19 de enero de 2017 en Barcelona, aconsejó como buena estrategia de mejora para la gestión de datos del estudio PREDIMED PLUS el disponer de un sistema con estas características.

Mientras no podamos disponer de este sistema de enclave de datos, el almacenado de datos, su actualización y el control de calidad de los mismos se realiza de manera controlada en el IMIM por los responsables del sistema para garantizar la integridad y siguiendo los protocolos de seguridad de datos con backups periódicos), privacidad y confidencialidad. De esta base de datos se generan ficheros de datos con distintas características que son los que se remiten a los nodos y se hacen accesibles a los investigadores.

Al ser el PREDIMED PLUS un estudio longitudinal, el almacenado, actualización, extensión y revisión de datos se va realizando continuamente. Por ello, hay que establecer una periodicidad para la generación de los archivos de datos para hacerlos accesibles a los investigadores con las garantías de que se han incorporado todos los casos disponibles para cada variable y se ha realizado el correspondiente control de calidad de los datos tras los queries y feedback de los distintos nodos. Para ello se estipulará una fecha realista en la que los nodos tendrán que haber remitido todos los datos que dispongan del proyecto general para las periodicidades indicadas.

En esta periodicidad o tandas marcadas, se establece la generación de un **archivo de datos basal** (correspondiente a los datos de las visitas S1, S2, S3 y basal) con los **datos de todos los cuestionarios y mediciones que están estipuladas en el Proyecto General**, incluyendo los **participantes aleatorizados de todos los nodos**.

A pesar de que se intente que este archivo de datos basal tenga todos los datos actualizados de los participantes, para algunas variables o nodos, esto no será posible totalmente en las fechas estipuladas, por lo que se cerrará la recepción de datos para no demorar más la creación del archivo de datos. A los archivos de datos correspondientes se les otorgará un número de versión en la fecha en que se generen. Este número de versión y fecha de generación tendrá que ser mencionado en los análisis de datos y publicaciones que se deriven de dichos archivos de datos, indicando también el número de participantes de los que se ha obtenido cada dato. De esta manera, se pueden ir añadiendo datos faltantes a la base de datos con posterioridad al cierre temporal en la generación de la misma para cada momento puntual, en lugar de descartar estos datos. La no inclusión posterior de los datos que no se remitieron en plazo se hacía para tener consistencia en las

variables a lo largo de los diferentes análisis y que no hubiera discrepancias en los resultados. Sin embargo, actualmente se prefiere ir incorporando los datos disponibles en el tiempo hasta el cierre final de la base de datos al finalizar el estudio. Mencionar en cada ocasión la versión y fecha del archivo de datos analizado es una buena práctica y garantiza la calidad global de los análisis.

Tras facilitar el acceso al archivo de datos basal, se seguirán actualizando los datos y su control de calidad, generando archivos de datos al año (incluyendo también las variables de los 6 meses), a los 3 años (incluyendo los datos de los 2 años) y cada año impar, hasta llegar a la base final completa con todos los años. Esta periodicidad se establece para tener los archivos de datos temporales lo más completos y actualizados posible y con el control de calidad realizado. Si fuera necesario realizar análisis con variables de años pares antes de generar el archivo de datos el año impar, se facilitaría el acceso a dicho archivo de datos con el subconjunto de variables específicas en las que se centraría el control de calidad.

Los archivos de datos del **proyecto general**, con la periodicidad indicada, se harán accesibles a los investigadores principales de cada nodo a través de una contraseña personalizada y la cumplimentación de un formulario en el que indiquen que se comprometen a mantener la seguridad, integridad, confidencialidad y privacidad del fichero. Serán también los investigadores principales los responsables de asegurar las buenas prácticas en el uso del mismo por los investigadores de su nodo y de garantizar la no distribución a personal no autorizado de otros grupos de investigación que nos sean de su nodo.

La generación de datos en los **proyectos específicos** dependerá de cuándo dispongan de financiación para ello. Ya que en los proyectos específicos se realiza un **esfuerzo adicional** para generar nuevos resultados, tiene que ser reconocido. Para ello, sin establecer un período formal de embargo, se considera oportuno establecer unas consideraciones generales para valorar el trabajo adicional realizado y dar prioridad al nodo/s generador/es de datos específicos en las publicaciones que se deriven de ellos, fundamentalmente cuando se trate de determinaciones muy costosas en tiempo y recursos. Así, en el caso de determinaciones analíticas muy costosas (metabólica, genómica, proteómica, metagenómica, epigenómica, análisis bioquímicos caros, etc.), con muestras de varios nodos, cuyos gastos se hayan sufragado mediante fondos conseguidos mayoritariamente por el nodo/s que hizo tales determinaciones, aunque dicho nodo tenga que remitir los resultados de las determinaciones a cada nodo que haya aportado muestras en el momento que se finalice y se circule el artículo correspondiente a dichas determinaciones, se respetará el período que transcurra hasta la aceptación del paper del nodo que ha realizado las determinaciones antes de aceptar propuestas de otros nodos para analizar dichos datos.

Si algún IP tiene interés en estos datos antes de la aceptación del paper, deberá contactar con el responsable de las determinaciones realizadas para establecer una colaboración y formular una propuesta conjunta de artículo/os según política de publicaciones. Este contacto se puede establecer a partir del inicio de las determinaciones. En todo caso, tanto antes como después de la publicación del paper inicial, el nodo que ha realizado las determinaciones tendrá prioridad en la propuesta de nuevos papers relacionados con dichas determinaciones y se aconseja la propuesta de papers co-liderados en caso de que otro nodo tenga gran interés en el mismo tema.

Estas consideraciones para los datos generados en proyectos específicos también se aplicarán en los casos de fondos conseguidos con financiación externa para realizar determinaciones con muestras de varios nodos. Si las condiciones de financiación externa obligaran a depositar los datos en algún repositorio al finalizar el proyecto, se priorizará que los investigadores PREDIMED-PLUS tengan acceso avanzado a dichos datos, antes que los investigadores externos que lo soliciten.

En general, todos aquellos datos que provengan de la información o muestras biológicas de la mayoría de los centros participantes en PREDIMED-PLUS (más del 50% de los participantes), o se haya determinado en los casos de eventos primarios, tendrán que ser integrados en la base de datos general custodiada por el IMIM que estará a disposición de todos los IPs de los nodos reclutadores. Periódicamente el Steering Committee definirá aquellas variables que son susceptibles de ser integradas en los archivos de datos para su acceso periódico por los IPs. Para publicar manuscritos con estos datos que no son propios del proyecto general el IP deberá ponerse en contacto con el/los responsables de estos datos generados para formular una propuesta de artículo y remitirla al Steering Committee según la política de publicaciones. Independientemente de esta periodicidad, los IPs podrán ponerse en contacto con los coordinadores de los proyectos específicos y solicitar una colaboración en un trabajo de compartido. Los detalles de este acceso general y específico se detallan en el epígrafe 2.2.2 (haciendo los datos accesibles).

### **2.1.E.3. Consideraciones éticas del origen de los datos**

Además de la recogida de datos de salud del participante mediante los cuestionarios y distintos instrumentos, periódicamente (con la periodicidad que se acuerde por los DSMB/SC) se recogerán datos de eventos de los participantes. Estos datos de eventos se documentarán desde la historia clínica del participante. Por ello, dado que se trata de datos personales de salud sujetos a la ley de protección de datos de carácter personal, es necesario obtener el consentimiento informado del participante, no sólo permitiendo la administración de cuestionarios y determinaciones, sino también aceptando expresamente que se consulte su historia clínica. Se dispone de un modelo consentimiento informado para el estudio general en el que se especifica también este punto que debería ser cumplimentado por todos los participantes de cada nodo. Adicionalmente debe solicitarse la cumplimentación de otro consentimiento informado para el estudio genético. El paciente firma el/los consentimiento(s) informado(s) para entrar en el estudio y puede retirar su consentimiento cuando quiera.

Cada centro es responsable de tramitar la autorización del comité de ética para los participantes que recluta y de guardar los consentimientos informados firmados en condiciones adecuadas de seguridad y confidencialidad.

Dado que en el estudio se administran también varios cuestionarios de salud, se recogen datos de analíticas y de enfermedades y factores de riesgo relacionadas así como variables del estilo de vida, los datos generados no es posible hacerlos accesibles de manera pública a todo tipo de personas. Al paciente se le informa que sus datos se tratarán de manera confidencial y se depositarán en una base de datos que se registra en la Agencia Española de Protección de Datos, estando obligados a seguir la normativa en este sentido. Cada IP de centro reclutador de participantes es el responsable de velar por la protección de datos



en dicha base o archivo de datos que ha registrado a su nombre (ver más detalle en los apartados **2.4** y **2.5**)

Según la Ley de Protección de datos, el participante tiene que estar informado del uso que se va a hacer con los datos, prestando o no su consentimiento. En el modelo general de consentimiento informado del estudio PREDIMED PLUS se informa al participante que parte de los datos que se generen con su información podrán ser compartidos con otros investigadores, indicando que se establecerá una colaboración para ello. En este sentido de cara al intercambio de datos, éste sólo puede realizarse en condiciones controladas de acceso a los datos, previa identificación del investigador y tras establecer un acuerdo de colaboración.

Para el control de calidad está previsto también realizar auditorías periódicas a los centros de reclutamiento de participantes para constatar las buenas prácticas en los aspectos éticos.

Ante cualquier participante del estudio PREDIMED que revoque su consentimiento y expresamente pida que sus datos o muestras biológicas sean destruidos, el Investigador principal del nodo implicado lo comunicará por escrito al IMIM, a los bancos de muestras biológicas, y centros específicos que manejen muestras (acelerometría, DEXA...), para que estos datos sean anulados de las bases de datos y de los bancos específicos.

#### **2.1.F. Especificar el tamaño de los datos generados**

La previsión es generar varios archivos de datos de distintos tamaños de manera temporal y longitudinal en el estudio. Finalmente se dispondrá de una base de datos final con todos los datos basales y seguimientos del proyecto general a la que también se podrán incorporar los datos generados en los proyectos específicos superando el orden de varios GB.

Además de los datos derivados de cuestionarios están los datos generados en los cuestionarios, manuales, imágenes, vídeos, que a su vez suponen cientos de GB. En proyectos específicos los datos generados pueden ocupar incluso mayor tamaño que la base del Proyecto general (algunos datos ómicos masivos), siendo necesario tener en cuenta esta característica y optar por disponer de una estructura de base única federada en distintas localizaciones para una mejor gestión de los datos.

El volumen total se estima en varios TeraB.

#### **2.1.F. Especificar la utilidad de los datos generados**

Los datos generados en la elaboración de cuestionarios, protocolo del estudio, manuales de procedimientos, material de intervención dietética, material de intervención de actividad física, material de intervención conductual, recogida de eventos, presentaciones en congresos, etc. pueden ser de interés para otros grupos de investigación, docencia e incluso para la población general.

Los datos generados en la base de datos pueden ser de gran interés para otros investigadores para colaborar en estudios comparativos, replicación de resultados, meta-análisis y otros tipos de estudios adicionales.

Los datos generados mediante los análisis estadísticos serán útiles en la generación de nuevo conocimiento en cada uno de los objetivos del estudio. Se publicarán en forma de

artículos originales y se intentará hacerlos accesibles de manera abierta en las publicaciones.

## **2.2. Datos FAIR (Descubribles, Accesibles, Interoperables y Reutilizables)**

### **2.2.1 Haciendo los datos Descubribles**

#### **2.2.1. A. Haciendo los datos descubribles. Provisión de metadatos**

Se utilizarán los metadatos estándar en la disciplina para describir las bases de datos generadas. Para ello se utilizarán los estándares de metadatos descritos en el Dublin Core Schema (<http://www.dcc.ac.uk/resources/metadata-standards/protocol-data-element-definitions>), que habitualmente se utilizan de referencia. Dentro de estos metadatos allí descritos se encuentran entre otros: Specification: <http://prsinfo.clinicaltrials.gov/definitions.html>. Standard's website: <http://clinicaltrials.gov/ct2/manage-recs/resources>.

Básicamente, crearemos metadatos para describir cada base de datos que generemos con los siguientes metadatos mínimos:

- Título: Nombre del proyecto del conjunto de datos o de investigación que lo produjo
- Nombres de los creadores y las direcciones de la organización o personas que han creado los datos.
- Código de identificación de los datos, incluso si es una referencia de uso interno.  
Palabras o frases que describen el tema o el contenido de los datos.
- Patrocinadores: Las organizaciones o agencias que financiaron la investigación.
- Derechos: Cualquier tipo de derechos de propiedad intelectual de los datos.
- Acceso a la información: ¿Dónde y cómo sus datos pueden ser accesibles por otros investigadores?
- Idioma del contenido.
- Fechas clave asociadas a los datos, incluyendo: inicio del proyecto y la fecha de finalización.
- Lanzamiento, período de tiempo cubierto por los datos, y otras fechas relacionadas con la vida útil de datos, por ejemplo, el ciclo de mantenimiento, actualización del programa.
- Lugar al que hacen referencia los datos (p.e, una ubicación física, cobertura espacial etc).
- Metodología: ¿Cómo se generaron los datos?
- Procesamiento de datos: toda la información acerca de cómo los datos se han procesado.
- Lista de nombres de archivo de la lista de todos los archivos de datos asociados con el proyecto, con sus nombres y extensiones de archivo.
- Formatos de archivo de los datos, necesario para leer los datos.
- Organización de archivos: estructura del archivo de datos (s) y la disposición de las variables.
- Lista de variables en los archivos de datos.
- Explicación de los códigos o abreviaturas utilizadas
- Versiones de fecha / fecha y hora para cada archivo, y usar un ID diferente para cada versión.

-Operaciones de comprobación para verificar si los archivos han cambiado a lo largo del tiempo (Algoritmos CheckSum Hash, MD5, SHA-1, etc para proteger la integridad de los datos).

### **2.2.1. B. Indicar la identificabilidad de los datos y referir mecanismos estándar de identificación ¿Se usarán Digital Object Identifiers: DOI?**

Los datos correspondientes a cuestionarios, manuales de uso, recetarios, protocolos de intervención dietética, de actividad física, etc, se depositarán en repertorios públicos como por ejemplo los que gestionan nuestras universidades. Entre ellos el repositorio RODERIC (<http://roderic.uv.es/>) en la universidad de Valencia y otros similares en los que se pueden depositar. Concretamente, RODERIC responde al compromiso de la Universidad de Valencia con su adhesión a la Declaración de Berlín (30 Septiembre de 2008). Utiliza protocolos internacionales normalizados que garantizan la visibilidad de los documentos en los buscadores de Internet, como Google, entre otros.

Estos repositorios institucionales proporcionan un único URL para acceder a los documentos correspondientes con el formato <https://repository/record/1234>.

Los datos depositados en algunos repositorios como Zenodo, Dryad, Pangaea o FigShare reciben un DOI, y en ese caso utilizaríamos el DOI como identificador unívoco para los datos allí depositados.

Lo ideal es disponer de un DOI para los documentos, metados y datos agregados depositados, y con el DOI Citation Formatter tenemos una interfaz sencilla para extraer los metadatos de forma automática a partir de un DOI y construir una cita completa. Es compatible con más de 500 diferentes estilos de citas en 45 idiomas diferentes.

### **2.2.1. C. Indicar cómo se denominan y se estructuran las carpetas y los archivos**

Las carpetas y los archivos del proyecto general tendrán un identificador inicial de denominación del estudio "PREDPLUS\_nombrearchivo\_fecha\_version\_XX" seguido del nombre específico del archivo, la fecha de creación, la versión y especificación (XX). Los dos últimos dígitos (XX) se utilizarán como tales si el archivo creado corresponde al total del proyecto general en todos los nodos reclutadores PREDIMED PLUS. Cuando se trate de archivos específicos con datos solamente de un nodo, los dos últimos dígitos corresponderán al nodo, según la numeración estandarizada de nodo establecida en PREDIMED PLUS (nodo 01, nodo 02, nodo 03, etc). De igual modo, se crean carpetas específicas temáticas para integrar los archivos correspondientes al mismo tema. Las carpetas y los archivos de los proyectos específicos tendrán un identificador inicial de denominación del estudio en el que se indique que es específico "ESPREDPLUS\_nombrearchivo\_fecha\_version", así como el nombre de archivo, fecha y versión.

### **2.2.1.D. Indicar cómo se identificarán las diferentes versiones de un mismo conjunto de datos**

El control de las distintas versiones de un conjunto de datos se realizará mediante la utilización de un postfijo compuesto por la fecha en formato aaaammdd seguido de los número de versión y subversión (para pequeñas modificaciones): ejempl V\_1\_S\_0; V\_1\_S\_1, etc.

### **2.2.1.E. Indicar cómo se capturan/crean los metadatos**

Se utilizará una forma mixta para capturar/crear los metadatos dependiendo del tipo de metadatos generado y de la existencia o no de estándares.

Cuando existan estándares se utilizarán los mismos siguiendo el Dublin Core o el Registro de metadatos ISO/IEC 11179, cuando sea pertinente.

Para otros casos, se utilizará un proceso manual de manera que cada archivo específico ira acompañado por sus propios metadatos para facilitar y clarificar su uso.

## **2.2.2 Haciendo los datos Accesibles**

### **2.2.2.A. Especificar de qué manera los datos generados estarán disponibles de manera interna para los investigadores del estudio multicéntrico y para otros investigadores externos y para el público en general.**

El estudio PREDIMED PLUS es un estudio multicéntrico, en el que cada nodo reclutador genera todos los datos del proyecto general correspondientes a los pacientes reclutados por el mismo y los remite electrónicamente o en papel a los centros procesadores de datos, que finalmente los integran en una base de datos general alojada en el IMIM. Se trata de un estudio longitudinal de larga duración, por lo que la base de datos final del proyecto estaría disponible a largo plazo (más de 10 años desde el inicio del proyecto), tras finalizar la recogida y validación de eventos primarios y secundarios y del control de calidad de los datos. Por ello, aunque en la mayoría de los estudios, las políticas de acceso a los datos se refieren a la base de datos final del proyecto, en el estudio PREDIMED PLUS contemplaremos también el acceso a bases de datos parciales, que se van generando a tandas de manera periódica en el tiempo. La periodicidad detallada de la generación de los archivos de datos se ha especificado en el epígrafe 2.1E2 (flujo de datos).

Paralelamente, existen proyectos específicos que irán generando datos de manera progresiva a medida que dispongan de financiación y cuyos resultados se irán integrando en la base central de manera paulatina (análisis genómicos; mediciones DEXA; determinaciones de microbiota; determinaciones de biomarcadores de distintos tipos en muestras biológicas de plasma, suero y orina; densitometría ósea, ecocardiografías, otras pruebas de imagen, etc. Estos datos son generados por nodos concretos, algunos de ellos de manera coordinada contribuyendo al proyecto específico.

Los datos pueden ser compartidos de manera interna o externa y en ambos casos se tendrán que garantizar las buenas prácticas en el uso de los datos, así como tener en cuenta que se trata de datos personales sujetos a privacidad y confidencialidad y que han sido cedidos para unos fines expresados en el consentimiento informado. Seguidamente se detallan los procedimientos de intercambio de datos en función del carácter interno, investigadores externos o público en general.

#### **2.2.2. A1. Procedimiento de acceso/intercambio interno de datos**

Los **investigadores principales de los nodos reclutadores** PREDIMED PLUS tendrán acceso a los ficheros de datos totales de todos los pacientes aleatorizados que se generen

en el estudio general PREDIMED PLUS, tanto al archivo final del estudio tras completar el mismo, como a los distintos archivos que se generen de manera periódica a medida que se vayan completando los cuestionarios y realizando el control de calidad de los datos: fichero de datos basales, al año, tres años, etc. (ver epígrafe 2.1E2: flujo de datos donde se especifica dicha periodicidad).

Los IPs de los nodos reclutadores también tendrán acceso a los datos del proyecto general correspondientes a una periodicidad distinta a la estandarizada, siempre que lo soliciten expresamente al Steering Committee para que se valore la viabilidad de dicha solicitud y se pueda hacer el control de calidad de las variables correspondientes para facilitar el archivo. Del mismo modo, los IPs de los nodos reclutadores también tendrán acceso a las variables generadas en los proyectos específicos (genéticas, biomarcadores, otro tipo de determinaciones bioquímicas, DEXA, microbiota, ecocardiografía, otras pruebas de imagen, etc), en el momento en que estas variables estén disponibles tal como se indica en el punto **2.1.E.2.**) En todo caso, para la solicitud de papers con dichos datos se tendrá que respetar el tiempo hasta la aceptación de los primeros papers propuestos por los nodos que han generado las determinaciones específicas costosas, o bien formular una solicitud en colaboración. Estos nodos responsables de las determinaciones específicas costosas tendrán prioridad en la propuesta nuevos papers derivados de la generación de los datos específicos. Reconocer este período es importante porque se está constatando que uno de los principales problemas en el rechazo al intercambio de datos entre los investigadores es el no reconocimiento en la generación de datos (Longo et al, 2016; Kalager et al, 2016). Por ello en el estudio PREDIMED PLUS es necesario cuidar este reconocimiento. Cuando los datos estén disponibles para los demás investigadores, se recomienda una colaboración activa con los grupos que han generado estos datos, siendo la situación ideal la de compartir los lugares principales de la autoría en los artículos (Longo et al, 2016). Del mismo modo, si desde un grupo investigador existiera un interés especial en realizar una investigación con los datos generados en un proyecto específico particular, se puede llegar a un acuerdo de colaboración entre los grupos en las primeras etapas del comienzo de las determinaciones analíticas y en general se recomienda realizar una propuesta conjunta de paper al Steering Committee.

Toda determinación nueva de una variable medida en los participantes de PREDIMED-PLUS que no sean de su nodo debe comunicarse a todos los investigadores de nodos reclutadores. El IP del nodo que efectúe tal determinación debe comunicarlo en un plazo nunca superior a dos meses.

El acceso a cada fichero de datos será realizado a través de un acceso Web del IMIM habilitado para ello a través de un usuario y contraseña personalizada. Los IPs de los nodos reclutadores accederán mediante descarga a dicho archivo de datos tras cumplimentar una hoja de solicitud indicando el archivo requerido (archivo del proyecto general basal, al año, tres años u otros archivos de distinta periodicidad o datos específicos) y el compromiso cumplir las buenas prácticas de manejo y gestión de datos. El IP de cada nodo será el responsable de velar por la seguridad, integridad, confidencialidad y privacidad de los datos que recibe, ya que además de los participantes de su nodo, recibe datos de los participantes de los demás nodos. En el Anexo 1 se presenta el formulario de solicitud de datos y compromiso para los IPs de nodo reclutador.

De manera similar, que los IPs de otros nodos no reclutadores PREDIMED PLUS podrán acceder a los datos que necesiten para su investigación formulando una petición concreta de datos al Steering Committee, que valorará la viabilidad de la misma. Para ello tienen que cumplimentar el formulario del Anexo 2. Básicamente, estos documentos indican el compromiso de respetar los siguientes puntos:

- Mantener la confidencialidad y la privacidad de los datos.
- Velar por la calidad de los datos y seguir buenas prácticas de análisis y procesado.
- No distribuir los datos a terceras personas que no dispongan de autorización para su uso.
- Reconocer el esfuerzo de los investigadores que han generado los datos a través de su autoría en las publicaciones que se deriven del acceso a los datos.
- Realizar una propuesta de publicación de paper de acuerdo con el modelo estandarizado y recibir la autorización pertinente del Steering Committee.
- Seguir la política de publicaciones PREDIMED PLUS.

**2.2.2. A2. Procedimiento de acceso/intercambio de datos con investigadores externos que solicitan datos** No se depositarán datos crudos del PREDIMED PLUS en repositorios de acceso totalmente abierto. PREDIMED PLUS se inició con antelación a las directrices del ICMJE (2016) y de la Comisión Europea sobre acceso universal a los datos, por eso los IPs no están obligados a facilitar el acceso abierto para toda la comunidad científica de los datos usados en publicaciones. Además, existe una limitación ética en facilitar el acceso libre a los datos, ya que en el consentimiento informado que firma el paciente no se autoriza tal acceso.

Si un grupo externo desea realizar una investigación utilizando datos de PREDIMED-PLUS, se podrá llegar a un acuerdo, pero se requerirá realizar antes una propuesta conjunta de colaboración (contando con al menos un IP PREDIMED-PLUS) y detallada. Sólo tras minucioso estudio por parte de Steering Committee del ensayo PREDIMED-PLUS, y de revisión de las implicaciones éticas, se podrá resolver si se puede firmar o no el acuerdo de colaboración, incluyendo la posibilidad de abonar una cuota.

En la colaboración con **investigadores externos que sólo solicitan datos**, se considerarán fundamentalmente dos modalidades.

**2.2.2.A:2.1: Colaboración en la que los investigadores externos sólo solicitan datos agregados de las variables de interés sin requerir acceso a las variables crudas.**

Este procedimiento es frecuentemente utilizado por Consorcios de investigadores para llevar a cabo meta-análisis siguiendo unos protocolos uniformes y un plan de análisis detallado y especificado. De esta manera sólo se solicita la colaboración para facilitar datos agregados teniendo en cuenta que serán los investigadores del estudio PREDIMED PLUS los que realizarán los análisis estadísticos con los datos crudos y proporcionarán a los investigadores externos los estadísticos descriptivos y las medidas de asociación requeridas para cada caso. Con este procedimiento no se plantean problemas adicionales de posible



mal uso de las variables crudas por los investigadores externos, ya que los análisis estadísticos y el control de calidad se realizarán de manera interna.

Para formalizar dicha colaboración se requerirá una carta de solicitud del investigador/res externos al SC del estudio PREDIMED PLUS y la cumplimentación de un formulario de solicitud de datos (**ver Anexo 3**). En dicho formulario se detallarán los datos del investigador/res solicitantes. El motivo por el cual solicitan los datos al estudio PREDIMED PLUS, detalles sobre los demás grupos a incluir en el meta-análisis, los objetivos del estudio propuesto, el tamaño de muestra que requieren, la hipótesis de partida, la metodología detallada, las variables de partida, los análisis estadísticos propuestos especificando detalladamente los modelos y las variables de ajuste y los datos agregados a aportar tras los análisis estadísticos. Se solicitará también información sobre el número de artículos que se pretenden publicar con la colaboración, las posibles revistas para su remisión, así como el número de autores PREDIMED PLUS que está previsto incluir en los posibles artículos que se deriven. En el **Anexo 4** se presenta la Guía de solicitud para acceso a datos.

El Steering Committee revisará la carta de solicitud con los criterios de excelencia en la investigación y se aprobará o no la colaboración en función de la información facilitada en el formulario de solicitud, teniendo también en cuenta que dicha propuesta no se solape ni interfiera con objetivos y propuestas de investigadores internos del estudio PREDIMED PLUS. En caso de ser aprobada dicha colaboración con investigadores externos se decidirá el grupo/s más indicado para formalizar la realización de los análisis estadísticos propuestos en base a la temática de la colaboración y a la implicación más directa con las variables solicitadas. La autoría PREDIMED PLUS se establecerá según la política de publicaciones, procurando que figuren como autores el máximo número posible de investigadores PREDIMED PLUS. Si se trata de una solicitud de variables del proyecto general con todos los participantes o la mayoría de ellos, se podrá establecer un turno rotatorio de investigadores de cada nodo en las distintas publicaciones generadas. Si se trata de datos de uno o pocos nodos, se incluirán los autores de dichos nodos, preferentemente. Si se trata de colaboración con variables generadas en un proyecto específico, también se considerarán con preferencia los autores de nodos que hayan generado dichas variables.

Las propuestas y borradores de los artículos que se generen para publicación con la integración de dichos datos, serán remitidos para su revisión al SC que gestionará dicha información de acuerdo con los procedimientos del estudio.

Teniendo en cuenta el volumen de trabajo que se requiera realizar en los análisis propuestos, se podrá valorar el establecimiento de una cuota a pagar por los investigadores que solicitan los datos agregados. El pago de dicha cuota se realizaría en la cuenta bancaria gestionada desde el CIBERobn creada para tal fin y con traslación directa a las necesidades del estudio PREDIMED PLUS (fundamentalmente destinada a contribuir al pago de gastos de publicación de los distintos artículos PREDIMED PLUS).

#### **2.2.2.A:2.2: Colaboración en la que los investigadores externos solicitan datos crudos individuales de los participantes en el estudio PREDIMED PLUS**

En el caso de solicitud de datos desglosados a nivel individual de los participantes en el estudio PREDIMED, dicha colaboración plantea más repercusiones relacionadas con la confidencialidad y buen uso de los datos, por lo que se requerirá adicionalmente a los investigadores que garanticen este cumplimiento. Actualmente, el sistema de gestión para el

manejo de datos que posee el estudio PREDIMED PLUS no permite un acceso seguro a los datos por investigadores externos tipo **enclave virtual** de datos (Vie et al, 2013). Por ello, al tratarse de un sistema no óptimo para el intercambio de datos (ya que hay que facilitar a los investigadores externos el fichero de datos con las variables requeridas, que pueden copiar, alterar y distribuir sin control), tenemos como objetivo de futuro intentar mejorar el sistema de gestión de datos. Como se ha indicado en el punto anterior, el sistema óptimo que tenemos como referencia para la gestión de datos es el implementado en la *Harvard University* para el “The Nurses’ Health Study” (Bao et al, 2016). Dicho sistema como tal, tiene un coste elevado de implementación y mantenimiento, pero se pueden plantear alternativas similares a menor escala. La implantación de dicho sistema de gestión de datos permitiría facilitar las colaboraciones al poder facilitar el acceso controlado a las variables requeridas por los investigadores sin facilitarles un fichero de datos, sino directamente realizando los análisis estadísticos bajo un login personalizado a través de las sintaxis pertinentes en un ordenador central dotado con el software específico de análisis. Mientras no se implante dicho sistema, el acceso a los datos tiene limitaciones de seguridad que intentarán ser minimizadas mediante distintas estrategias.

El documento de solicitud para datos será el Anexo 3 y la Guía de solicitud será la especificada en el Anexo 4. **El procedimiento general para la solicitud de datos crudos por investigadores externos** será el siguiente:

*Debido a la alta demanda interna de análisis de datos por los numerosos grupos que forman parte del estudio PREDIMED PLUS, las propuestas de los investigadores externos tienen que estar claramente justificadas y no solaparse con propuestas similares realizadas por investigadores internos ya en curso o que figuren como objetivos en los distintos proyectos de investigación.*

*Las propuestas de colaboración pueden formularse tanto para datos generales del estudio PREDIMED PLUS como para proyectos específicos. En el caso de proyectos específicos se procurará que el coordinador del proyecto específico correspondiente esté en contacto más directo con el investigador externo que formula la solicitud de colaboración.*

### **I.-Presentación de una propuesta de colaboración**

*Para presentar una propuesta de colaboración y de acceso a datos hay que seguir los siguientes pasos:*

#### **I-1 Elaboración de una carta de solicitud de colaboración:**

*En esta carta dirigida al Steering Committee del estudio PREDIMED PLUS, el investigador indicará las generalidades de la solicitud de colaboración, aportará un breve curriculum sobre sus actividades de investigación y describirá los beneficios de la colaboración (ver más detalle en el impreso ANEXO 4, formulario de solicitud de colaboración de investigadores externos).*

**I.2. Cumplimentar el formulario de solicitud de colaboración de investigadores externos** al estudio PREDIMED PLUS para acceso a datos generados en el estudio PREDIMED PLUS.

*En este formulario se solicitan más detalles sobre la hipótesis del estudio, objetivos, metodología a utilizar, variables solicitadas, tamaño de muestra, duración del estudio, etc.*

En el caso de **meta-análisis agregados** en los que el investigador no solicita acceso a las variables crudas sino sólo a los coeficientes y valores de P correspondientes a modelos de asociación concretos. En este apartado, debe indicar las variables que se van a incluir en el análisis así como los análisis concretos a realizar, especificando las variables de control, y las sintaxis correspondientes a los modelos.

### **I.3-Remisión de la propuesta y evaluación**

La propuesta se remitirá electrónicamente al Steering Committee del estudio la primera semana de cada mes. Se evaluarán las propuestas que podrán ser aceptadas, pendientes de subsanación o denegadas.

Los criterios para la evaluación de la propuesta serán fundamentalmente los siguientes:

- La excelencia científica de la propuesta;
- La prioridad estratégica de la propuesta, con respecto a los estudios en curso en el PREDIMED PLUS, evitando la duplicación innecesaria de trabajo;
- El historial científico de investigación de la persona o grupo que presenta la solicitud,
- Calidad científica y capacidad para administrar el proyecto;
- Cumplimiento del estudio propuesto con lo especificado en el consentimiento informado de los pacientes de los que se requieren datos.

### **II- Condiciones de acceso los datos solicitados del estudio PREDIMED PLUS**

Si la propuesta es aceptada, entonces se firmará en primer lugar un **Acuerdo de colaboración** (Anexo 9) que detalla las condiciones bajo las cuales se permite el acceso a los datos solicitados, siendo necesario la aceptación de las mismas por parte del investigador colaborador. Según sea el tipo de estudio planteado podrá ser necesaria una nueva autorización del Comité de Ética del Estudio PREDIMED PLUS (centralizado o en distintos nodos participantes si se requieren datos de los mismos).

Dado que en la actualidad el estudio PREDIMED PLUS no dispone de un sistema de gestión de datos basado en enclave de datos, a través del cual el investigador externo podría acceder a las variables de modo seguro, sin capacidad de modificación o copia, se especificarán unos requisitos adicionales para velar por la seguridad, integridad y confidencialidad de los datos.

Al investigador externo se le proporcionará una contraseña para acceder a un fichero de datos conteniendo las variables solicitadas. Este fichero de datos estará anonimizado sin utilizar ningún identificador de variable para proteger mejor la identidad del participante. Se facilitarán también los metadatos correspondientes. El investigador externo se compromete a:

- No tratar de identificar los pacientes individuales y mantener la confidencialidad
- Mantener la integridad de los datos
- Disponer de entornos seguros de almacenamiento de datos
- No distribuir los datos a terceras personas no autorizadas
- Utilizar los datos facilitados solamente para los objetivos concretos de la colaboración

*-Destruir los datos facilitados tras el período activo de establecimiento del Acuerdo de colaboración*

*En el caso de **propuestas de colaboración para realizar meta-análisis agregados**, como el investigador externo no solicitará en este caso las variables crudas, sino solamente solicita que se realicen los análisis estadísticos y se faciliten los coeficientes, la situación de seguridad e integridad de las variables se mantiene en el propio estudio PREDIMED PLUS.*

### **III. Análisis de datos y publicaciones**

*El investigador externo, al igual que los investigadores internos tiene que garantizar la calidad y las buenas prácticas en la realización de los análisis estadísticos. Detallará el software utilizado, guardará y aportará la sintaxis correspondiente a cada artículo, así como los datos utilizados*

*En el caso de propuestas de **meta-análisis**, el análisis estadístico será realizado por un investigador interno del estudio PREDIMED PLUS, designado en función de la experiencia en la colaboración requerida y facilitará los resultados agregados requeridos en forma de Tablas.*

*Cumplimentará el formulario de propuesta de publicación (o publicaciones) e indicará los autores de la misma. En el caso de estudios originales, se seguirá la política de publicaciones para fijar las autorías, teniendo en cuenta si se trata de una colaboración con los datos del estudio general en la muestra total, o con proyectos específicos.*

*En caso de estudios colaborativos o meta-análisis en los que participaran muchos autores de otras cohortes y estuviera limitado el número de autores, se seguirá como norma general incluir el máximo número de autores del estudio PREDIMED PLUS, estableciendo también un rotatorio acumulable para distintos meta-análisis.*

*La propuesta de publicación tiene que ser aprobada por el Steering Committee y los resultados en forma de Tablas y el artículo completo tiene que ser presentado antes de remitir para publicación para realizar los oportunos controles de calidad.*

### **IV. Coste**

*El trámite de revisión de la solicitud de acceso a datos, la preparación de los mismos, la facilitación del acceso a datos y metadatos, análisis estadísticos en su caso, la revisión de la propuesta de publicación y resultados tiene un coste en tiempo para los investigadores del estudio PREDIMED PLUS. Por lo tanto, en consonancia con el acceso a datos de otras grandes cohortes internacionales, se establece una tasa que se hará efectiva tras la aprobación de la solicitud. El ingreso se realizará en la cuenta PREDIMED PLUS (CIBER OBN) habilitada para ello a través de transferencia bancaria. Las tasas varían según el tipo de colaboración contemplando desde la situación exenta de coste hasta 5000 euros (ver documento de Tasas para datos y muestras según tipo de colaboración solicitada y volumen de datos requeridos; Anexo 10)*

#### **2.2.2. A3. Procedimiento de acceso/intercambio de datos con investigadores externos que solicitan muestras**

Para los investigadores externos que sólo solicitan muestra, se cumplimentará el formulario de solicitud del Anexo 5 y se seguirá la Guía correspondiente en el Anexo 6. También se tendrá que cumplimentar un acuerdo de colaboración específico para solicitud de muestras que se detalla en el Anexo 9.

Básicamente, el proceso sería el siguiente:

*En este caso, el investigador externo sólo requiere muestras para realizar determinados tipos análisis sin coste adicional para el estudio PREDIMED PLUS, existiendo el compromiso de algunos investigadores del estudio PREDIMED PLUS que posteriormente realizarán el link entre los resultados de las muestras analizadas y los datos correspondientes del estudio PREDIMED PLUS. El investigador interno tendrá que solicitar a su vez autorización para formalizar la colaboración a los centros de cuyos pacientes se van a analizar datos y muestras.*

*Se valorarán especialmente las colaboraciones con objetivos relevantes y novedosos y que puedan implicar aportación de otros datos por parte del investigador colaborador con fines comparativos o para incrementar el tamaño de muestra. Las propuestas para evaluar hipótesis altamente especulativas o con limitaciones éticas no se consideran apropiadas.*

*Debido a la alta demanda interna de análisis de muestras por los numerosos grupos que forman parte del estudio PREDIMED PLUS, y dado el carácter finito de las mismas, las propuestas de los investigadores externos tienen que estar claramente justificadas y no solaparse con propuestas similares realizadas por investigadores internos ya en curso o que figuren como objetivos en los distintos proyectos de investigación.*

### ***I.-Presentación de una propuesta de colaboración***

*Para presentar una propuesta de colaboración y de acceso a muestras hay que seguir los siguientes pasos:*

#### ***I-1.Elaboración de una carta de solicitud de colaboración:***

*En esta carta dirigida al Steering Committee del estudio PREDIMED PLUS, el investigador externo indicará las generalidades de la solicitud de colaboración, aportará un breve curriculum sobre sus actividades de investigación y describirá los beneficios de la colaboración (ver más detalle en el impreso ANEXO 5, formulario de solicitud de colaboración de investigadores externos para acceder a muestras).*

#### ***I.2. Cumplimentar el formulario de solicitud de colaboración de investigadores externos*** al estudio PREDIMED PLUS para acceso a muestras generadas en el estudio PREDIMED PLUS.

*En este formulario se solicitan más detalles sobre la hipótesis del estudio, objetivos, metodología a utilizar, investigador PREDIMED PLUS que realizará el estudio de datos, variables que se van a analizar, tamaño de muestra, duración del estudio, tipo de muestras que se solicita, cantidad, concentración, condiciones de transporte, lugar de transporte, determinaciones a realizar, etc.*

#### ***I.3. Remisión de la propuesta y evaluación***



*La propuesta se remitirá electrónicamente al Steering Committee del estudio la primera semana de cada mes. Se evaluarán las propuestas, y éstas podrán ser aceptadas, pendientes de subsanación o denegadas.*

*Los criterios para la evaluación de la propuesta serán fundamentalmente los siguientes:*

- *-Disponibilidad de muestras*
- *-La excelencia científica de la propuesta;*
- *-La prioridad estratégica de la propuesta, con respecto a los estudios en curso en el PREDIMED PLUS, evitando la duplicación innecesaria de trabajo;*
- *El historial científico de investigación de la persona o grupo que presenta la solicitud,*
- *-Calidad del grupo que va a realizar las determinaciones con las muestras.*
- *-Oportunidad del investigador o investigadores PREDIMED PLUS que avalan la propuesta de intercambio de muestras y su compromiso para realizar el análisis de datos y/o colaboración con otros investigadores internos PREDIMED.*
- *Tipo y disponibilidad de los datos que se pretender combinar con los resultados del análisis de muestras.*
- *-Cumplimiento del estudio propuesto con lo especificado en el consentimiento informado de los pacientes de los que se requieren datos.*

## **II. Condiciones de acceso los datos solicitados del estudio PREDIMED PLUS**

*Si la propuesta es aceptada, entonces se firmará en primer lugar un **Acuerdo de colaboración para acceso a muestras** (Anexo 9) que detalla las condiciones bajo las cuales se permite la cesión de muestras al investigador externo, siendo necesario la aceptación de las mismas por parte del investigador colaborador. También será necesario un acuerdo con los investigadores internos PREDIMED PLUS y la solicitud de permiso de éstos para el uso de datos que requieren los análisis de las muestras.*

*Según cuál sea el tipo de estudio planteado podrá ser necesaria una nueva autorización del Comité de Ética del Estudio PREDIMED PLUS (centralizado o en distintos nodos participantes si se requieren datos de los mismos).*

*Dado que en este caso son los investigadores internos PREDIMED PLUS los que realizarán el análisis de datos, no es necesario facilitar el fichero de variables al investigador externo, por lo que los principios de seguridad, confidencialidad, privacidad, integridad de datos y no distribución de los mismos a terceros, los tienen que garantizar los investigadores internos PREDIMED cumplimentando el formulario de compromiso de acceso a los datos.*

*Los datos requeridos pueden ser del Proyecto General o de Subproyectos y requerir que los investigadores internos tengan que realizar la solicitud de dichos datos. Será necesario para ello contar con la colaboración de los investigadores que han generado dichos datos (generales y/o específicos). El acceso a los datos se proporcionará facilitando la transferencia del archivo de datos correspondiente, tras la realización de estos acuerdos.*

*El investigador externo que accede a muestras (que le serán remitidas a la dirección especificada) se compromete a mantener:*

*-Seguridad e integridad de las muestras, garantizando el transporte en las condiciones adecuadas de congelación/refrigeración y almacenamiento.*



*-Exclusividad de las muestras para las determinaciones aprobadas en el estudio PREDIMED PLUS, sin poder realizar determinaciones adicionales ni ceder las muestras a otros investigadores.*

*-Retornar en condiciones óptimas de conservación las muestras sobrantes tras las determinaciones.*

*-Remitir una copia de los resultados de las determinaciones realizadas en la base central de datos del estudio PREDIMED PLUS o en la base de datos federada que se designara para ello, si existieran limitaciones técnicas de formato, para integrar los datos tras un tiempo de procesado que se establece en 2 años tras la recepción. En casos excepcionales se podría solicitar una prórroga. Además, cada vez que se elabore un paper con las determinaciones realizadas con muestras de varios nodos, se remitirá a cada nodo los datos correspondientes a las muestras aportadas.*

*-Establecer a su vez los acuerdos pertinentes de colaboración con los investigadores internos PREDIMED PLUS que realizarán los análisis de datos con las determinaciones efectuadas.*

*-Seguir la política de publicaciones del estudio PREDIMED PLUS y el plan de gestión e intercambio de datos de dicho estudio.*

### **III. Análisis de datos y publicaciones**

*El investigador externo tiene que garantizar la calidad de las determinaciones realizadas en las muestras. Detallará la metodología utilizada, el tipo de datos generados, el control de calidad realizado y aportará los metadatos correspondientes a cada variable medida, así como los datos a cada nodo del que haya utilizado muestras para las determinaciones.*

*Los investigadores internos realizarán los análisis estadísticos tras el link con las variables PREDIMED PLUS pertinente. Aportarán las sintaxis y datos correspondientes para cada artículo que se remita para publicación.*

*Los investigadores internos junto con el externo cumplimentarán el formulario de propuesta de publicación (o publicaciones) e indicarán los autores de la misma. En el caso de estudios originales, se seguirá la política de publicaciones para fijar las autorías, teniendo en cuenta si se trata de una colaboración con los datos del estudio general en la muestra total, o con proyectos específicos, así como la contribución del investigador externo y los internos implicados.*

*En caso de estudios colaborativos o meta-análisis en los que participaran muchos autores de otras cohortes y estuviera limitado el número de autores, se seguirá como norma general incluir el máximo número de autores del estudio PREDIMED PLUS, estableciendo también un rotatorio acumulable para distintos meta-análisis.*

*La propuesta de publicación tiene que ser aprobada por el Steering Committee y los resultados en forma de Tablas y el artículo completo tiene que ser presentado y aprobado antes de remitir para publicación para realizar los oportunos controles de calidad.*

### **IV. Coste**

*El trámite búsqueda de muestra, preparación, envío con hielo seco, etc. tiene un coste en tiempo para los investigadores del estudio PREDIMED PLUS, así como un coste económico. Por lo tanto, en consonancia con el acceso a datos de otras grandes cohortes internacionales, se establece una tasa que se hará efectiva tras la aprobación de la solicitud. El ingreso se realizará en la cuenta PREDIMED PLUS (CIBER OBN) habilitada para ello a través de transferencia bancaria. Las tasas varían según el tipo de colaboración contemplando desde la situación exenta de coste hasta 5000 euros (ver documento de Tasas para datos y muestras según tipo de colaboración solicitada y volumen de datos requeridos)*

#### **2.2.2. A4. Procedimiento de acceso/intercambio de datos con investigadores externos que solicitan muestras y datos**

Este caso normalmente implica la solicitud adicional de un proyecto de investigación entre investigadores PREDIMED PLUS e investigadores externos. Se cumplimentará el documento de solicitud en el Anexo 7 y se seguirá la Guía detallada en el Anexo 8. Tiene unas características similares a la solicitud combinada de ambas situaciones de datos y muestras anteriores más la solicitud de datos del proyecto de investigación. En este caso también habrá que cumplimentar un **acuerdo de colaboración** para solicitud conjunta de datos y muestras que se detalla en el Anexo 9. Se aplicarán los postulados de los documentos de acceso a datos y a muestras de manera integrada.

#### **2.2.2. A5. Procedimiento de intercambio de datos con el público en general**

Dadas las limitaciones éticas de los datos que se generan en el estudio PREDIMED PLUS, no está contemplado el intercambio de datos crudos con el público en general. Por ello, no se depositarán datos crudos del PREDIMED PLUS en repositorios de acceso totalmente abierto. Si se remiten los artículos a alguna revista que lo exija, se comunicará a la revista que los datos analizados están sujetos a restricciones debido a consideraciones éticas y se indicará que se podrán facilitar los datos a investigadores acreditados previa solicitud formal (se cumplimentará el impreso de solicitud de datos).

Con el público en general se podrán compartir metadatos, artículos originales de los resultados del estudio, documentos, videos, material didáctico y similar.

Como extensión de las consideraciones anteriores, de manera extraordinaria también se podría contemplar la posibilidad de que en paralelo a la publicación de los resultados principales finales del proyecto en una revista de alto impacto, se pudiera facilitar una base de datos reducida con las variables específicas del artículo en formato abierto al público general si así lo exigiera la revista, y se tratara de un importante valor añadido que justificara esta decisión. Para ello se solicitaría la autorización específica correspondiente al Comité de Ética de cada nodo reclutador, detallando de manera específica las variables que se liberarían y los procedimientos que asegurarían la anonimización. Si el pronunciamiento de cada Comité de ética fuera favorable se procedería. Si de alguno no lo fuera, no se podrían incluir los datos correspondientes a dicho nodo.

#### **2.2.2. B. Especificar qué métodos o herramientas de software serán necesarias para acceder a los datos.**

Actualmente el sistema de gestión de datos del estudio PREDIMED PLUS sólo permite la distribución de archivos de datos. Estos archivos de datos se pueden facilitar en formato texto o en formatos más utilizados de los distintos paquetes estadísticos (SPSS, STATA, etc). Se accede a los mismos online mediante contraseña personalizada.

En un futuro, si se dispone del enclave de datos, el acceso a los datos será a través de dicho sistema que ha incluye el software necesario de acceso y análisis de datos.

### **2.2.2. C. Especificar dónde (en qué repositorios, enclaves, etc.) se depositarán los datos y los correspondientes metadatos**

Los datos están depositados en los distintos nodos que los generan y en la base de datos central del IMIM. Se podrá establecer un sistema más dinámico de base de datos federada y así fuera requerido. Los archivos de datos que se van generando periódicamente y aquellos a los que se accede mediante solicitud autorizada se distribuyen también a los distintos nodos que son responsables de su seguridad.

Se creará un repositorio propio del estudio PREDIMED PLUS para depositar, mediante acceso cerrado, tras identificación y autorización del investigador, los datos correspondientes a archivos de datos generados para publicaciones que se hayan comprometido a facilitar un acceso controlado a datos a los investigadores que lo requieran.

Los metadatos y documentos se depositarán en repertorios públicos abiertos (los de las universidades correspondientes, Zenodo, etc). Si se obtiene financiación de proyectos NIH los datos implicados tendrán que ser depositados en los repertorios que especifique la financiación, siempre manteniendo un acceso controlado por el tipo de dato y su restricción ética. Entre estos repositorios mencionamos dBGap (ver más detalle de los repositorios en los epígrafes siguientes).

En un futuro, para garantizar la mejor seguridad, confidencialidad, privacidad y buenas prácticas de acceso y uso de datos, se pretende crear un enclave de datos virtual en el estudio PREDIMED PLUS (ver detalle en epígrafes siguientes).

### **2.2.3. Haciendo los datos interoperables**

#### **2.2.3.A. Establecer la interoperabilidad de los datos. Especificar qué datos, metadatos, vocabularios, estándares o metodologías se seguirán para facilitar la interoperabilidad**

Para facilitar la máxima interoperabilidad de los datos, éstos se almacenarán en formatos agnósticos de plataforma. Así los datos en hojas de cálculo se almacenarán como .csv; los datos en archivos de texto libre se almacenarán como .txt y las imágenes se almacenarán como .txt. En los casos en los que existan datos que de manera particular o adicionalmente a los ya expresados, necesiten almacenarse en formato propietario, tendrán asociado el software correspondiente (incluyendo número de versión), así como la INFO.txt.file al respecto.

Otros modelos de datos se ajustarán a las estructuras de intercambio ampliamente definidas en los estándares, incluyendo: a) Archivos semánticos en formato XSD (XML Schema Definition), clasificados según su contenido; b) Documentos explicativos en formato PDF (Portable Document Format).

Los metadatos seguirán los códigos estándar siempre que sea posible.

### **2.2.3.B. Establecer si se utilizarán vocabularios estándar para todos los tipos de datos generados**

Cuando sea posible se utilizarán vocabularios estándar para los datos generados. Se tendrá en cuenta el conjunto de estándares que facilita el HL7 (Health Level Seven) para el intercambio electrónico de información clínica, cuando sea pertinente. Del mismo modo, se utilizarán vocabularios específicos para conceptos clínicos tomando como referencia SNOMED-CT (Systematized Nomenclature of Medicine – Clinical Terms), cuando esté indicado. Para otros tipos de datos se utilizarán codificaciones internas, que serán claramente específicas en el archivo correspondiente.

### **2.2.4. Haciendo los datos reutilizables**

#### **2.2.4. A. Especificar cómo se permitirá la reutilización de los datos y que licencias se utilizarán**

La diseminación y reutilización de los datos se ajustará al marco ético y legal en el que se lleva a cabo el estudio PREDIMED PLUS.

**Privacidad y confidencialidad:** En PREDIMED PLUS se utilizan datos de carácter personal (identificativos y de salud) y se tienen que cumplir los preceptos de la legislación de protección de datos (regulados en España por la Ley de Protección de Datos de Carácter Personal). Esta ley es de aplicación a los datos de carácter personal registrados en cualquier soporte físico. El tratamiento de los datos cubre las actividades de recolección, registro, almacenamiento, recuperación, consulta, uso y diseminación. Este hecho, junto con lo especificado en el consentimiento informado del participante, determina la limitación de la reutilización de datos de manera abierta, siendo necesario implementar restricciones.

Teniendo en cuenta estos factores, los datos (en sentido amplio, incluyendo documentos y resúmenes de datos) en el PREDIMED PLUS podrán responder a las **tres categorías** generales reconocidos **de datos**:

**a-Datos públicos:** pueden ponerse sin restricciones a disposición de cualquier usuario en acceso abierto. Esta opción podrá utilizarse en pocas ocasiones en el estudio PREDIMED PLUS dadas las características de los datos.

**b-Datos restringidos:** sólo pueden ser consultados por determinados usuarios. Se utilizará esta opción preferentemente en el PREDIMED PLUS. Para ello en lugar de hacer disponibles los datos de manera pública en repositorios de acceso libre o como archivos suplementarios en los artículos publicados, se indicará siempre que el acceso a los datos será restringido y se podrá solicitar bajo petición por investigadores acreditados.

**c-Datos privados:** no se pueden hacer públicos. Son confidenciales. Algunos datos del PREDIMED PLUS tendrán que mantenerse siempre bajo esta categoría.

Además de la confidencialidad, los usuarios que reutilicen los datos deben cumplir las condiciones de la licencia y los permisos de utilización, reconociendo los derechos de **propiedad intelectual** de los investigadores productores de los datos. Según la legislación española, la propiedad intelectual es el conjunto de derechos que corresponden a los

autores y a otros titulares respecto de las obras y prestaciones fruto de su creación. En España la norma principal que regula los derechos de propiedad intelectual es la Ley de Propiedad Intelectual (Real Decreto Legislativo 1/1996 de 12 de abril por el que se aprueba el Texto Refundido de la LPI) que ha sufrido varias modificaciones, entre ellas la operada por la Ley 23/2006 de 7 de julio con el objeto de adaptar la normativa española a las nuevas circunstancias creadas por la sociedad de la información, y otra modificación posterior en 2014.

Por lo que respecta a los derechos que conforman la propiedad intelectual se distinguen los derechos morales y los derechos patrimoniales. En referencia a los Derechos morales, en comparación con los sistemas anglosajones, la legislación española es claramente defensora de los derechos morales, reconocidos para los autores y para los artistas intérpretes o ejecutantes. Estos derechos son irrenunciables e inalienables, acompañan al autor durante toda su vida. Entre ellos destaca el derecho al reconocimiento de la condición de autor de la obra, y el de exigir el respeto a la integridad de la obra o actuación y la no alteración de las mismas. Entre los Derechos de carácter patrimonial hay que distinguir entre los Derechos relacionados con la explotación de la obra o prestación protegida, que a su vez se subdividen en Derechos exclusivos y en Derechos de remuneración.

Faltaría clarificar qué tipos de datos se consideran “obras” para perfilar mejor estos derechos. Según el artículo 12 del mencionado TFPIR, las colecciones de datos y las bases de datos están protegidas por propiedad intelectual, mediante el denominado derecho sui generis, en cuanto que constituyen creaciones intelectuales. “La protección se refiere únicamente a su estructura en cuanto forma de expresión de la selección o disposición de contenidos”, no a los datos mismos. Los derechos de autor pertenecen a sus creadores, siempre que se trate de trabajos originales.

Los derechos de explotación o **copyright** son transferibles. El titular de estos derechos posee su ejercicio exclusivo y no pueden ser realizados sin su autorización, salvo en los límites que establece la ley. Los derechos de explotación constituyen una serie de actos como el de reproducción, distribución, comunicación pública y transformación.

El depósito de los **datasets** en un repositorio implica el ejercicio de los derechos de explotación, por lo que se requiere el permiso explícito del titular de dichos derechos mediante un acuerdo de cesión no exclusiva de los derechos necesarios.

**Licencias alternativas** al copyright: Existen licencias estándares y libres que el autor puede aplicar a sus datos de investigación para proporcionar los términos en los que compartir y reutilizar dichos datos en el ámbito de Internet. Un ejemplo de dichas licencias son las **Creative Commons**. Creative Commons es una corporación sin ánimo de lucro basada en la idea de que algunas personas pueden no querer ejercer todos los derechos de propiedad intelectual que les permite la ley. Las licencias Creative Commons permiten a las obras ser compartidas y reutilizadas en términos que son flexibles y legales. Creative Commons ofrece una base de seis licencias que permiten la copia, distribución, descarga y transformación de los documentos digitales. El Enlace para descargar los logos de alta resolución de Creative Commons se encuentra en <https://creativecommons.org/about/downloads/>. Las licencias Creative Commons (CC) se pueden aplicar a cualquier tipo de trabajo, incluyendo recursos educativos, fotografías, bases de datos y muchos otros tipos de contenido creativo. Los cuatro preceptos son: RECONOCIMIENTO (Attribution): En cualquier explotación de la obra

autorizada por la licencia hará falta reconocer la autoría (BY). NO COMERCIAL (Non Commercial): La explotación de la obra queda limitada a usos no comerciales (NC). SIN OBRAS DERIVADAS (No Derivate Works): La autorización para explotar la obra no incluye la transformación para crear una obra derivada (ND) y COMPARTIR IGUAL (Share alike): La explotación autorizada incluye la creación de obras derivadas siempre que mantengan la misma licencia al ser divulgadas (SA). La combinación de estos cuatro preceptos da lugar a los seis tipos de licencias. Reconocimiento de autoría (**CC BY**); Reconocimiento de autoría – compartir en idénticas condiciones (**CC BY-SA**); Reconocimiento de autoría – Sin obra derivada (CC BY-ND); Reconocimiento de autoría – Sin uso comercial (**CC BY-NC**); Reconocimiento de autoría – Sin uso comercial – compartir en idénticas condiciones (CC **BY-NC-SA**) y Reconocimiento de autoría – Sin uso comercial – Sin obra derivada (**CC BY-NC-ND**).

**CC0**: Acceso totalmente abierto. Esta licencia se utiliza cuando no se quiere disponer de ningún derecho, ni autoría ni similares, lo que se suele llamar un archivo completamente público.

Normalmente la descripción de los datos y metadatos que se remite a los repositorios o se libera está bajo CC0.

De acuerdo con el tipo de dato que se estime oportuno compartir de esta manera (metadatos, documentos de texto sobre manuales, recomendaciones, datos agregados de los análisis estadísticos o algunas variables que no requieran confidencialidad), en el PREDIMED PLUS para su reutilización, teniendo en cuenta las anteriores consideraciones de Confidencialidad, privacidad y propiedad intelectual, se optará por un tipo u otro de licencia cuando se deposite en el repositorio pertinente.

Sin embargo, la política general de re-uso de datos en el PREDIMED PLUS será utilizar licencias y repositorios que permitan un acceso cerrado a los datos depositados. En general se preferirá que el acceso controlado a los datos sea autorizado por el PREDIMED PLUS.

La solicitud para la autorización a la reutilización de datos será formulada al SC del PREDIMED PLUS de acuerdo con los formularios diseñados para ello. Tras examinar la solicitud se decidirá la pertinencia o no de facilitar el acceso a los datos por los investigadores solicitantes.

#### **2.2.4. B. Especificar cuándo los datos estarán disponible para su re-utilización. Especificar qué período de embargo es necesario**

En general, los datos no estarán disponibles para su utilización inmediata tras su generación. Se especificará un período de embargo que dependerá de cada situación y si responde o no a un proyecto de investigación financiado con fondos internacionales que especifique en su convocatoria un determinado período de embargo que hay que cumplir. Igualmente, cuando se trate de datos que constituyen los datos crudos de un artículo publicado en una revista que especifique una determinada disponibilidad para su reutilización, se atenderá con preferencia la temporalidad especificada, siempre que sea compatible con la política de intercambio de datos en los artículos publicados del PREDIMED PLUS que implica el requisito de realizar una solicitud previa para el acceso a dichos datos.



**2.2.4. C. Especificar de qué manera los datos generados por el proyecto pueden utilizarse por terceras partes, en particular cuando el proyecto termine. Especificar si algunos datos no permiten su reutilización.**

En el apartado 2.2.4. A. se ha especificado de manera general cómo se permitirá la reutilización de los datos y las licencias que se utilizarán. Esta reutilización de datos se podrá realizar tanto durante el desarrollo del proyecto como cuando el proyecto esté ya finalizado. El acceso a datos durante el período de vigencia del proyecto estará más restringido por la propia finalidad del proyecto y las posibles interferencias en los resultados finales del ensayo clínico. Tras finalizar el proyecto, el acceso a los datos podrá ser más oportuno.

Como principios generales del estudio PREDIMED PLUS, los datos crudos correspondientes a los distintos cuestionarios y mediciones realizadas en los participantes que sean generados en el proyecto serán utilizados por los investigadores del estudio PREDIMED PLUS y no serán liberados de manera pública para su uso por terceras partes sin ninguna restricción. Dado que se trata de datos de salud para los que hay que respetar los principios de confidencialidad y seguridad reconocidos por la legislación española y por las políticas de intercambio de datos de la Comisión Europea, estos **datos crudos no** se depositarán en **repertorios públicos de acceso totalmente abierto**.

Sin embargo, sí que se podrá permitir un acceso restringido a ciertos datos para aquellos **investigadores acreditados** que **formulen una solicitud específica** de datos cumplimentando el formulario de solicitud de datos que se ha diseñado al respecto y recibiendo la autorización correspondiente por el Steering Committee del estudio PREDIMED PLUS. Estos datos serán completamente anonimizados eliminando cualquier tipo de identificador de paciente, aunque sea numérico, con objeto de minimizar las probabilidades de identificación del mismo.

No se autoriza a ningún investigador principal a que pueda distribuir los archivos de datos generales y específicos que se le faciliten, mencionados en 1.2, a personal no autorizado que no pertenezca a su nodo. Los IPs velarán para que los archivos de datos no se distribuyan fuera de su grupo de investigación.

También está contemplado el facilitar el **uso de datos agregados** de distintas variables originales en el PREDIMED PLUS para aquellos **investigadores que soliciten dicha colaboración** a través del formulario correspondiente. El uso de datos agregados plantea menos problemas de confidencialidad y de mal uso de variables, por lo que se contempla como una opción de preferencia para el intercambio de datos con otros investigadores internacionales. Estas colaboraciones también requerirán aprobación previa por el SC del PREDIMED PLUS.

Debido al tipo de datos de salud que se generan, la forma recomendada por los DMPs de las distintas agencias es utilizar un portal Web creado y mantenido por los investigadores del proyecto, en este caso una versión mejorada de la página Web oficial del estudio PREDIMED PLUS. Tras solicitud mediante la cumplimentación del formulario oficial y la aprobación pertinente, el investigador externo al proyecto recibirá una contraseña única para acceder a las variables autorizadas y anonimizadas. El actual sistema de base de datos y extracción de ficheros implica que el usuario externo tendrá acceso a las variables crudas que puede copiar, modificar y analizar con su propio software. Este acceso y uso requiere el

compromiso de buenas prácticas, no alterar los datos y respetar los plazos y destrucción de las variables una vez agotado en plazo estipulado para su utilización.

Sin embargo, como alternativa ideal se recomienda que el proyecto posea el denominado “**enclave de datos**” (Vie et al, 2013; Bao et al, 2016), que aunque requiere de mayor inversión económica, garantiza un acceso seguro a los datos, sin posibilidad de copia, alteración o distribución posterior de las variables autorizadas. Un enclave de datos es un entorno seguro de datos en el que las personas autorizadas pueden realizar análisis estadísticos de los datos sin capacidad de modificarlos bajo acceso restringido ([https://grants.nih.gov/grants/policy/data\\_sharing/data\\_sharing\\_guidance.htm#enclave](https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#enclave)).

Existen múltiples ejemplos de enclaves de datos que pueden ser tomados de referencia para el estudio PREDIMED PLUS, entre ellos el de la escuela de Salud Pública de Harvard incluyendo el denominado “The Channing Cohorts computer system” (Bao et al, 2016), el de la Universidad de Michigan (<https://micda.psc.isr.umich.edu/enclave/>) y el de la Universidad de Chicago (<http://www.norc.org/Research/Capabilities/Pages/data-enclave.aspx>), entre otros.

El enclave de datos permite al usuario externo tener un espacio de trabajo en un ordenador central, donde también se incluye distinto software de análisis estadístico, de manera que el investigador externo ejecuta las instrucciones de análisis mediante programación específica y obtiene el output de resultados correspondientes. El **acceso puede ser “in situ”** o también mediante **acceso remoto “online”**. Aunque en la actualidad el estudio PREDIMED PLUS no cuenta con este sistema, se está trabajando para poder implementarlo en un futuro.

Los datos generados en presentaciones públicas de resultados y artículos publicados, la elaboración de cuestionarios, protocolo del estudio, material de intervención dietética, material de intervención de actividad física, material de intervención conductual, recogida de eventos, etc. pueden ser de interés para otros grupos de investigación por lo que sí que se pondrán poner en acceso abierto y estarán disponibles para toda la comunidad científica a través de la página web del proyecto ([predimedplus.com](http://predimedplus.com)), tras registro de su propiedad intelectual.

Aunque como primera opción se pretende depositar los datos crudos en un portal o enclave creado y administrado por los investigadores PREDIMED PLUS, de manera parcial para ciertos conjuntos de variables y también como opción a largo plazo y una vez finalizado el estudio, se contempla depositar subconjuntos de datos en el repositorio público **Zenodo** patrocinado por Unión Europea.

Se ha elegido este repositorio (<https://zenodo.org>) porque permite depositar datos en acceso cerrado además del acceso restringido, bajo embargo y el acceso abierto (<https://zenodo.org/policies>). En Zenodo se pueden depositar archivos restringidos con la posibilidad de compartir el acceso con otros si se cumplen ciertos requisitos. Estos archivos no se pondrán a disposición del público y el intercambio sólo será posible con la aprobación del depositante del archivo original. El repositorio está alojado en el CERN y su soporte técnico procede del software Invenio. El acceso a metadatos y archivos de datos se proporciona a través de protocolos estándar como HTTP y OAI-PMH. Zenodo es la alternativa cerrada a OpenAirPlus, ya que la Comisión Europea es consciente de que ciertos datos generados en la investigación médica en humanos no pueden compartirse en abierto

y necesitan un acceso restringido o cerrado. Zenodo ofrece inicialmente una garantía de almacenamiento a largo plazo (20 años), así como una capacidad de 50GB por dataset.

Aunque se afirma que Zenodo todavía está en su infancia, y pueden existir algunas dudas sobre su mantenimiento a largo plazo, desde la Unión Europea se nos indica que se puede confiar en Zenodo ya que en principio está comprometida la financiación como infraestructura relevante y además el CERN está familiarizado con la preservación de grandes conjuntos de datos de investigación de varios petabytes. En el improbable caso de que Zenodo tenga que cerrar, se garantiza una migración de todo el contenido a otros repositorios adecuados, y como todas las subidas tienen DOI, no se verán afectadas todas las citas y enlaces a los recursos de Zenodo. Zenodo está desarrollado por el CERN bajo el proyecto FP7 de la UE OpenAIREplus (acuerdo de subvención nº 283595) y OpenAIRE2020. En este repositorio se garantiza la preservación de archivos de manera que los archivos de datos y los metadatos se respaldan todas las noches y se replican en múltiples copias en el sistema en línea. Además, todos los archivos de datos se almacenan junto con una suma de comprobación MD5 del contenido del archivo. Los archivos se comprueban regularmente contra sus sumas de comprobación para asegurar que el contenido del archivo permanece constante.

Adicionalmente a los datos generados en el proyecto PREDIMED PLUS general, existen otros subproyectos que pueden generar nuevos datos con financiación adicional que tendrá que solicitarse. De estos proyectos financiados se obtendrán nuevos datos que pueden ser genómicos, como por ejemplo GWAs a gran escala, que si son financiados por los NIH, los datos genómicos de estos GWAs, así como las principales variables fenotípicas tienen que depositarse en la base de datos **dbGap** (The Database of Genotypes and Phenotypes). dbGaP (Tryka et al, 2014)) es un repositorio patrocinado por los NIH encargado de archivar, curar y distribuir información producida por estudios que investigan la interacción entre el genotipo y el fenotipo. Fue lanzado en 2006 como respuesta al desarrollo de la política del NIH sobre GWAS y proporciona acceso a estudios genéticos financiados por los NIH y otras agencias en todo el mundo.

**dbGap** tiene un **acceso público** en el que se puede acceder libremente a la información general sobre los estudios presentados, los datos de nivel de resumen y los documentos relacionados con los estudios (<http://www.ncbi.nlm.nih.gov/gap>). El depositar en dbGap datos agregados del PREDIMED PLUS, metadatos o datos resumen, no plantea ningún problema para el acceso abierto que se implementa a los mismos.

En la parte de acceso público de dbGap no se puede acceder a los datos individuales. A ellos se accede a través de la parte de **acceso restringido de dbGap** (<https://dbgap.ncbi.nlm.nih.gov/aa>). A esta parte sólo pueden acceder investigadores acreditados tras solicitud y cumplimentación de un formulario en el que se hayan indicado los objetivos de la investigación y se demuestre la capacidad de proteger adecuadamente los datos. Actualmente PREDIMED PLUS está en los primeros pasos de la formación de un Consorcio impulsado por los NIH, y que incluye a ensayos clínicos aleatorizados de intervención con estilo de vida para la pérdida de peso, y es probable que se pueda obtener algo de financiación para realizar GWAs financiados por los NIH por lo que podríamos encontrarnos en la situación de tener que depositar parte de los datos del estudio PREDIMED PLUS (genéticos de genoma completo y los correspondiente fenotipos de variables antropométricas) en dbGap, tanto en su parte pública como de acceso restringido.

El depósito de los datos en el repositorio de acceso restringido requiere la autorización institucional previo informe favorable del Comité de Ética correspondiente. Existe también un período de embargo en este repositorio. Para acceder al sistema de acceso autorizado, los usuarios que no son NIH deben tener una cuenta de NIH eRA Commons.

Versiones similares a dBGap también se pueden encontrar en otros repositorios europeos. Entre ellos destaca "The European Genome-phenome Archive (EGA)" (Lappalainen et al, 2015). EGA fue lanzado en el año 2008 por el European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) para apoyar el archivo voluntario de datos que requerían un almacenamiento seguro. Recientemente, se han ampliado sus capacidades con la colaboración del Centre for Genome Regulation (CRG).

Finalmente, es importante tener en mente la existencia de la recientemente creada plataforma **vivli** ([www.vivli.org](http://www.vivli.org)). A pesar de que se han postulado nuevos requisitos que los datos de ensayos clínicos para intercambio de datos, todavía no ha habido un esfuerzo organizado para coordinar las plataformas y servidores existentes y proporcionar una plataforma básica para permitir a la mayoría de los generadores de datos compartir sus datos de ensayo de forma sencilla. Actualmente se está creando una entidad de este tipo, patrocinada por el Centro Multi-Regional de Ensayos Clínicos de Brigham and Women's Hospital y la Universidad de Harvard (MRCT Center). Esta plataforma se denomina "Vivli", (el griego "vivlithíki", o biblioteca, y el latín "viv", o vida.). El objetivo de Vivli es conectar las plataformas y comunidades de intercambio de datos existentes, al tiempo que recibirá datos de investigadores que aspiran a compartir datos, pero que no tienen recursos para hacerlo (Bierer et al, 2016).

Como **consideración importante** en el depósito de datos crudos en distintos repositorios o soportes es que las **variables se tienen que depositar sin ningún tipo de identificador**. Por seguridad se recomienda eliminar la variable "identificador" (Tucker et al, 2016), aunque se trate de un identificador numérico, ya que siempre cabe el riesgo de que se pueda localizar el archivo que relaciona los ID del PREDIMED PLUS con el nombre y apellidos del participante.

#### 2.2.4. D. Describir el control de calidad de los datos generados

En el estudio PREDIMED PLUS se ha implementado un control de calidad interno de los datos generados que incluye todas las etapas desde la producción de los mismos siguiendo protocolos y manuales de recogida de datos estandarizados y detallados, hasta el envío, el almacenamiento, el control de *outliers* y *missings* y el doble chequeo por parte de los responsables de la base de datos central y los responsables de cada nodo que los genera en base a *queries* periódicas. Paralelamente, también se aplican otros procedimientos de control como auditorías selectivas en los distintos nodos y procesos. Básicamente cada nodo es el responsable de verificar la calidad de sus datos, estableciendo controles de calidad para asegurar que los datos remitidos a la base de datos se corresponden con los datos obtenidos. Será responsable de revisar los outliers de su nodo y comunicar su rectificación a la base central, si se trata de errores o confirmar el valor revisado. Paralelamente, se establecerá un control de calidad de datos al azar en todos los nodos para comprobar la validez de los mismos que se realizará en las visitas de calidad a los nodos. Este control de calidad sigue las directrices de la Guía de buenas prácticas para la

preparación de datos y archivo de los mismos a través de todo el ciclo de vida útil, publicada por La Inter-University Consortium for Political and Social Research (ICPSR), 2012.

Uno de los aspectos importantes a decidir para el uso consistente de los datos generados es el tratamiento de los **missings mediante imputación** o no. Inicialmente los ficheros se generarán con los datos faltantes no imputados. Sin embargo, ya que para muchos análisis se requieren datos imputados de manera complementaria a los datos reales para análisis de sensibilidad u otro tipo de análisis, se estima conveniente generar también datos imputados de los valores faltantes para determinadas variables. La imputación de estos datos se realizará de manera centralizada por personas expertas en el tema, y a ser posible encargando la imputación a dos o más grupos para que puedan comparar y consensuar los mejores métodos y estrategias de imputación. El fichero de imputación resultante se hará accesible para todos los investigadores PREDIMED PLUS para que las publicaciones que se deriven se realicen con los mismos valores de imputación si es pertinente.

Para asegurar el control de calidad del procesamiento de datos y generación de resultados, se tendrá que generar y conservar una sintaxis para el tratamiento y análisis estadístico de los datos correspondientes a cada artículo, que junto con los datos correspondientes, tendrán que estar disponibles para consulta de los autores del artículo si así lo requieren. Los autores principales del artículo serán los responsables de velar por la calidad de los datos y análisis.

También está previsto como control de calidad y seguridad que una vez publicado un artículo, el investigador principal del artículo facilite la sintaxis del análisis estadístico de los datos del artículo, así como los mismos al IP de estudio que se encargará de habilitar un repositorio con las garantías de seguridad informática necesarias que garanticen el acceso restringido.

#### **2.2.4. D. Preservación de los datos**

En el estudio PREDIMED PLUS, se irá mejorando el sistema de gestión de datos con recursos adicionales para que los datos sean preservados y puedan ser utilizados para la investigación futura el mayor número de años posible. Dado que se trata de un estudio prospectivo, los datos del proyecto general se irán generando paulatinamente durante un período de varios años (hasta 2020). Posteriormente se seguirán generando datos a través de proyectos específicos así como la recopilación de un mayor número de eventos en un potencial “*extended follow-up*”. Al menos se prevén unos 15 años de preservación activa de datos. El sistema de gestión de datos se diseñará para facilitar el plan de preservación de datos a largo plazo conforme a estándares internacionales. En principio, todos los datos generados se van a preservar. Se realizarán copias de seguridad de forma regular, para que puedan ser utilizadas en caso en que haya que restaurar los ficheros originales. De manera estándar como mínimo se realizará un backup incremental de los ficheros a diario y un backup completo de manera semanal. Se comprobará también la integridad de los ficheros mediante la comprobación del código *MD5 checksum value*, el tamaño del fichero y la fecha. En la política de preservación de datos también contemplaremos la posible obsolescencia del hardware y del software.

### **2.3. Proceso de almacenamiento de datos y recursos**



Recientemente, el estudio PREDIMED PLUS inauguró un nuevo sitio web para facilitar el acceso a los usuarios y participantes en el proyecto. Tiene un área pública y un acceso privado para el personal de investigación. El contenido privado está protegido por una contraseña y sólo está disponible en español. Actualmente la contraseña es común para cada centro y miembro del personal de investigación. Esto es una limitación importante para la calidad, integridad y seguridad y en breve tendría que mejorarse para proporcionar una contraseña diferente para cada nodo que introduce los datos.

La plataforma central de Teleform se encuentra en el centro coordinador de datos, IMIM, en Barcelona. Esta plataforma permite administrar tanto formularios electrónicos como en papel y documentos. Los formularios electrónicos son cumplimentados por el personal del centro de reclutamiento, utilizando un navegador de Internet común con el complemento de Acrobat Reader, y se envían electrónicamente al IMIM. Los formularios en papel son descargados y administrados por el personal del centro de reclutamiento y enviados al IMIM por correo postal cuando fuera necesario. Todo el conjunto de formularios electrónicos se desarrollan y administran con la plataforma Teleform Enterprise para Captura Automática para Formas y Documentos. Paralelamente existen cuestionarios en papel de lectura óptica que se remiten por correo postal y se procedan en los nodos procesadores de datos generales que a su vez lo transfieren al IMIM para guardarlo en la base de datos central. De manera similar los datos de acelerómetros y electros de remiten de manera electrónica a los respectivos centros procesadores a la base de datos general del IMIM. En el IMIM se integran todos los datos en la base de datos central, se almacenan allí los datos y se realizan controles de calidad y copias de seguridad. Existe un equipo de administradores y gestiones de la base de datos en contacto directo con el coordinador del estudio PREDIMED PLUS. El almacenamiento de datos del Proyecto General es continuo a lo largo del estudio y diariamente se realiza en control de calidad anotando las incidencias diarias y remitiendo a los nodos distintas preguntas sobre datos missing, incompletos, outliers, etc.

Periódicamente se extraen ficheros de datos de cada nodo y se remiten al mismo para el control de calidad local, que una vez realizado contactará con el IMIM para notificar incidencias y subsanaciones.

Los datos revisados por los nodos se remiten al IMIM y se guarda la subsanación en la base de datos obteniendo una versión más depurada que se utilizará para extraer los archivos de datos completos del proyecto general con la información de todas las variables y participantes en cada tanda acordada.

Los datos generados por los proyectos específicos se almacenan inicialmente en el nodo que los genera, en bases de datos diseñadas para ello y de acuerdo con las características de cada dato generado. Tras un periodo de embargo, específico para cada tipo de proyecto y datos, se remiten los datos a la base central del IMIM para que queden allí almacenados y se puedan generar ficheros de datos con los mismos cuando sea oportuno. Según la naturaleza del dato generado en el proyecto específico, por ejemplo si se trata de datos ómicos de diferentes arrays de gran densidad, ocupan varias decenas de GB y por su estructura no se pueden almacenar en las bases de datos convencionales, por lo que se puede optar por guardar en las tablas los metadatos de los resultados, y estos en un repositorio propio no estructurado fuera de las bases de datos. El repositorio o repositorios estarían interconectados con las bases de datos dentro del esquema de base de datos federada. Actualmente en el mercado se comercializan unidades de discos de 2 a 5 GB de



3,5", con interface SATA de 6 GB/s, velocidad de rotación de 7.200 rpm, y buffer de 128 MB y tecnología NativeCommandQueuing. Montando varios de estos discos en racks o backplanes para discos duros intercambiables en caliente de múltiples bahías y gestión de redundancia, se puede empezar con una capacidad inferior a 10 TB e ir subiendo a medida que aumente la necesidad de almacenamiento. Probablemente, la tecnología de discos Solid State Disc (SSD) logrará destacables avances en los próximos años y se podrá disponer de unidades de capacidad similar a los discos convencionales, con mejores prestaciones de velocidad, por un coste sensiblemente inferior, y así se podrán sustituir los discos duros de rotación por discos SSD.

### **2.3.1 Mejoras a introducir y estimar los costes de hacer los datos FAIR**

El sistema de gestión de datos del que dispone actualmente el estudio PREDIMED PLUS es mejorable y tiene muchas limitaciones en el intercambio de datos debido a la necesidad de distribuir ficheros de datos con acceso directo a todos los datos del estudio, con lo cual se puede poner en peligro la seguridad, integridad, confidencialidad y privacidad, a la vez que facilita que pueda haber un mal uso de estas variables, en particular, en el caso de investigadores externos, a los que también se les facilitan los datos con el compromiso de destruirlos tras el período de vigencia del convenio, extremo que no se puede comprobar realmente.

Por ello, la mejor opción para hacer los datos FAIR sería implantar un sistema de gestión de datos basado en el modelo de enclave de datos (in situ/virtual) tal como se ha descrito en el punto 2.2.4. C. Este **enclave de datos** se gestionaría con distintos privilegios de acceso a determinadas variables según el perfil del usuario, tanto internos como externos. Un sistema de enclave de datos con funcionamiento a tiempo real es muy costoso, pero para el estudio PREDIMED sería necesario que funcionara todos los días de la semana permitiendo consultas y análisis a la base en cualquier momento, sino que para minimizar los costes podría operar en períodos especificados reduciendo el coste de infraestructura y de recursos humanos. Este coste no está estimado ahora pero lo estamos valorando para el futuro. Este sistema debe garantizar:

- La seguridad de la información almacenada en las bases de datos.
- El acceso controlado a los datos.
- Se debe supervisar en todo momento la actividad de los investigadores que acceden a los datos.
- Se debe proporcionar en la medida de lo posible el acceso a herramientas de datos internas para evitar la difusión incontrolada de los datos o el acceso no autorizado de programas.
- Se debe determinar el alcance exacto en el acceso de datos en función de los privilegios asociados al perfil de cada usuario.
- Se debe garantizar la integridad de los datos evitando su dispersión y asegurando el control de actualizaciones.
- Se debe garantizar la eficiencia en el acceso, gestión y presentación de los datos mediante la normalización del diseño de la base de datos y de las aplicaciones en una misma plataforma.

## **2.4. Seguridad de los datos**

Disponer de políticas de seguridad supone garantizar la integridad, disponibilidad y confidencialidad de los datos. La integridad nos asegura que los datos no van a sufrir modificaciones no autorizadas. Este aspecto es muy importante en PREDIMED PLUS, ya que actualmente, el sistema de gestión de datos que está vigente en el estudio no garantiza la integridad de datos cuando se distribuyen a los distintos investigadores (ya sean internos o externos), ya que se permite un acceso directo a las variables.

Para ello los responsables tienen que asignar contraseñas individuales para identificar y autenticar a las personas autorizadas para consultar o tratar datos, establecen qué personas pueden acceder al lugar físico en el que se encuentran los expedientes, etc, los sistemas tienen que disponer de recursos de timing out que cierran una sesión de trabajo si no se está haciendo nada, o asegurarse de que nadie pueda manipular un ordenador a través de Internet.

La disponibilidad permite que los datos estén siempre a disposición de las personas autorizadas, pudiendo ser recuperados, cuando algún evento físico o de cualquier otro tipo afecta a su funcionamiento normal.

La confidencialidad comporta que los datos sólo sean conocidos y accesibles a los usuarios autorizados del sistema de información. Cualquier problema de seguridad en éste ámbito puede afectar al deber de secreto. Por todo ello, la mejor opción sería utilizar el llamado “enclave de datos”.

Este epígrafe se complementa con el epígrafe 2.5 sobre aspectos éticos, en los que se detallan los requisitos de seguridad que tienen que tener los ficheros de los participantes en el estudio PREDIMED PLUS al incluir datos de salud, sujetos a la Ley de Protección de datos (La Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal y su Reglamento de desarrollo, aprobado por el Real Decreto 1720/2007, de 21 de diciembre). Esta Ley ha sido complementada por el reciente **REGLAMENTO (UE) 2016/679 DEL PARLAMENTO EUROPEO Y DEL CONSEJO** de 27 de abril de **2016** relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la directiva 95/46/CE. Reglamento general de protección de datos.

El nuevo Reglamento General de Protección de Datos (RGPD) entró en vigor en mayo de 2016 a nivel europeo y será aplicable a partir de mayo de 2018. En este periodo transitorio y aun cuando siguen vigentes las disposiciones de la Directiva 95/46 y las correspondientes normas nacionales de desarrollo, los responsables y encargados de tratamiento deben ir preparando y adoptando las medidas necesarias para estar en condiciones de cumplir con las previsiones del RGPD en el momento en que sea de aplicación. El RGPD es una norma directamente aplicable, que no requiere de normas internas de trasposición ni tampoco, en la mayoría de los casos, de normas de desarrollo o aplicación. Por ello, los responsables deben ante todo asumir que la norma de referencia es el RGPD y no las normas nacionales, como venía sucediendo hasta ahora con la Directiva 95/46. No obstante, la ley que sustituirá a la actual Ley Orgánica de Protección de Datos (LOPD) sí podrá incluir algunas precisiones o desarrollos en materias en las que el RGPD lo permite. En noviembre de 2017 se aprobó el anteproyecto de la nueva Ley de Protección de Datos en España, pero está todavía en trámite para su aplicación a partir de mayo de 2018. Las novedades de este RGPD se

comentan y detallan en la “Guía Reglamento General de Protección de Datos para responsables de tratamiento” que ha elaborado la Agencia de Protección de Datos de manera digital.

Desde el punto de vista ético y de seguridad es crucial mantener la confidencialidad de los datos a través la utilización de distintas estrategias que no permitan en modo alguno la identificación del paciente. Además de los identificadores directos, incluido el ID numérico que pueden identificar a una persona de manera directa, existen otros **identificadores, denominados indirectos**, de manera que el cruce de varios de estos identificadores podría permitir la identificación de una persona. Además de los datos clínicos y de estilo de vida, los datos genéticos, fundamentalmente si se trata de secuenciación directa o de análisis de alta densidad del genoma, pueden permitir identificar a la persona (Check-Hayden, 2013), por lo que por seguridad se aconseja mantener en archivos separados los datos genéticos de genotipado denso de otros datos clínicos o de estilo de vida.

Como norma general, se emplearán los archivos anonimizados, sin microdatos de nombre, apellidos, DNI, número de la seguridad social, etc. Aunque de manera interna se utilice el ID generado de manera numérica en el estudio PREDIMED para cada paciente. Este ID es susceptible de ser utilizado mediante malas prácticas para identificación del paciente ya que incluye clave de nodo y números correlativos de pacientes. Además, si no se ha extremado la seguridad en los archivos que contienen el link entre el nombre del paciente y su ID PREDIMED PLUS, se podrían filtrar fácilmente. Por ello, para intercambio con otros investigadores es necesario eliminar el ID PREDIMED PLUS en el archivo de datos, y en caso de que fuera necesario incluir algún ID, se utilizarán distintos algoritmos de anonimización como por ejemplo los algoritmos de HASH, como SHA1 o MD5, en el que el ID puede reemplazarse por la denominada clave o huella digital aleatoria que inequívocamente sirve como identificador no vulnerable de la persona.

Según la Agencia de Protección de Datos Personales en su Guía para la anonimización de datos (2016), el proceso de anonimización no puede asegurar la imposibilidad de reidentificación de las personas en términos absolutos, motivo por el cual se deben de tener en cuenta las garantías jurídicas necesarias para preservar los derechos de los interesados.

En cualquier Sistema de Gestión de Bases de Datos se debe garantizar su seguridad en los tres aspectos fundamentales: (a) La protección del sistema frente a ataques externos; (b) La protección del sistema frente a caídas o fallos en el software o en el hardware; (c) La protección frente a manipulación por parte de usuarios no autorizados. Además, dado que se debe garantizar la seguridad de las comunicaciones y de la información transmitida entre la base central y los nodos.

La seguridad del sistema frente a caídas o fallos en el software o en el hardware está basada en la redundancia y el respaldo. El valor operacional del sistema de gestión de base de datos no es lo suficientemente crítico para exigir la disposición de un hot site que pueda suplir el fallo del servicio en unas pocas horas. Sin embargo, sí que se dispone de un cold site para el sistema informático que soporta la Base de Datos Central y el Repositorio Público de Datos que se pretende generar en el PREDIMED PLUS, y de cierto grado de redundancia para las bases de datos bajo la responsabilidad de los nodos. En todos los centros se aplica un protocolo de copias de seguridad que consiste, en general, de la realización de copias incrementales en horas nocturnas, copias completas semanales, un

ciclo de seis copias incrementales (semana laboral más un día) y cinco copias completas (mes más una semana), todo ello automatizado. Se guarda un duplicado de cada copia en el ciclo en un armario ignífugo en una dependencia distinta y con acceso también restringido.

Para la protección frente a manipulaciones se implemente la estrategia de acceso compartimentado a las bases de datos en la escala de usuarios. Para cifrar y proteger la información que se transmite en la red entre las distintas bases de datos componentes del SGBDF, mediante el protocolo HTTPS, se aplica la tecnología Transport Layer Security (TLS) Protocol Version 1.2 de clave pública, con claves TLS\_RSA\_WITH\_AES\_128\_CBC\_SHA o AES\_256\_CBC\_SHA256, de 128 bytes de longitud. Con ello se garantiza la seguridad de cifrado, interoperabilidad, extensibilidad y eficiencia relativa en la transmisión de datos.

## **2.5. Aspectos éticos**

Las reglas que se aplican a los datos personales van más allá de la imposición de deberes para su obtención y obligaciones que garantizan que el responsable los tratamientos actuará adecuadamente. Para ello debe garantizarse la calidad de los datos, la seguridad y el secreto.

Tal como se ha detallado en los epígrafes anteriores, el estudio PREDIMED PLUS es un ensayo clínico multicéntrico en humanos en el que se recogen datos personales y de salud tanto de manera presencial como a través de la historia clínica, previo consentimiento. Además, cada IP de nodo ha registrado el fichero de datos en la Agencia Española de Protección de Datos, siendo el responsable del fichero y teniendo la obligación legal de cumplir las especificaciones al respecto recogidas en la Ley de Protección de Datos. La Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal y su Reglamento de desarrollo, aprobado por el Real Decreto 1720/2007, de 21 de diciembre, definen al responsable del fichero o tratamiento como la "persona física o jurídica, de naturaleza pública o privada, u órgano administrativo, que sólo o conjuntamente con otros decida sobre la finalidad, contenido y uso del tratamiento, aunque no lo realice materialmente. Podrán ser también responsables del fichero o del tratamiento los entes sin personalidad jurídica que actúen en el tráfico como sujetos diferenciados". Por fichero se entiende "todo conjunto organizado de datos de carácter personal, que permita el acceso a los datos con arreglo a criterios determinados, cualquiera que fuere la forma o modalidad de su creación, almacenamiento, organización y acceso" (artículo 5.1.k). Por otra parte, el tratamiento se define como "cualquier operación o procedimiento técnico, sea o no automatizado, que permita la recogida, grabación, conservación, elaboración, modificación, consulta, utilización, cancelación, bloqueo o supresión, así como las cesiones de datos que resulten de comunicaciones, consultas interconexiones y transferencias" (artículo 5.1.t).

Así, el principio de seguridad de datos establecido en el artículo 9 de la Ley Orgánica 15/1999, impone al responsable del fichero adoptar las medidas de índole técnica y organizativas necesarias que garanticen la seguridad de los datos de carácter personal y eviten su alteración, pérdida, tratamiento o acceso no autorizado. Estas medidas han sido desarrolladas en el Título VIII – del Reglamento de desarrollo de la LOPD, aprobado por el Real Decreto 1720/2007, de 21 de diciembre. Para facilitar a los responsables de los ficheros la adopción de las disposiciones que marca el reglamento, existe una Guía de Seguridad de Datos que resume las medidas de seguridad, las comprobaciones a realizar y

un modelo de documento de seguridad, que tienen que ser tenidos en cuenta por los IPs de cada nodo reclutador PREDIMED PLUS, así como por el SC al custodiar el archivo general integrado de los distintos nodos. De conformidad con lo dispuesto en el artículo 44.3.h) de la Ley Orgánica 15/1999, constituye infracción grave el "Mantener los ficheros, locales, programas o equipos que contengan datos de carácter personal sin las debidas condiciones de seguridad que por vía reglamentaria se determinen".

El tipo de población analizada y los datos obtenidos hacen que sea obligatorio también cumplir con los principios éticos nacionales e internacionales, así como la legislación aplicable al respecto. Cada IP ha tenido que tramitar el correspondiente permiso al Comité de Ética de su centro y/o región, obteniendo un informe favorable.

De acuerdo con el Artículo 6, el tratamiento de los datos de carácter personal requerirá el consentimiento inequívoco del afectado, salvo que la ley disponga otra cosa. Los participantes en el estudio han tenido leer la hoja de información al paciente sobre los detalles del estudio y firmar el modelo de consentimiento informado general del estudio PREDIMED PLUS. Este modelo consta de dos apartados: el general del estudio y el consentimiento para análisis genéticos. Lo que firma el participante en el consentimiento informado determina el alcance del uso posterior de los datos. Si no se ha especificado en el consentimiento informado que se va a realizar un intercambio abierto de los datos del paciente, éstos no pueden depositarse en repositorios públicos. Por ello, actualmente esta limitación ética en el estudio PREDIMED PLUS limita el intercambio de datos. Por lo que sólo podrá contemplarse la opción de intercambio de datos tras solicitud de los mismos por parte de un investigador acreditado, con acceso restringido y siempre que el motivo de intercambio de datos esté relacionado con los objetivos iniciales del proyecto. Este aspecto es importante, ya que muchas veces los datos personales que se recogieron para un objetivo determinado, se destinan para un propósito que es diferente de la colección inicial (también llamado "uso secundario" o "procesamiento adicional"). Como principio general enunciado en el artículo 5 del GDPR, el tratamiento de datos personales con fines distintos de aquellos para los que se recopilaron inicialmente los datos personales sólo debe admitirse cuando el nuevo objetivo del tratamiento sea compatible con los fines para los que se recogieron inicialmente los datos. A efectos prácticos, el nuevo anteproyecto de Ley de Protección de Datos, todavía restringe más estos aspectos así como los usos del consentimiento informado, no admitiendo un consentimiento informado genérico, sino que indica que el consentimiento informado tiene que ser más específico detallando los usos concretos que consiente.

Por todo ello, mientras no puedan implementarse alternativas de obtención de consentimientos informados más acordes con el actualmente propuesto intercambio de datos, como por ejemplo la propuesta de consentimiento informado dinámico (Budin-Ljøsne et al, 2017), es necesario preservar los datos del PREDIMED PLUS de un acceso abierto.

Así, en el estudio PREDIMED PLUS seguiremos las recomendaciones generales para el intercambio de datos cuando se trata de datos de salud de pacientes (Tucker et al, 2016), basadas en:

a) Anonimización/desidentificación de los datos: Los titulares de los datos son responsables de generar conjuntos de datos des-identificados que tienen por objeto ofrecer mayor

protección a la privacidad del paciente mediante el enmascaramiento o la generalización de otros identificadores;

b) Acceso controlado a los datos, incluido el uso de acuerdos de intercambio de datos por investigadores acreditaos. Debería establecerse un acuerdo de intercambio de datos jurídicamente vinculante, incluidos acuerdos para no descargar o compartir más datos y no tratar de identificar a los pacientes. Deben utilizarse niveles adecuados de seguridad para transferir datos o proporcionar acceso; Una solución es el uso de un sistema seguro de “enclave de datos” que proporciona salvaguardias adicionales.

## **2.6. Otros aspectos**

### **2.6. A. Indicar si se han seguido otros procedimientos para el DMP auspiciados por otros organismos financiadores o directrices nacionales o internacionales**

Para el desarrollo del DMP del estudio PREDIMED PLUS se han seguido las directrices generales de la Comisión Europea. Estas directrices generales están también impulsadas a nivel nacional por la Fundación Española para la Ciencia y la Tecnología (FECYT), y la Ley de la Ciencia e Innovación.

Actualmente, ninguno de los organismos financiadores del estudio PREDIMED requiere como obligatorio para nuestro proyecto disponer de un DMP o de compartición de datos en abierto. El esfuerzo de elaboración del DMP se ha realizado con expectativas de futuro.

## **2.7. Soporte adicional en el desarrollo del DMP**

Para crear este DMP se ha realizado una profusa revisión bibliográfica y normativa de los principales aspectos relacionados con el DMP tanto a nivel nacional como internacional, fundamentalmente en el marco de la Unión Europea y de Estados Unidos. También se ha contado con el asesoramiento puntual de expertos en cada tema, del Data and Safety Monitoring Board (DSMB) del PREDIMED PLUS (reunión del 20-1-2017) y con la colaboración de un ingeniero informático, auditor certificado CISA (Certified Information Systems Auditor). Ha aconsejado sobre las limitaciones actuales del sistema informático y las posibilidades de mejora del sistema de gestión de datos del estudio PREDIMED PLUS, aconsejando sobre posibles alternativas y viabilidad de las propuestas para facilitar el DMP. Este plan también ha sido revisado por el Steering Committee y por los investigadores PREDIMED PLUS que han aportado mejoras sustanciales.

El presente DMP para el PREDIMED PLUS ha sido creado siguiendo el modelo propuesto por la Comisión, publicado en 2016 (European Commission, Directorate General for Research & Innovation. H2020, 2016), utilizando la herramienta facilitada por el Consorcio CSUC.



# PREDIMED-PLUS Data Management & Sharing

## RESUMEN DE PUNTOS CLAVE

Barcelona 27 de abril de 2017

(revisado en Barcelona el 15-1-2018, revisado en Madrid el 22-1-18 y el 16-4-18)

### 1. Acceso interno a los datos:

1.1 Para asegurar la privacidad de los datos, la base de datos general nunca incluirá nombres, apellidos o DNI. Sólo incluirá un código numérico (ID) para la identificación del participante. La correspondencia entre el ID y el nombre apellidos y DNI queda restringida al nodo reclutador. Se implantarán los mecanismos necesarios para que la relación entre nombre y apellidos del paciente, DNI y su ID no pueda ser desvelada de manera pública. Tal correspondencia entre el ID y el respectivo nombre, apellidos y DNI se almacenará en soportes seguros sin conexión a internet o con posibilidad de fácil distribución externa.

1.2 Bajo solicitud, siguiendo el procedimiento indicado en el documento de “data management and sharing”, el IMIM proporcionará a todos los investigadores principales de PREDIMED-PLUS un archivo de datos completo de las variables generales del estudio. Este archivo tendrá un número de versión y una fecha de generación. Se prevé generar distintos archivos a medida que se vayan procesando los datos. Se comenzará por la generación del archivo de datos basales, y se continuará con la generación de datos al año, a los 3, 5 y 7 años. Este número de versión y fecha de generación tendrá que ser mencionado en los análisis de datos y publicaciones. La solicitud de otro tipo de datos o de otros períodos distintos a los mencionados ser podrá realizar de manera específica, siguiendo también el procedimiento indicado en el documento de “data management and sharing”.

1.3 Cuanto sea posible, se sustituirá este proceso de envío periódico de archivos de datos por otro sistema de enclave virtual de datos, en el que los datos son depositados y gestionados de manera segura, por acceso remoto para visualizarlos y analizarlos on-line sin permitir descargarlos a ningún ordenador personal, garantizando así una integridad y seguridad de los datos que son limitadas con el sistema actual.

1.4 Todo dato que se haya recogido en el 50% o más de los participantes de PREDIMED PLUS o se haya determinado en los casos de eventos primarios, se incluirán en la base de datos general que estará a disposición de todos los IPs de los nodos reclutadores. Para publicar manuscritos con estos datos que no son propios del proyecto general el IP deberá ponerse en contacto con el/los responsables de estos datos generados para formular una propuesta de artículo y remitirla al Steering Committee según la política de publicaciones.

1.5 En el caso de determinaciones analíticas muy costosas (metabólica, genómica, proteómica, metagenómica, epigenómica, etc.), con muestras de varios nodos, cuyos gastos se hayan sufragado mediante fondos conseguidos mayoritariamente por el nodo/s que hizo tales determinaciones, aunque dicho nodo tenga que remitir los resultados de las determinaciones a cada nodo que haya aportado muestras en el momento que se finalice y se circule el artículo correspondiente a dichas determinaciones, se respetará el período que transcurra hasta la aceptación del paper del nodo que ha realizado las determinaciones antes de aceptar propuestas de otros nodos para analizar dichos datos. Si algún IP tiene interés en estos datos antes de la aceptación del paper, deberá contactar con el responsable de las determinaciones realizadas para establecer una colaboración y formular una propuesta conjunta de artículo/os según política de publicaciones. Este contacto se puede establecer incluso a partir del inicio de las determinaciones. En todo caso, tanto antes como después de

la publicación del paper inicial, el nodo que ha realizado las determinaciones tendrá prioridad en la propuesta de nuevos papers relacionados con dichas determinaciones y se aconseja la propuesta de papers co-liderados en caso de que otro nodo tenga gran interés en el mismo tema.

1.6 Toda determinación nueva de una variable medida en los participantes de PREDIMED-PLUS que no sean de su nodo debe comunicarse a todos los investigadores de nodos reclutadores. El IP del nodo que efectúe tal determinación debe comunicarlo en un plazo nunca superior a dos meses.

## 2. Acceso externo a los datos

2.1 No se depositarán datos crudos del PREDIMED PLUS en repositorios de acceso totalmente abierto. PREDIMED PLUS se inició con antelación a las directrices del ICMJE (2016) y de la Comisión Europea sobre acceso universal a los datos, por eso los IPs no están obligados a facilitar el acceso abierto para toda la comunidad científica de los datos usados en publicaciones. Además, existe una limitación ética en facilitar el acceso libre a los datos, ya que en el consentimiento informado que firma el paciente no se autoriza tal acceso.

2.2 No se autoriza a ningún investigador principal a que pueda distribuir los archivos de datos generales y específicos que se le faciliten, mencionados en 1.2, a personal no autorizado que no pertenezca a su nodo. Los IPs velarán para que los archivos de datos no se distribuyan fuera de su grupo de investigación.

2.3 Los datos generados en presentaciones públicas de resultados y artículos publicados, la elaboración de cuestionarios, protocolo del estudio, material de intervención dietética, material de intervención de actividad física, material de intervención conductual, recogida de eventos, etc. pueden ser de interés para otros grupos de investigación por lo que sí que se pondrán poner en acceso abierto y estarán disponibles para toda la comunidad científica a través de la página web del proyecto (predimedplus.com), tras registro de su propiedad intelectual.

2.4 Si un grupo externo desea realizar una investigación utilizando datos de PREDIMED-PLUS, se podrá llegar a un acuerdo, pero se requerirá realizar antes una propuesta conjunta de colaboración (contando con al menos un IP PREDIMED-PLUS) y detallada. Sólo tras minucioso estudio por parte de Steering Committee del ensayo PREDIMED-PLUS, y de revisión de las implicaciones éticas, se podrá resolver si se puede firmar o no el acuerdo de colaboración, incluyendo la posibilidad de abonar una cuota.

## REFERENCIAS

- 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.
- Agencia Española de Protección de Datos. Orientaciones y garantías en los procedimientos de anonimización de datos personales. 2016. Documento digital.
- Agencia Española de Protección de Datos. Guía del Reglamento General de Protección de Datos para responsables de tratamiento. 2016. Documento digital [<http://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/index-ides-idphp.php>].
- Amarnath GuptaEmail authorWilliam BugLuis MarengoXufei QianChristopher ConditArun RangarajanHans Michael MüllerPerry L. MillerBrian SandersJeffrey S. GretheVadim AstakhovGordon ShepherdPaul W. SternbergMaryann E. Martone. Federated Access to Heterogeneous Information Resources in the Neuroscience Information Framework (NIF). *Neuroinformatics*. 2008; 6(3):205–217.
- ÁREA DE PROFESIONALES [<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. *Ecanermedicalscience*. 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](http://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. 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. Recomanacions per seleccionar un repositori per al dipòsit de dades de recerca. Versió 2, Novembre 2016. [Doc. 16/57: B6SR\GDR\1611RecomanacionsSeleccionarRepositoriDades.docx, 15.11.16].
- Grupo de Trabajo de “Depósito y Gestión de datos en Acceso Abierto” del proyecto RECOLECTA. La conservación y reutilización de los datos científicos en España. Informe del grupo de trabajo de buenas prácticas. Madrid: Fundación Española para la Ciencia y la Tecnología, 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.
- Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación (LCTI). BOE. 02/06/2011; 131. Actualización de 10/09/2015. [https://www.boe.es/buscar/pdf/2011/BOE-A-2011-9617-consolidado.pdf].
- Ley 21/2014, de 4 de noviembre, por la que se modifica el texto refundido de la Ley de Propiedad Intelectual, aprobado por Real Decreto Legislativo 1/1996, de 12 de abril, y la Ley 1/2000, de 7 de enero, de Enjuiciamiento Civil: [https://www.boe.es/diario\\_boe/txt.php?id=BOE-A-2014\\_11404](https://www.boe.es/diario_boe/txt.php?id=BOE-A-2014_11404)

- Ley Orgánica 15/1999 de 13 de diciembre de Protección de Datos de Carácter Personal, (LOPD). BOE. 14/12/1999; 298. Actualización de 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.mesa.nhlbi.org/PublicDocs/MESA\\_DeidentifiedDataDistribution\\_PolicyStatement\\_04122016.pdf](https://www.mesa.nhlbi.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 randomized 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](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](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](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, versión 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](http://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.