Aspectos de seguridad en el manejo de datos sensibles

IMPaCT

Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología

Programa	IMPaCT: Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología			
Nombre Proyecto	IMPaCT-Data: Programa de Ciencia de Datos de IMPaCT			
Expediente	IMP/00019			
Duración	Enero 2021 – Diciembre 2023			
Página web	impact-data.bsc.es			
Paquete Trabajo	WP6 – Casos de uso para la aplicación de la metodología y la evaluación de su Calidad			
Tarea	T6.2. Evaluación de aspectos de seguridad en el manejo de datos sensibles, como procesos de pseudonimización, accesibilidad, trazabilidad y privacidad de los datos.			
Entregable	E6.4. Aspectos de seguridad en el manejo de datos sensibles			
Versión	1.1.1			
Fecha Entrega	30/06/2022		***Fecha Aprobación***	17/05/2023
Responsable	SAS-HUVR			
Nivel Diseminación	X	PU	Público	
		CO-IMP	Confidencial, sólo participantes de los pilares de IMPaCT, incluyendo la comisión de evaluación de IMPaCT.	
		CO-DATA	Confidencial, sólo participantes de IMPaCT-Data, incluyendo la comisión de evaluación de IMPaCT.	

Autores		
Organización	**Nombre**	**Rol**
SAS-HUVR	Carlos Luis Parra Calderón	Coordinador/Autor
SAS-HUVR	Sara González García	Autor
FJD	Carmen Ayuso	Autor (IMPaCT-Genómica)
FISABIO-UMIB	Silvia Nadal	Autor
FISABIO-UMIB	María de la Iglesia	Autor
HCB	Santiago Frid	Autor
HCB	Guillem Bracons	Autor
IDIAPJGOL	María Aragón	Autor
IDIBELL-ICO	David Cordero	Autor
IdiPAZ	Carlos Rodríguez	Autor
IdiPAZ	Ángela del Pozo	Autor
IIS La Fe-GIBI230	Pedro Mallol	Autor
ISCIII-UITeS	Adolfo Muñoz	Autor
UMA	Juan Antonio García	Autor
UPV/EHU	Guillermo Lazcoz	Autor (IMPaCT-Genómica)
UPV/EHU	María Pilar Nicolás	Autor (IMPaCT-Genómica)
UAM	Fernando Rodríguez Artalejo	Autor (IMPaCT-Cohorte)
ISCIII	Marina Pollán	Autor (IMPaCT-Cohorte)
VHIR	Berta Miró	Autor
Universitat de Valencia	Ricard Martínez Martínez	Revisor
BSC	Martin Krallinger	Revisor

Historial de versiones			
Nro.	**Fecha**	**Descripción**	**Autor**
V 0.1	26/01/2022	Borrador Índice	CLPC (SAS-HUVR) SGG (SAS-HUVR)
V 0.2	05/05/2022	Índice para compartir con revisores	CLPC (SAS-HUVR) SGG (SAS-HUVR)
V 0.3	20/05/2022	Índice revisado	CLPC (SAS-HUVR) SGG (SAS-HUVR) R. Martínez (UVA) M. Krallinger (BSC)
V0.4	09/06/2022	Aportación de Contenido	S. Frid (HCB) G. Bracons (HCB) S. Nadal (FISABIO) M. de la Iglesia (FISABIO) M. Aragón (IDIAPJGOL) C. Rodríguez (IdiPAZ)

			Á. del Pozo (IdiPAZ) P. Mallol (IIS La Fe) B. Miró (VHIR) D. Cordero (IDIBELL) Adolfo Muñoz (ISCIII) Juan Antonio García (UMA)
V0.5	24/06/2022	Revisión previa al 1º envío a revisores formales, estando pendiente de contribuciones de otros programas de IMPaCT.	CLPC (SAS-HUVR) SGG (SAS-HUVR)
V0.6	30/06/2022	Aportaciones IMPaCT-Genómica	G. Lazcoz (UPV/EHU) M.P. Nicolás (UPV/EHU) C. Ayuso (FJD)
V0.7	03/07/2022	Aportaciones IMPaCT-Cohorte	F. Rodríguez (UAM) M. Pollán (ISCIII)
V0.8	25/07/2022	Actualización de contenido revisión Ricard Martínez	CLPC (SAS-HUVR) SGG (SAS-HUVR)
V1.0	02/09/2022	Actualización de contenido revisión Martin Krallinger	CLPC (SAS-HUVR) SGG (SAS-HUVR)
V1.1	17/05/2023	Cambio visibilidad a público y aprobado	Comité Dirección
V1.1.1	14/06/2023	Cambio de formato para publicar en la Web de IMPaCT	David Velasco (ISCIII)

Contenido

Tablas

Figuras

Resumen Ejecutivo

El objetivo de este entregable es disponer de un panorama extenso de los aspectos de seguridad que habrá que tener en cuenta en IMPaCT-Data respecto a los datos sensibles, teniendo en cuenta la normativa vigente, las responsabilidades del tratamiento, los riesgos que conlleva, el diseño de la infraestructura definida, y las necesidades de confiabilidad ante los proveedores de datos sobre todo del Sistema Nacional de Salud, exponiendo en detalle las principales técnicas de anonimización y pseudonimización utilizadas en datos clínicos, imagen médica y datos genómicos. Para la realización del mismo se ha tenido en cuenta el marco legal y los requisitos de cumplimiento del Esquema Nacional de Seguridad de España.

Introducción

Audiencia

Este entregable es de utilidad a todos los miembros y usuarios potenciales de IMPaCT-Data. Especialmente relevante para los miembros de los paquetes de trabajo de datos clínicos, de imagen médica y de información genómica, ya que deberán conocer la normativa vigente para su tratamiento en base a la RGPD y LOPDGDD. Por otro lado, para los miembros de los paquetes de trabajo encargados del diseño e implementación de la infraestructura, ya que la seguridad de la información debe asegurarse en todo el flujo del tratamiento de dichos datos sensibles.

Ámbito

Este entregable consiste en una descripción de los aspectos a tener en cuenta para cumplir con la seguridad de los datos sensibles a utilizar en los 3 programas canónicos de IMPaCT, teniendo en cuenta la tipología de estos datos y los recursos computacionales y de comunicaciones en implementación en el programa IMPaCT-Data.

No entra en el ámbito de este entregable la realización de un análisis de riesgos o de una Evaluación del Impacto en la Protección de Datos, así como tampoco la realización de un plan de implementación del Esquema Nacional de Seguridad en el ámbito de la infraestructura de IMPaCT-Data.

Relación con otros Entregables

Este entregable guarda relación a su vez con los siguientes entregables: E2.2 que define el diseño Inicial de la Arquitectura de la Infraestructura, E3.1. Requisitos de un Nodo Local EGA. Requisitos técnicos, humanos y acuerdos legales para la puesta en marcha de un nodo de local EGA, E4.1. Normas Internacionales de Anotación de Información de HCE. Revisión de las normas internacionales para la anotación de la información extraída como mecanismo para minimizar la generación de nuevos estándares de interoperabilidad, E4.4. Normas Internacionales de Anotación de Información de Imagen Médica. Revisión de las normas internacionales para la anotación de la información extraída como mecanismo para minimizar la generación de nuevos estándares de interoperabilidad y E5.1. Técnicas de Integración de Datos Biomédicos. Informe con las aproximaciones para la integración de datos genómicos, de imagen médica y médicos estructurados considerando las experiencias existentes dentro del programa de Ciencia de Datos y de IMPaCT

Estructura Entregable

El entregable está divido en 5 secciones, integradas a su vez por diversos subapartados.

1. Contexto: en esta primera sección se identifican los aspectos relevantes de la seguridad para el tratamiento de datos clínicos, imagen médica y datos genómicos.
2. Activos de soporte: en esta segunda sección se describen los componentes identificados, que integrarán el primer demostrador de la infraestructura de IMPaCT-Data.
3. Aspectos fundamentales: sección dedicada a identificar y detallar los aspectos relacionados con la normativa vigente para el tratamiento de Datos Personales, así como para su uso en investigación biomédica, contemplando los distintos escenarios que se presentan en IMPaCT-Data.
4. Riesgos: en esta cuarta sección se profundiza en los riesgos que pueden presentarse en el Tratamiento de Datos Personales de acorde al RGPD identificando las posibles medidas para la mitigación de estos.
5. Aspectos relacionados con las medidas de aplicación: sección dedicada al análisis de las distintas técnicas de anonimización y pseudonimización utilizadas en datos clínicos, imagen médica y datos genómicos.

Y por último se exponen las conclusiones extraídas tras la realización de dicho entregable.

1 Contexto

1.1 Descripción de tratamiento de datos sensibles

En esta sección se incluyen los aspectos relevantes a los 3 tipos de datos que se incluyen en IMPaCT-Data (datos clínicos, imagen médica y datos genómicos) y los casos de uso del proyecto (cáncer, enfermedades raras e IMPaCT-Cohorte).

1.1.1 Aspectos relevantes de la seguridad del uso secundario de datos clínicos

Para investigar a partir de datos clínicos procedentes de las Historias Clínicas Electrónica (HCE) es necesario asegurar la privacidad de los pacientes, pudiendo optar por una anonimización total donde se eliminen todas las referencias demográficas o por una pseudonimización donde se mantienen algunos de los datos demográficos (sexo, fecha de nacimiento, y lugar de residencia) con una precisión seleccionada por los usuarios (cumplimiento del k-anonimato), en caso de que dicha información fuese relevante para el estudio. La aplicación de estándares puede ayudar a realizar estas tareas, es el caso de la herramienta propuesta por Somolinos R et al. (1) la cual realiza el proceso de anonimización manteniendo tres cuasi identificadores basándose en la norma ISO/EN 13606 y utilizando sus características favorables para la anonimización.

1.1.2 Aspectos relevantes de la seguridad del uso secundario de imagen médica

Para investigar con datos de imagen médica es necesario asegurar que no aparecen datos sensibles en ella que puedan identificar al paciente si no son necesarios para la propia investigación. Estos datos pueden aparecer en forma de texto como datos asociados, por lo que se debería eliminar de la imagen. También es necesario eliminar de las imágenes aquellos elementos gráficos identificativos, por ejemplo, las características faciales como ojos o nariz, siempre que estas no sean relevantes para la investigación en las que se van a utilizar, así como otros identificadores indirectos que pueda contener la imagen al margen del objeto de estudio, como por ejemplo malformaciones singulares o existencia de prótesis, cicatrices o tatuajes. No se debe olvidar tampoco anonimizar los metadatos DICOM que se encuentran en las imágenes.

1.1.3 Aspectos relevantes de la seguridad del uso secundario de datos genómicos

IMPaCT-Genómica se estructura en dos circuitos diferenciados validados favorablemente por el CEI-ISCIII.

Por un lado, el circuito diagnóstico-asistencial permite el acceso a la secuenciación genómica una vez agotado sin éxito el proceso asistencial estándar, con el fin de descubrir alteraciones en el genoma y avanzar en el diagnóstico. En este circuito participan los centros clínico-asistenciales, los centros de secuenciación y CIBER, y se utilizan las muestras extraídas, los datos clínicos y los datos genómicos generados. La utilización de datos y muestras en este circuito se rige por la normativa aplicable en el ámbito asistencial. El paciente firma un CI en el que se le informa sobre los procesos de extracción y utilización de muestras y datos pseudonimizados con fines de diagnóstico. Se establece un marco contractual común (Material Transfer Agreements and Data Transfer Agreements MTA/DTA) entre los centros clínico-asistenciales, los centros de secuenciación y CIBER para la realización de la secuenciación genómica y del posterior análisis con fines exclusivamente diagnósticos de los resultados obtenidos de la secuenciación.

Por otro lado, el circuito de investigación permite el uso secundario de los datos generados en el ámbito asistencial para su uso con fines de investigación científica. Este circuito tiene por objetivo impulsar la investigación a través de la compartición de los datos generados en su estudio genómico, integrados en un repositorio, y así fomentar la innovación orientada a la implementación de la Medicina de Precisión como instrumento que contribuye a la sostenibilidad y eficiencia del SNS. Las muestras biológicas no son incluidas en el circuito de investigación. La cesión y uso de los datos pseudonimizados en este circuito se rige por la normativa aplicable en el ámbito de la investigación biomédica. El paciente firma un CI en el que se le informa sobre las condiciones de su participación en el proyecto de investigación. Se establece un marco contractual común (MTA/DTA) entre los centros clínico-asistenciales, los centros de secuenciación y CIBER para la cesión por parte de los centros clínicos a CIBER de los datos generados con objetivos asistenciales para su almacenamiento y custodia en una base de datos y su posterior uso secundario con fines de investigación biomédica. Una política de acceso y uso definirá los términos en los que dichos datos serán compartidos con fines de investigación para la realización de los objetivos del proyecto.

1.1.3.1 Aspectos específicos de datos genómicos relevantes teniendo en cuenta la historia familiar

Las pruebas genómicas pueden generar hallazgos adicionales, inesperados o incidentales, predecir la salud futura y diagnosticar problemas actuales tanto de la persona implicada como de la familia, por ejemplo, paternidad incorrecta, detección de enfermedades hereditarias no diagnosticadas en la familia, etc. Por ello es necesario especificar el correcto tratamiento de estos posibles hallazgos en el consentimiento que acepta el paciente, recogiendo los siguientes supuestos:

- La interpretación de los resultados genómicos puede actualizarse en el futuro y puede necesitar una reevaluación periódica.
- Las muestras de ADN se almacenan de forma rutinaria y permanente (en contraste con la mayoría de las otras muestras biológicas que se desechan una vez finalizada la prueba).
- Las muestras de ADN almacenadas de un miembro de la familia se utilizan habitualmente como control de calidad para pruebas clínicas en otros miembros de la familia.
- A veces es necesario compartir estos datos más ampliamente en el sistema sanitario u ocasionalmente fuera de él para recopilar evidencia para informar la interpretación variante o evaluación de la historia familiar; la anonimización absoluta puede no ser posible y podría comprometer la utilidad de intercambio.

En el programa de IMPaCT-Genómica varios WPs se encuentran trabajando en esta cuestión para el circuito diagnóstico y aún por desarrollar la política de hallazgos incidentales en el circuito de investigación.

1.1.4 Descripción y finalidad del tratamiento de datos de cáncer

Recientemente, un gran aumento en la producción de datos requiere de nuevas estrategias para compartir tanto los datos como las estrategias de análisis y flujos de trabajo. Los datos tratados en cáncer se encuentran en formatos muy diversos (2):

- Datos de anatomía patológica (Datos clínicos, datos de bioespecimen, reporte de patología).
- HCE: datos hospitalarios, atención primaria, ambulatorio, hospital de día datos de terapia (radioterapias, quimioterapia), datos diagnósticos (radiológicos), mortalidad, incidencia.
- Microarray, High Throughput Sequencing (HTS), NanoString, Serial Analysis of Gene Expression (SAGE), Reverse Transcriptase – Polymerase Chain Reaction (RT-PCR), Microsatélites, Metilación.
- Imagen: Imagen diagnóstica, imagen de tejido, imagen radiológica.
- Metadatos: hay de dos tipos, metadatos de las muestras que serían el peso de la biomuestra, tipo de tejido, método de extracción del material genético, y metadatos de la plataforma utilizada, así como diseño del array en relación al tipo de sondas utilizadas, número de copias, información específica de la plataforma de secuenciación.

Así mismo, los datos también se encuentran en formatos muy diversos, ejemplos de extensiones de archivo serían las siguientes que corresponderían a distintos tipos de datos: TXT, CSV, TSV, XML, SOFT, CEL, FASTQ, BAM, BED, VCF, MAF, SVS, DCM, TIFF (2). Una vez recogidos, los datos tratados en cáncer pasan por unos estadios de proceso antes de que puedan ser utilizados. El primero es el preprocesado de los datos, donde se hace un control de calidad inicial, alineamiento y anotación, por ejemplo, en el caso de datos genómicos. El paso siguiente, muy importante, es el de o bien anonimización, o más comúnmente de pseudononimización (ejemplo Figura 1)., requisito para el posterior almacenaje, análisis y compartición de datos entre unidades distintas A continuación, se procede a realizar un agregado de datos para datos asociados (3).

Una vez preprocesados, los datos están listos para el manejo, manipulación e integración de datos, en este punto es cuando se realizan posibles transferencias de datos por ejemplo a través de FTP (utilizados en repositorios abiertos como NCBI, GEO (4), SFTP, o privados como Aspera Connect (EGA), o portales específicos del repositorio en cuestión (GDC data transfer tool (5)) entre otros, todos tienen en común que funcionan con usuario y contraseña. En estos repositorios tanto internos del centro como externos se realiza el almacenaje de datos tanto raw como intermedios.

Para facilitar el manejo de datos, manipulación e integración se utilizan herramientas para interoperabilidad, flujos de trabajo y containerización. Aunque distintos registros utilizan herramientas y anotaciones distintas (6), estos incluyen guías y protocolos de trabajo comunes, ontologías y diccionarios, lenguajes y sistemas de intercambio entre otras. Ejemplos de estándares (como por ejemplo MIAME para datos de microarray) que se utilizan para registrar y reportar datos ómicos y con la finalidad de facilitar el acceso a los datos y la interpretación (5). Con fines parecidos, ejemplos de ontologías y diccionarios utilizados son: International classification of Diseases for Oncology (7), Human disease ontology (DOID), HPO, NCI Thesaurus (NCIt), Bioscientific data analysis ontology (EDAM), Microarray and gene expression data ontology (MO), ABAP CDS, entre otras. Y también se utilizan lenguajes y sistemas de intercambio como OMOP, HL7 o FHIR.

En relación a la periodicidad de recogida de los datos tratados en cáncer, esta es variable en diferentes estudios y registros, en algunos casos se realiza una sola recogida, o una recogida pre y post tratamiento, y también una recogida periódica cada cierto número de estadios/ tratamientos/ años. En la estrategia en cáncer del sistema nacional de salud español, por el momento no se ha fijado una periodicidad de recolección (8). Actualmente, el tiempo de conservación es permanente en repositorios especializados y open Access (datos genómicos TCGA, GEO) y también en los registros de cáncer (de momento la recogida en el registro de cáncer Europeo es de los años 1950 en adelante, ENCR (6). Por lo que a los datos de investigación se refiere, sobre todo a datos intermedios utilizados en informes y publicaciones, se recomienda conservar los datos un mínimo de 5 años después de ser publicados (9). En centros de investigación, la conservación oscilaría entre 6 meses a 2 años para datos de análisis y datos intermedios en almacenaje local (centro específico), los tiempos dependiendo del volumen de datos que se manejan y de la capacidad de almacenaje de cada centro. Todos los repositorios donde se almacenan datos tienden al cumplimiento de los principios FAIR y ofrecen un identificador único que permite localizar los datos, así como licencias de uso, regulación de acceso cuando sea necesario, y trazabilidad de usos.

Flujo de los datos (Adaptación de Crossfield et al., 2022 (10)):

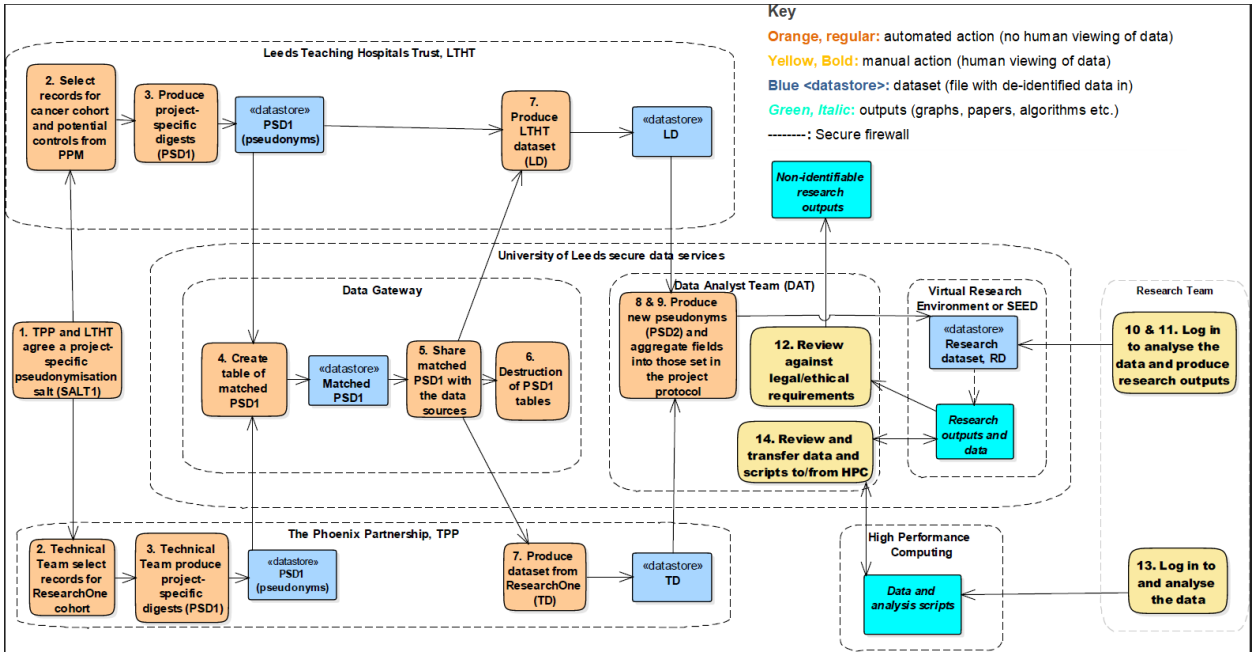

Figura 1.- Diagrama de las partes implicadas en el estudio de caso y de las acciones realizadas durante el proceso de flujo de datos, vinculación y acceso, tal y como se define mediante el protocolo de vinculación y anonimización de datos

1.1.5 Descripción y finalidad del tratamiento de datos de enfermedades raras

Los datos se almacenarán durante el periodo de duración de la infraestructura creada por el proyecto IMPaCT, cuya utilidad y objetivos es previsible que se extiendan durante un largo plazo. Los participantes podrán revocar su consentimiento en cualquier momento, los datos se mantendrán en los proyectos de investigación en que se hayan utilizado hasta ese momento, pero quedarán anonimizados, sin vínculo alguno con su identidad, salvo obligaciones legales o razones científico-técnicas en otro sentido.

Los datos clínicos y genómicos tienen su origen en el circuito diagnóstico-asistencial definido anteriormente.

- Información de Datos clínicos recogidos en la plataforma IMPaCT- Genómica
  - Datos de contacto: detalles del gestor que remite el caso, facultativo de referencia, datos del subproyecto IMPaCT al que se asocia (se adjunta el modelo de consentimiento informado del paciente).
  - Datos del paciente: información demográfica del paciente, que se mantendrá anonimizada.
  - Motivo de la consulta: característica clínica primaria, edad de inicio, clasificación de la patología.
  - Curso de la patología y otros síntomas.

- Antecedentes pre/perinatales: gestaciones previas, TPAL, enfermedades maternas, exposición a teratógenos, hallazgos ecográficos, parto, diagnóstico prenatal.
- Desarrollo psicomotor: sostén cefálico, volteo en decúbito, sedestación, bipedestación, pinza con dedos, deambulación independiente, control esfínteres lenguaje.
- Antecedentes familiares: posible patrón de herencia, árbol genealógico en los casos que proceda.
- Somatometría: edad, altura, percentil altura, peso, percentil peso, índice de masa corporal, perímetro cefálico, percentil perímetro cefálico.
- Disformología.
- Exploración otros órganos
- Neurocognitivo/Conductual: datos de psicometría (escalas de conducta, socialización, cognitivo (Batelle, Vineland, CBCL, K-BIT, WISC-IV, etc.).
- Bioquímica y otras analíticas
- Pruebas complementarias: EEG, Ecografías, RMN, etc.
- Cáncer somático: información relativa a casos de cáncer somático (caracterización del tumor, tratamientos, supervivencia, etc.)
- Datos genéticos: pruebas realizadas y descripción de resultados (cariotipo, array, panel, exoma, etc.)
- Pruebas previstas
- Tratamiento e intervenciones
- Historia de derivación
- Disponibilidad de muestra
- Fenotipo HPO: codificación de los síntomas mediante terminología Human Phenotype Ontology.
- En todos los apartados se podrá introducir un texto libre y adjuntar documentación anonimizada.

- Datos Genómicos
  - Datos derivados de la secuenciación genómica, RNAseq y genotipado (lecturas de secuencias, listado de variantes, niveles expresión, genotipos, métricas de secuenciación y/o genotipado, etc.).

1.1.6 Descripción y finalidad del tratamiento de datos de IMPaCT Medicina Predictiva

El programa de Medicina Predictiva (IMPaCT-Cohorte) tiene por objetivo el diseño y creación de una cohorte de base poblacional representativa de la población residente en España, su variabilidad étnica, diversidad geográfica y ambiental, con la participación de todas las CC.AA. y seguimiento prospectivo. En este caso, la información no procederá de las HCEs, sino que corresponderá a una recolección de datos específica, mediante una serie de cuadernos de recogida de datos (REDCAP), que las instituciones participantes comenzarán a poblar en los próximos meses, tratándose así de un estudio longitudinal. Por ello debe existir un sistema de pseudonimización para la investigación o para el análisis, pero en el origen debe mantenerse la identidad de los individuos para completar dichos cuadernos de manera

continua en el tiempo. En este caso el dato de filiación permitirá mantener la traza del sistema de pseudonimización

Los datos recogidos se corresponderán con la siguiente selección de variables, sujeta a posibles modificaciones:

- Información procedente de cuestionarios básicos:
  - Datos de filiación
  - Variables demográficas
  - Nivel socioeconómico, clase social
  - Historial médico personal
  - Historial médico familiar
  - Medicación de los últimos 7 días
  - Participación en programas de detección precoz de enfermedades
  - Salud sexual y reproductiva en hombres y mujeres
  - Consumo de tabaco
  - Exposición al humo ambiental del tabaco
  - Consumo de bebidas alcohólicas, patrones de consumo de alcohol
  - Consumo de otras drogas
  - Adiciones sin sustancias
  - Calidad de vida relacionada con la salud
  - Discapacidades en actividades instrumentales de la vida diaria (60 + años)
  - Discapacidad en actividades básicas de la vida diaria (60 + años)
- Información procedente de cuestionarios específicos
  - Factores neurológicos y psiquiátricos (síntomas de depresión, trastorno de ansiedad, sueño)
  - Factores psicosociales: personalidad, estrés crónico, estrés laboral, red social
  - Conflictos en el trabajo y en la familia
  - Inseguridad laboral
  - Situación inmunitaria e infecciones pasadas
  - Dolor crónico
  - Salud oral
  - Funcionamiento de los órganos de los sentidos: vista, oído, olfato y gusto
  - Actividad física, sedentarismo, condición física
  - Alimentación: consumo de alimentos, preferencias y hábitos alimentarios
  - Factores del medio ambiente construido: características de la vivienda, características del barrio, barreras a los estilos de vida saludables
  - Ocupación: exposiciones especiales en el trabajo
- Información procedente de exámenes físicos:
  - Sistema cardiovascular: presión arterial y frecuencia cardiaca, electrocardiografía, ecocardiografía 3D, índice tobillo-brazo, medidas de rigidez arterial (onda de pulso)
  - Diabetes: prueba de tolerancia oral de glucosa,

- Función cognitiva: memoria (semántica, episódica, de trabajo), atención/función ejecutiva, coordinación motriz, razonamiento numérico (inteligencia fluida), vocabulario.
- Función pulmonar: espirometría, óxido nítrico en el aire exhalado.
- Sistema musculoesquelético: examen médico para identificar artrosis y artritis reumatoide, cadera.
- Salud bucal: caries, enfermedad periodontal, número de dientes, trastornos de la articulación temporomandibular
- Exploración oftalmológica: prueba de agudez visual
- Audición: comprensión verbal mediante prueba de los tres dígitos, audiometría tonal
- Actividad y función física: acelerometría de 7 días, ergometría submáxima en bicicleta, fuerza de agarre de la mano, velocidad de la marcha, Short Physical Performance Battery (mayores de 65 años)
- Antropometría: peso, talla, circunferencia de cintura y cadera, bioimpedancia eléctrica, DEXA.

1.2 Responsabilidades vinculadas al tratamiento

De acuerdo con el *Articulo 11. La seguridad como función diferenciada* del Esquema Nacional de Seguridad (ENS) 11, «En los sistemas de información se diferenciará el **responsable de la información**, el **responsable del servicio**, el **responsable de la seguridad** y el **responsable del sistema**». Las funciones de cada uno de ellos, indicadas en el *Artículo 13. Organización e implantación del proceso de seguridad* del ENS son las siguientes:

- Responsable de la información: determinará los requisitos de la información tratada.
- Responsable del servicio: determinará los requisitos de los servicios prestados.
- Responsable de la seguridad: determinará las decisiones para satisfacer los requisitos de seguridad de la información y de los servicios, supervisará la implantación de las medidas necesarias para garantizar que se satisfacen los requisitos y reportará sobre estas cuestiones.
- Responsable del sistema: se encargará de desarrollar la forma concreta de implementar la seguridad en el sistema y de la supervisión de la operación diaria del mismo, pudiendo delegar en administradores u operadores bajo su responsabilidad.

Asimismo, según indica la Guía CCN-STIC 801 del ENS - Responsabilidades y Funciones, así como el RGPD (art. 4.7, 4.8,19) y la LOPDGDD (Título V y art. 34 a 37), al tratar datos de carácter personal es necesario la presencia de un:

- Responsable del tratamiento: persona física o jurídica, autoridad pública, servicio u otro organismo que determine los fines y medios del tratamiento.

- Encargado del tratamiento: persona física o jurídica, autoridad pública, servicio u otro organismo que trate datos personales por cuenta del Responsable del Tratamiento.

La diferencia entre el responsable del tratamiento y el encargado del tratamiento radica en que mientras el primero es quién decide el uso y la finalidad de los datos personales, el segundo debe seguir sus instrucciones respecto a dicho uso y finalidad.

- Delegado de Protección de Datos (DPD): persona encargada de informar al responsable o al encargado del tratamiento sobre sus obligaciones legales en materia de protección de datos.

El trabajo de la Línea Estratégica Transversal de Ética e Integridad Científica definirá la limitación de responsabilidades asociadas a los distintos modelos de participación de la infraestructura (nodo central, nodos proveedores de computación, nodos gestores de datos, nodos proveedores de datos sensibles, detallados a continuación en el apartado *2.1 Descripción de los activos de soporte identificados*).

1.2.1 Responsabilidades comunes

La Guía CCN-STIC 801 del ENS diferencia tres grandes bloques de responsabilidad (Figura 2):

- Responsabilidad legal y especificación de necesidades o requisitos, que corresponde a la Dirección de la entidad y a los responsables del tratamiento, de la información y del servicio.
- La supervisión, que corresponde al Responsable de Seguridad y al Delegado de Protección de Datos, en sus respectivos ámbitos.
- La operación del sistema de información, que corresponde al Responsable/s del Sistema/s

Figura 2.- Niveles de la estructura de seguridad - CCN-STIC-801 seguridad - CCN-STIC-801

1.2.2 Responsabilidades específicas de datos genómicos

En los circuitos diagnóstico-asistencial y de investigación en IMPaCT-Genómica el responsable del tratamiento de los datos varía. En el circuito diagnóstico asistencial la responsabilidad del tratamiento recae sobre cada uno de los centros clínicos en los que se presta la relación asistencial a los pacientes, mientras que los centros de secuenciación y CIBER actúan como encargados del tratamiento. El contrato MTA/DTA que realizan los centros clínicos, los centros de secuenciación y CIBER incluye como anexo un contrato de encargado del tratamiento que define las condiciones bajo las que se realiza el tratamiento de datos personales.

En el circuito de investigación el responsable del tratamiento es el CIBER. La cesión de los datos generados en el circuito asistencial de un responsable (centros clínicos) a otro (CIBER) se regula por el contrato MTA/DTA. Las condiciones de uso y acceso a estos datos con fines de investigación científica se establecerán conforme a la política de acceso y uso IMPACT-Genómica. CIBER realizará una Evaluación de Impacto del Tratamiento de Datos (EIPD) conforme al artículo 35 RGPD. Se suscribirán los contratos de encargado del tratamiento con las instituciones que sean necesarias para llevar a cabo los objetivos del proyecto, entre otros, con los centros de secuenciación que dispondrán inicialmente de los datos genómicos generados en el circuito asistencial.

1.2.3 Responsabilidades específicas en cáncer

Para los casos de cáncer hereditario, en el circuito diagnóstico-asistencial se realizará una preselección de casos a través de una encuesta REDCAP de CIBER. La encuesta tiene por objetivo determinar si procede su inclusión en el proyecto conforme a criterios clínicos. En este circuito asistencial el responsable del tratamiento de los datos es el centro clínico pertinente que firmará, a su vez, el MTA/DTA con los centros de secuenciación y CIBER. Este contrato incluye a CIBER como encargado del tratamiento dentro del circuito asistencial, con lo cual está habilitado legalmente para el tratamiento de datos personales por cuenta del responsable para la realización de los objetivos que define el contrato. El tratamiento de datos personales se realiza de forma pseudonimizada. Los datos solicitados en la encuesta son adecuados a la finalidad del tratamiento y cumplen con el principio de minimización. Se prevé la eliminación de los datos por el encargado del tratamiento desde el momento en que se determine que no son aptos.

1.3 Estándares vinculados al tratamiento

Actualmente el estándar internacional más relevante vinculado al tratamiento de datos es la ISO/IEC 27701:2019 (https://www.iso.org/standard/71670.html). Esta norma internacional, conocida anteriormente como la ISO/IEC 27552, nace para garantizar el correcto cumplimiento de la RGPD con cobertura en nuestro caso para la LOPDGDD, puestas en marcha en 2018 para la protección de los datos personales y privados, por parte de los encargados de tratamiento de datos y de los responsables de éstos en la Unión Europea.

La certificación necesaria es una extensión de la ISO/IEC 27701 y de la ISO/IEC 27002, que de forma separada aseguran respectivamente la seguridad de los sistemas de gestión y de la información.

Desde el programa IMPaCT-Data se valorará la certificación de los sistemas puestos en marcha y del almacenamiento de los datos incluidos, aunque esto no delimita el cumplimiento del ENS.

2 Activos de soporte

2.1 Descripción de los activos de soporte identificados

En este apartado describimos los activos identificados para el primer demostrador de la infraestructura de IMPaCT-Data, previsto en el entregable E2.2 Diseño Inicial de la Arquitectura de la Infraestructura".

En este sentido, se ha optado por la utilización de componentes ya funcionales en infraestructuras similares, en particular los utilizados en el proyecto Galaxy Europa:

- Autentificación basada en OIDC, gestionada por KeyCloak.
- Servidor central e interfaz Galaxy incluyendo las herramientas necesarias para la realización del flujo de trabajo de demostración
- Infraestructura federación de repositorios de datos sensibles basada en tecnología EGA
- Infraestructura de computación distribuida basada en Pulsar (Pulsar Network)
- Infraestructura de compartición de datos basada en CMVFS
- Entornos Virtuales de Investigación basados en OpenVRE
- En esta fase de la demostración cada centro usará el modelo de almacenamiento local que considere más adecuado.

La descripción de estos componentes se desarrolla en dicho E2.2., donde se identifica igualmente los 4 modelos de participación a nivel de nodo de la infraestructura: (i) nodo central, (ii) nodos proveedores de computación, (iii) nodos gestores de datos, (iv) nodos proveedores de datos sensibles.

Se empleará la tecnología local EGA implementada en los nodos proveedores de datos para la gestión de los datos sensibles en la infraestructura computacional, la cual proporciona un protocolo de ingestión de datos, metadatos que describen los datos alojados mediante su API, un módulo de gestión de credenciales de acceso sincronizado con el módulo central de gestión de acceso y un módulo de distribución que proporcionara datos encriptados a los nodos computacionales. En el caso de transferencia de datos entre un nodo proveedor de datos y un nodo computacional o al nodo central de la infraestructura, se establecerá una red privada virtual (VPN) entre ambos centros para la transmisión de datos con las garantías de privacidad requeridas, y mediante la realización de los acuerdos legales necesarios.

2.2 Percepción del riesgo de accesos no deseados para cada activo de soporte identificado

Para facilitar una aproximación a la percepción del riesgo de accesos no deseados para cada activo de soporte se propone una serie de cuestiones que AEINSE 10/21 “Guía de buenas prácticas de ciberseguridad en proyectos de seguridad física” de diciembre de 2021, para facilitar una estimación de esta:

- ¿Qué conexiones y datos entran y salen de los sistemas electrónicos de seguridad?
- ¿Hay algún problema conocido con los sistemas?
- ¿Cuáles son los proyectos en curso o programados?
- ¿Qué dependencias tiene la ubicación?
- ¿Hay resúmenes y diagramas detallados de los sistemas y la red?
- ¿Está protegida toda la documentación y se siguen procedimientos de gestión de cambios?

3 Aspectos fundamentales

3.1 Proporcionalidad y necesidad

3.1.1 Evidencias propuestas para la legitimidad del tratamiento

El considerando 35 del Reglamento General de Protección de Datos, en adelante RGPD, (RGPD 2016/679) define los datos relativos a la salud de una persona física como «todos los datos relativos al estado de salud del interesado que dan información sobre su estado de salud física o mental pasado, presente o futuro»; siendo el dato relativo a la salud «todo número, símbolo o dato asignado a una persona física que la identifique de manera unívoca a efectos sanitarios; la información obtenida de pruebas o exámenes de una parte del cuerpo o de una sustancia corporal, incluida la procedente de datos genéticos y muestras biológicas, y cualquier información relativa, a título de ejemplo, a una enfermedad, una discapacidad, el riesgo de padecer enfermedades, el historial médico, el tratamiento clínico o el estado fisiológico o biomédico del interesado, independientemente de su fuente, por ejemplo un médico u otro profesional sanitario, un hospital, un dispositivo médico, o una prueba diagnóstica in vitro».

El art. 9 RGPD, regula el tratamiento de los datos de salud, de los datos genéticos y de los datos biométricos dirigidos a identificar de manera unívoca a una persona física bajo la denominación de **categorías especiales de datos personales**. Conforme al art. 9.1., el tratamiento de estos está prohibido como regla general, salvo que se den algunas de las circunstancias contempladas en el apartado segundo del precepto art. 9.2 RGPD. En nuestro caso, el tratamiento legitimo se realizará acogiéndose al primer supuesto del art. 9.2.a RGPD, el cual establece que la prohibición del tratamiento de datos personales no será de aplicación

cuando «el interesado de su **consentimiento explícito** para el tratamiento de dichos datos personales con uno o más de los fines especificados, exceptuando cuando el Derecho de la Unión o de los estados miembros establezca que la prohibición del tratamiento de datos personales no pueda ser levantada por el interesado».

Excepciones a este consentimiento las encontramos en el considerando 54 del RGPD «El tratamiento de categorías especiales de datos personales, sin el consentimiento del interesado, puede ser necesario por razones de interés público en el ámbito de la salud pública. Ese tratamiento debe estar sujeto a medidas adecuadas y específicas a fin de proteger los derechos y libertades de las personas físicas». En esta línea, la Ley Orgánica de Protección de Datos, LOPDGDD 3/2018, en su disposición adicional decimoséptima permite a las autoridades sanitarias e instituciones públicas con competencias en vigilancia de la salud pública llevar a cabo estudios científicos sin el consentimiento de los afectados en situaciones de excepcional relevancia y gravedad para la salud pública.

Se destaca, que en el ámbito de IMPaCT-Data se va a trabajar básicamente con conjuntos de datos anonimizados y pseudonimizados, siendo relevante dicha diferenciación para determinar a qué nivel aplica la normativa de protección de datos personales:

- **Conjuntos de datos anonimizados**: el RGPD y la LOPDGDD no son de aplicación para los datos anonimizados, ya que tras el proceso de anonimización dejan de considerarse datos personales. Precepto recogido en el Considerando 26 del RGPD «Por lo tanto los principios de protección de datos no deben aplicarse a la información anónima, es decir información que no guarda relación con una persona física identificada o identificable, ni a los datos convertidos en anónimos de forma que el interesado no sea identificable, o deje de serlo. En consecuencia, el presente Reglamento no afecta al tratamiento de dicha información anónima, inclusive con fines estadísticos o de investigación».
- **Conjunto de datos pseudonimizados**: el RGPD y la LOPDGDD sí son de aplicación para los conjuntos de datos pseudonimizados y para la información adicional vinculada con dicho conjunto de datos, así como el tratamiento que los genera. Por ello, será necesario el consentimiento explícito del paciente y cumplir con los siguientes requisitos recogidos en la disposición adicional decimoséptima de la Ley Orgánica de Protección de Datos, LOPDGDD 3/2018, por la que se admite el tratamiento de datos pseudonimizados:
    - Haber realizado una EIPD (Anexo A: Lista de tipos de tratamientos de datos que requieren evaluación de Impacto relativa a la Protección de Datos de la AEPD).
    - Separar funcionalmente el control sobre los identificadores y garantizar el modelo con compromisos jurídicos de no re-identificación.
    - Adoptar medidas de seguridad (técnicas de análisis de riesgos de re-identificación)

  Asimismo, se cumplirá con lo dispuesto en el art. 16.3 de la Ley 41/2002, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica, el cual determina que el acceso a la historia clínica con fines de investigación, entre otras, obliga a preservar los datos de identificación personal del paciente, separados de los de carácter clínico - asistencial, quedando asegurado el anonimato, salvo que el propio paciente haya dado su consentimiento para no separarlos.

3.1.2 Bases jurídicas del tratamiento de datos en IMPaCT-Data

El tratamiento de datos en IMPaCT-Data está regulado en las siguientes leyes y sus disposiciones de desarrollo:

- 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.
- Directiva (UE) 2019/1024 del Parlamento Europeo y del Consejo, de 20 de junio de 2019, relativa a los datos abiertos y la reutilización de la información del sector público.
- Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales.
- La Ley 14/1986, de 25 de abril, General de Sanidad.
- La Ley 31/1995, de 8 de noviembre, de Prevención de Riesgos Laborales.
- La Ley 41/2002, de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica.
- La Ley 16/2003, de 28 de mayo, de cohesión y calidad del Sistema Nacional de Salud.
- La Ley 44/2003, de 21 de noviembre, de ordenación de las profesiones sanitarias.
- La Ley 14/2007, de 3 de julio, de Investigación biomédica.
- La Ley 33/2011, de 4 de octubre, General de Salud Pública.
- La Ley 20/2015, de 14 de julio, de ordenación, supervisión y solvencia de las entidades aseguradoras y reaseguradoras.
- El texto refundido de la Ley de garantías y uso racional de los 105 medicamentos y productos sanitarios, aprobado por Real Decreto Legislativo 1/2015, de 24 de julio.

3.1.3 Principio de minimización de datos en IMPaCT-Data

Entre los 6 principios de Protección de Datos que define la RGPD se encuentra el de minimización o proporcionalidad de datos, según el cual los datos han de ser «*adecuados, pertinentes y limitados a lo necesario en relación con los fines para los que son tratados*». De esta manera, el acceso y tratamiento a los distintos conjuntos de datos dentro del ámbito del proyecto IMPaCT-Data cumplirá con los siguientes requisitos:

- Sólo serán recogidos los datos personales que se vayan a utilizar, los adecuados, relevantes y mínimos, es decir, los que sean estrictamente necesarios para el tratamiento
- Se justificará la necesidad de los datos sensibles necesarios incluidos.
- Los datos sólo serán recogidos cuando vayan a ser tratados, por tanto, no se recogerán datos para un uso posterior.
- La información personal incluida sólo será utilizada para la finalidad con la que fue recogida inicialmente, no para ningún otro objetivo no especificado a priori.

El cumplimiento de dicho principio asegura que se cumplirán las condiciones de transparencia, respeto de privacidad del paciente y la confidencialidad de la información.

3.2 Controles para proteger los derechos personales de los propietarios de los datos sensibles

3.2.1 Consentimiento explícito

El consentimiento explícito que legitima el tratamiento de categorías especiales de datos personales debe ser libre, específico, informado e inequívoco para garantizar el cumplimiento de la normativa vigente (considerandos 32, 42, 43 y 171 y art. 6.1.a), 7 y 9 RGPD). Asimismo, según la normativa específica en materia de investigación biomédica, para la realización de cualquier actividad de investigación biomédica será necesario obtener previamente su consentimiento expreso y escrito una vez recibida la información adecuada (art. 4.1, 5, 13, 45 y 48.1 Ley 14/2007, de 3 de julio), o bien por cualquier medio que, en caso de incapacidad para prestar el consentimiento por parte del paciente, permita manifestar expresamente la voluntad del interesado; dicha necesidad también está recogida en los principios éticos para la investigación médica que incluya sujetos humanos de la Declaración de Heslinki publicada por la World Medical Association, donde se menciona además el derecho de los sujetos participantes en la investigación a que se salvaguarde su integridad, aplicando toda precaución necesaria para respetar su privacidad y confidencialidad de la información sobre el paciente.
Este consentimiento explícito, donde se recogen las normas éticas y legales de los pacientes participantes del proyecto son validados por los Comités Éticos.

En la Declaración de Taipei (12) de 2017 se enumeran los aspectos sobre los que el sujeto debe ser informado respecto a la recopilación y almacenamiento de sus datos para asegurar la seguridad de su información personal, así como la validez del consentimiento. Estos son:

- La finalidad de la base de datos o biobanco de salud.
- Los riesgos y cargas asociados con la recopilación, el almacenamiento y el uso de datos y materiales.
- La naturaleza de los datos o material a recopilar.
- Los procedimientos para la devolución de resultados, incluidos los hallazgos incidentales.
- Las reglas de acceso a la Base de Datos de Salud o Biobanco.
- Cómo se protege la privacidad.
- Los arreglos de gobernanza.
- Que en caso de que los datos y el material se hagan no identificables, es posible que la persona no pueda saber qué se hace con sus datos/material y que no tendrá la opción de retirar su consentimiento.
- Sus derechos fundamentales y las garantías establecidas en esta Declaración.
- Cuando corresponda, el uso comercial y la distribución de beneficios, las cuestiones de propiedad intelectual y la transferencia de datos o material a otras instituciones o terceros países.

En cuanto a los arreglos de gobernanza, estos deben incluir los siguientes elementos:

- La finalidad de la Base de Datos o Biobanco de Salud.

- La naturaleza de los datos de salud y material biológico que estarán contenidos en la Base de Datos de Salud o Biobanco.
- Arreglos por el período de tiempo durante el cual se almacenarán los datos o el material.
- Arreglos para regulaciones de disposición y destrucción de datos o material.
- Acuerdo sobre cómo se documentarán y rastrearán los datos y el material de acuerdo con el consentimiento de las personas interesadas.
- Acuerdo sobre cómo se tratarán los datos y el material en caso de cambio de propiedad o cierre.
- Acuerdo para obtener el consentimiento apropiado u otra base legal para la recopilación de datos o material.
- Arreglos para la protección de la dignidad, la autonomía, la privacidad y la prevención de la discriminación.
- Criterios y procedimientos relacionados con el acceso y el intercambio de datos de salud o material biológico, incluido el uso sistemático del Acuerdo de Transferencia de Material (MTA) cuando sea necesario.
- La persona o personas que tengan a su cargo el gobierno.
- Las medidas de seguridad para evitar el acceso no autorizado o el intercambio inapropiado.
- Los procedimientos para volver a contactar a los participantes cuando corresponda.
- Los procedimientos para recibir y atender consultas y quejas.

El Comité de Revisión de Ética de la Investigación (ERC) de la Organización Mundial de la Salud (OMS) propone unas plantillas de consentimiento informado con ejemplos y explicaciones en cada sección (13). Estas plantillas incluyen ejemplos de preguntas clave que pueden formularse al final de cada sección, que podrían garantizar la comprensión de la información proporcionada, especialmente si el estudio de investigación es complejo. Estos son solo ejemplos y sugerencias, y los investigadores tendrán que modificar las preguntas según su estudio
En el contenido del Consentimiento Informado se recogen también los derechos ARCO-POL o ARSULIPO (Acceso, Rectificación, Supresión, Limitación, Portabilidad y Oposición al tratamiento). Como se comenta en el apartado 3.2.3, estos son derechos que puede ejercer cualquier persona sobre el tratamiento de sus datos personales.

3.2.2 Excepciones al consentimiento informado

Como se ha comentado previamente, el art. 9.2.a RGPD permite el tratamiento de categorías especiales de datos si el interesado da su consentimiento explícito para ello, sin embargo, nos encontramos con las siguientes excepciones:

- Cuando no sea posible la identificación del paciente, porque sus datos han sido anonimizados previamente, siempre previa autorización del Comité Ético de investigación.
- Cuando los fines de la investigación sean de interés público en el ámbito de la salud pública, con fines de investigación científica – art. 9.2 i) y j) RGPD- o con fines de archivo en interés público, fines de investigación científica o histórica o fines estadísticos –art. 89 RGPD- cuando así lo exija la gestión de los sistemas y servicios de asistencia sanitaria y social, pública y privada, o la ejecución de un contrato de

seguro del que el afectado sea parte (art. 9.2 LOPDGDD), y siempre que se ofrezcan las garantías adecuadas para los derechos y libertades de los pacientes. Dentro de estas garantías destacan todas aquellas medidas técnicas y organizativas necesarias para cumplir con el principio de minimización de los datos personales que permita alcanzar tales fines a través de un tratamiento ulterior sin la identificación del interesado.

- Cuando en casos excepcionales y de interés sanitario general, la autoridad competente, previo informe favorable de la autoridad en materia de protección de datos, haya autorizado la utilización de datos genéticos codificados, siempre asegurando que no puedan relacionarse o asociarse con el sujeto fuente por parte de terceros (art 50. Ley 14/2007, de 3 de julio, de Investigación biomédica).

3.2.3 Soporte a la aplicación derechos ARCOPOL en IMPaCT-Data

Los derechos ARCO (Acceso, Rectificación, Cancelación y Oposición) son aquellos que puede ejercer cualquier persona en cuanto al tratamiento de sus datos personales, protegiendo a los ciudadanos y evitando que su información personal sea incorrectamente tratada. En el caso de IMPaCT-Data, dichos derechos ARCOPOL, son de posible aplicación para aquellos conjuntos de datos pseudonimizados, ya que sobre datos anonimizados no es posible ejercer dichos derechos.

Si bien no existe una ley ARCO específica, el ejercicio de estos derechos se regula en el RGPD y en la LOPDGDD. Aunque en un principio las siglas ARCO se referían al derecho de Acceso, Rectificación, Cancelación y Oposición, con la entrada en vigor del RGPD y la LOPDGDD han sido ligeramente modificados, manteniendo el derecho de Acceso, Rectificación y Oposición, sustituyendo el de Cancelación por el de Supresión y el Derecho al Olvido y añadiendo el derecho a la Limitación del Tratamiento y la Portabilidad. Por ello, la normativa española ha dado en llamarlos derechos ARSULIPO o ARCOPOL:

- **Acceso**: permite que una persona pueda solicitar información al responsable de un fichero sobre si sus datos personales están siendo tratados, con qué finalidad, cuál el origen de esos datos y cuáles son las comunicaciones realizadas o previstas de los mismos.
- **Rectificación**: permite a la persona afectada solicitar la modificación de datos que sean inexactos o incompletos, aportando documentación justificativa de la rectificación solicitada.
- **Cancelación**: permite al afectado solicitar la supresión de los datos que resulten inadecuados o excesivos, aportando documentación justificativa de la rectificación solicitada.
- **Oposición**: derecho a oponerse al tratamiento de sus datos personales. Deberán hacerse constar los motivos fundados y legítimos, relativos a una concreta situación personal de la persona afectada, que justifiquen el ejercicio de este derecho.
- **Portabilidad de datos**: posibilidad de solicitar al responsable del tratamiento que se le faciliten los datos personales en un formato estructurado y claro a otro responsable.
- **Olvido o supresión**: supresión de los datos personales del interesado sin dilación cuando dichos datos ya no sean necesarios conforme a su finalidad, se retire el consentimiento o se ejerza el derecho a oposición sobre los mismos.

- **Limitación del tratamiento**: posibilidad de exigir al responsable del tratamiento la limitación del tratamiento de sus datos personales.

El responsable debe facilitar el ejercicio de los derechos ARCO por un canal fácilmente accesible y, si es posible, por el mismo medio por el que recabó los datos personales del interesado.
Es habitual que las instituciones de salud tengan implementados procedimientos para que los pacientes ejerzan sus derechos ARCO (con foco en el uso asistencial), a partir de documentos de denegación de consentimiento ajustados a la normativa de protección de datos. Uno de los grandes problemas que conllevan estas situaciones es que las personas tienen derecho a consentir de forma selectiva el uso de los datos (por ejemplo, según ámbito de atención, especialidad, patología, etc.). Esto implica un gran desafío en la gestión de esta especificidad por parte de las instituciones de salud.

4 Riesgos

La realización de cualquier tratamiento de datos personales requiere de un análisis de riesgos RGPD en protección de datos, necesidad que aparece implícita en el RGPD y en la LOPDGDD art. 24 RGPD, 25.1 RGPD, art25.2 RGPD, art.32.2 RGPD. Dicho proceso de análisis de riesgos RGPD en protección de datos consiste en un análisis previo al tratamiento de datos personales que permite identificar los riesgos derivados de este para la protección de datos, y establecer las correctas acciones preventivas, correctivas y reductivas que minimicen el nivel de exposición al riesgo.

Se diferencian tres etapas para la correcta elaboración de un análisis de riesgos RGPD: **identificación de amenazas, evaluación de los riesgos y tratamiento de los riesgos**.

A continuación, se describen las principales amenazas identificadas dentro del ámbito de IMPaCT-Data, las cuales se corresponden con los riesgos potenciales que pueden darse desde el punto de vista de la protección de datos y que violarían la seguridad de la información: acceso ilegitimo a los datos, modificación no autorizada de los datos, eliminación de los datos y re-identifcación a partir de los datos.

4.1 Accesos ilegítimos a los datos

El acceso ilegítimo a los datos implica la pérdida de confidencialidad de la información, es decir, la cualidad de la información para no ser divulgada a personas o sistemas no autorizados. Las amenazas más comunes que originan dicho riesgo son la fuga de información, el uso de operaciones de tratamiento no autorizadas por el personal investigador y los ataques intencionados (hacking, suplantación de identidad, etc.).

Entre las posibles medidas a adoptar para mitigar dicho riesgo se encuentran la gestión de altas y bajas en el registro de usuarios, la gestión de los derechos de acceso con privilegios especiales, la revisión de los derechos de acceso de los usuarios, la retirada o adaptación de los derechos de acceso, el uso de la información confidencial para la autenticación, la

restricción del acceso a la información, los controles de red y la notificación de los eventos de seguridad de la información.

En el caso de IMPaCT-Data, existe el riesgo específico de malentendidos con el alcance de la legitimación de ciertos consentimientos para el uso de los datos en el entorno de IMPaCT-Data, sobre todo para datos que vienen con consentimientos y son recogidos desde fuera del propio IMPaCT-Data.

4.2 Modificación no deseada de los datos

La modificación de los datos implica la pérdida de integridad de la información, es decir, la cualidad que asegura su veracidad y autenticidad, garantizando la inexistencia de errores y la no alteración de los datos. Las amenazas más comunes que originan dicho riesgo son la perturbación de la información no intencionada, la modificación no autorizada de datos intencionada y los errores en los procesos de recopilación y captura de la información.

Entre las posibles medidas a adoptar para mitigar dicho riesgo se encuentran la creación y actualización de un registro de actividades del tratamiento, una política de control de accesos (gestión de los derechos de acceso con privilegios especiales, revisión de los derechos de los usuarios, retirada o adaptación de los derechos de acceso), la segregación de tareas, la realización de controles que aseguren la integridad de la información en la entrada, almacenamiento y procesamiento de los datos y la realización de copias de seguridad de la información.

4.3 Desaparición de los datos

La desaparición de los datos implica la pérdida de la disponibilidad de la información, entiéndase como la cualidad que posibilita el acceso a la información cuando sea necesario a través de los canales adecuados para ello. Las principales amenazas que originan dicho riesgo son la pérdida o borrado no intencionado de los datos, errores humanos o ataques intencionados que provocan pérdidas o borrado de datos, cortes de suministro eléctrico o fallos en los servicios de comunicaciones y desastres naturales.

Entre las posibles medidas a adoptar para mitigar dicho riesgo se encuentran la planificación e implantación de la continuidad de la seguridad de la información, la realización de copias de seguridad, respuestas a los incidentes de seguridad, notificación de los eventos de seguridad de información, recopilación de evidencias, controles de red, segregación de redes, firewalls externos IDs/IPS, Protección DDos (Denegación Distribuida de Servicio), uso aceptable de los activos, mantenimiento de los equipos y protección contra las amenazas externas.

4.4 Re-identificación a partir de los datos

El riesgo de re-identifcación de los datos personales está estrechamente relacionado con los procesos de anonimización y pseudonimización utilizados, ya que ninguna técnica de anonimización garantiza en términos absolutos la imposibilidad de la re-identificación, existiendo siempre una cierta probabilidad que debe intentar atenuarse. Las principales amenazas que lo originan son la inadecuada implantación de procedimientos de anonimización (pudiendo influir también una inadecuada formación o información del personal implicado en la anonimización o en el tratamiento de los datos anonimizados), así como una inadecuada gestión de las claves utilizadas en métodos basados en algoritmos de cifrado o huella digital.

Entre las posibles medidas a adoptar para mitigar dicho riesgo se encuentran los mecanismos y protocolos de anonimización que conlleven la definición del equipo de trabajo, las medidas organizativas, la formación del personal, las medidas de confidencialidad, el uso de posibles estándares, la utilización de códigos de buenas prácticas, etc.

4.4.1 Perfilado

En caso de que la investigación implique la elaboración de perfiles, el responsable deberá explicar cómo se informará a los interesados de la existencia de la elaboración de perfiles, de sus posibles consecuencias y de cómo se salvaguardarán sus derechos fundamentales.

Una cuestión fundamental es la necesidad de conocer a priori los mecanismos de salvaguarda de los derechos fundamentales en caso de la necesidad de perfilado.

El artículo 4.4 del RGPD define perfilado como una forma de tratamiento de datos personales que permite inferir más información acerca de una persona física, evaluando, analizando o prediciendo aspectos personales. Un tratamiento que implique la elaboración de perfiles se caracteriza por tres elementos:

- Debe ser una forma automatizada de tratamiento, incluyendo aquellos tratamientos que tienen participación parcialmente humana.
- Debe llevarse a cabo respecto a datos personales.
- Y el objetivo de la elaboración de perfiles debe ser evaluar aspectos personales sobre una persona física.

La aplicación de métodos de Inteligencia Artificial permite este tipo de tratamiento de los datos.
Algunas potenciales medidas para la salvaguarda de los derechos fundamentales de las personas son asegurar que las personas tomen la decisión final (incluyendo los perfiles), informar y ofrecer oportunidades para que los participantes/pacientes puedan objetar/excluirse de la elaboración de perfiles e impugnar la decisión y expresar su propio punto de vista.

5 Aspectos relacionados con las medidas de aplicación

5.1 Adecuación del modelo de análisis de riesgos al diseño funcional de IMPaCT-Data

Se desarrollará el análisis de riesgos cuando se disponga de la versión final del diseño funcional de IMPaCT-Data, empleando para ello la metodología PILAR, herramienta de Entorno de Análisis de Riesgos (EAR), que soporta el análisis y la gestión de riesgos siguiendo la metodología *Magerit* (Metodología de Análisis y Gestión de Riesgos de los Sistemas de Información), desarrollada y financiada parcialmente por el CCN

5.2 Medidas de anonimización y pseudonimización

Las principales diferencias entre ambas técnicas radican en las garantías que protegen los derechos de los pacientes. Como ya se ha comentado en apartados anteriores, aquellos conjuntos de datos anonimizados no se encuentran bajo el ámbito de aplicación del RGPD y la LOPDGDD, mientras que los conjuntos de datos pseudonimizados si lo están. E igualmente, los conjuntos de datos anonimizados impiden vincular los resultados con personas concretas evitando por tanto cualquier uso adicional de los datos personales.

Según el considerando 26 del RGPD, el **proceso de anonimización** genera un nuevo y único conjunto de datos que no guarda relación con una persona física identificable, considerándose compatible con el fin original del tratamiento de datos personales del que proceden los datos (Dictamen 05/2014 sobre técnicas de anonimización WP246 apartado 2.2.1. Legitimación del proceso de anonimización). El conjunto de datos estará anonimizado en la medida que no exista una probabilidad razonable de re-identificación de los individuos que lo integran mediante «el conjunto de los medios que puedan ser razonablemente utilizados» por el responsable del tratamiento o por terceros, es decir, se debe tratar de un proceso irreversible equivalente al borrado. La aplicación de las técnicas de anonimización debe diseñarse adecuadamente, con una clara definición de los requisitos previos y los objetivos del proceso. La solución óptima debe decidirse para cada caso y puede implicar la combinación de diversas técnicas. El proceso de anonimización no debe entenderse como un proceso esporádico, debe existir una evaluación regular de los riesgos existentes donde se valoren los costes, el tiempo requerido o los medios tecnológicos actuales, así como los avances tecnológicos en años venideros para alcanzar la reversión de la anonimización, y la realización de auditorías tanto internas como por equipos independientes. En cambio, el **proceso de pseudonimización** genera dos nuevos conjuntos de datos, la información pseudonimizada y la información vinculada con dicho conjunto de datos, la cual sí permite relacionar dichos datos pseudonimizados con una persona física identificable. El conjunto de datos pseudonimizados debe cumplir con las garantías del RGPD que establecen determinadas limitaciones, como el periodo de conservación, la comunicación de los datos pseudonimizados y las finalidades del uso de estos. Así como cumplir con las garantías

técnicas y jurídicas de la disposición adicional decimoséptima sobre tratamientos de datos de salud de la LOPDGDD y las garantías técnicas y organizativas dispuestas al efecto de impedir la materialización de brechas de datos personales, requisito que también concierne al conjunto de datos de la información adicional.
A continuación, se detalla las técnicas más comunes de anonimización y pseudonimización recogidas por la Agencia Española de Protección de Datos, en adelante AEPD.

5.2.1 Medidas de Anonimización

LA AEPD define el proceso de anonimización como «la ruptura de la cadena de identificación de las personas» y se diferencian las siguientes técnicas para su aplicación:

- **Generalización:** técnica que consiste en la conversión de los datos individuales en datos genéricos mediante el uso de escalas o magnitudes más amplias y generales, por ejemplo, sustituyendo una ciudad por una región, o una semana por un mes.
    - **K-anonimidad**: evitan que un interesado sea singularizado cuando se le agrupa junto con, al menos, un número k de personas. Para ello, los valores de los atributos se generalizan hasta el punto de que todos los individuos del conjunto de datos acaban compartiendo el mismo valor.
    - **Diversidad l:** aporta complejidad a la técnica anterior, asegurando que en cada clase de equivalencia, todos los atributos tienen al menos l valores diferentes. Es muy útil para proteger los datos ante ataques por inferencia, siempre y cuando los valores de los atributos estén bien distribuidos.
    - **Proximidad t:** Consiste en la creación de clases equivalentes que se parezcan a la distribución inicial de los atributos, es muy útil cuando hay que conservar los datos lo más próximo posible a los originales.
- **Aleatorización**: técnica que modifica la veracidad de los datos con el objetivo de eliminar el vínculo existente entre estos y el individuo al que pertenecen. Si los datos se hacen lo suficientemente ambiguos, no podrán remitir a una persona concreta. Para ello se emplean diferentes técnicas
    - **Adición de ruido:** Consiste en modificar los atributos del conjunto de datos para que sean menos precisos, conservando su distribución general.
    - **Permutación**: Se mezclan valores de los atributos de un conjunto de datos para que algunos de ellos se vinculen a distintos interesados, teniendo cuidado de no cambiar la relación lógica existente.
    - **Privacidad diferencial:** técnica que actúa sobre el proceso de transformación o algoritmo de consulta y publicación de los datos analizados sin alterar los datos originales. Se basa en funciones matemáticas que añaden ruido aleatorio a los resultados de la consulta realizada. La característica de este método es cómo se calcula la intensidad del ruido que se añade a los resultados, ya que mucho ruido resta utilidad a los datos y poco ruido, en cambio, permite estimar la realidad con relativo poco esfuerzo. Este equilibrio viene determinado por el parámetro ε o presupuesto de privacidad, que establece el equilibrio entre la precisión del resultado de la consulta (exactitud) y la protección de la información consultada (privacidad).

5.2.2 Medidas de Pseudonimización

En el tratamiento de la pseudonimización como se ha comentado lo más importante es determinar cómo se va a proteger la información adicional, ya que está es la que permite la identificación del paciente. Se diferencian las siguientes técnicas para su aplicación:

- **Cifrado con clave secreta:** técnica que consiste en que tanto el emisor como el receptor encriptan y desencriptan la información con una misma clave k, clave secreta, que ambos comparten. Su principal ventaja es la eficiencia al emplear algoritmos de rápida ejecución, y su desventaja la necesidad de compartir la clave, precisando de un medio con alta seguridad para ello.
- **Función Hash:** técnica que empleando la función hash y a partir de un valor de entrada de cualquier tamaño devuelve un resultado de tamaño fijo. Esta puede ser reversible si se conocen los valores de entrada de la función hash y se aplican ataques de fuerza bruta, por ello para reducir el riesgo de re-identificación se añade un valor aleatorio, al atributo al que se aplica la función hash, técnica conocida como función hash con *salt.*
- **Función con clave almacenada**: técnica que se diferencia de la anterior en el uso de una clave secreta como valor de entrada suplementario, aumentando así el riesgo de revertir el resultado.
- **Cifrado determinista o función hash con clave con borrado de clave:** técnica que consiste en la generación de un número aleatorio para cada atributo personal del conjunto de datos y eliminación de la tabla de correspondencia posteriormente.
- **Descomposición en tokens**: técnica que consiste en la aplicación de mecanismos de cifrado unidireccionales o en la asignación de un número de secuencia o número generado aleatoriamente que no derive matemáticamente de los datos originales.

5.2.3 Medidas de anonimización y pseudonimización en textos clínicos

El primer paso para iniciar el proceso de deidentificación de datos clínicos consiste en la detección y tratamiento de datos sobre Información de Salud Protegida (PHI) presentes en el texto. Los datos PHI se refieren a entidades que contienen información demográfica, fechas, instituciones o resultados de pruebas médicas, entre otros. Según el nivel de deidentificación requerido, se pueden tener en cuenta unas categorías u otras.

5.2.3.1 Detección de entidades sensibles

La detección de las categorías PHI puede llevarse a cabo de forma manual o automática. La detección manual, mediante la revisión de cada uno de los textos de forma independiente, permite una deidentificación de gran precisión a costa de una gran inversión de tiempo y esfuerzo. No obstante, la gran velocidad en la que se crean datos de origen clínico obliga a la exploración de métodos de deidentificación automáticos, tales como el uso de expresiones regulares, listas de conceptos y modelos de aprendizaje automático y procesamiento de lenguaje natural. Estos métodos también requieren de cierta supervisión humana para

garantizar la eliminación de todos los datos sensibles del texto, pero son de gran ayuda para agilizar el proceso.

Las **expresiones regulares** se pueden utilizar para detectar entidades con un patrón más o menos definido y fácilmente distinguible. Entre estas entidades destacan las fechas, números de teléfono o códigos identificadores. El uso de expresiones regulares requiere un gran trabajo manual para poder capturar todas las maneras posibles de expresar un tipo de entidad. Esto presenta un problema cuando existen muchas variantes de un mismo patrón. Por ejemplo, las fechas se pueden escribir utilizando diferentes elementos de separación, distinto orden entre los elementos que las componen o alterando números y texto. Además, las expresiones regulares no son robustas frente a cualquier error tipográfico que se pueda cometer en la redacción; esto provoca que, en caso de error tipográfico, la entidad quede sin ser reconocida en el texto.

Por otro lado, las **listas de términos** se pueden emplear para detectar nombres propios o de personas o lugares, entre otros. El primer problema que puede presentar este método es que la cobertura de la lista no sea suficiente: es decir, que existan términos pertenecientes a la categoría a detectar que no aparecen en el listado de términos disponible. Otro problema es la posibilidad de que en un término recogido en la lista de una categoría concreta sea ambiguo, provocando falsos positivos. Un ejemplo de este segundo problema podría ser el apellido español “Cabeza”, que también puede referirse al área anatómica de mismo nombre.

Entre los métodos más novedosos para detectar y categorizar términos destaca el reconocimiento de entidades nombradas o ***named entity recognition*** **(NER)**. Esta metodología forma parte del conjunto de herramientas del procesamiento de lenguaje natural y permite identificar en un texto entidades clave previamente definidas, que pueden estar compuestas de una o más palabras. Estos métodos están basados en el entrenamiento y la aplicación de modelos predictivos de aprendizaje automático (*machine learning*). Estos sistemas se generan con conjuntos de datos de entrenamiento en los que los expertos han marcado la categorías sensibles que deben ser detectadas por los modelos automáticos. Las entidades que pueden ser reconocidas son bastante versátiles, pudiendo contemplar desde conceptos generales como nombres propios o de lugares a conceptos técnicos como nombres de genes o compuestos. La desventaja de este método con respecto a los nombrados anteriormente es que los modelos predictivos necesitan datos anotados para poder aprender y la creación de estos datos requiere tiempo y esfuerzo humano.

5.2.3.2 Tratamiento de entidades sensibles

Una vez reconocidos y marcados los datos sensibles, se pueden seguir tres estrategias para su tratamiento. La primera de ellas consiste en la **sustitución** de los términos a anonimizar por datos completamente ficticios; de esta forma se garantiza que la privacidad del paciente no se vea comprometida, pero se imposibilita la distinción entre un informe deidentificado de uno original. Esta estrategia suele ser la más óptima si queremos crear un conjunto de datos cercanos a los reales, pero conlleva un mayor esfuerzo para crear sustitutos adecuados. La segunda estrategia posible a seguir para el tratamiento de los datos sensibles consiste en la **eliminación** de todos los términos marcados, de tal manera que no haya forma de recuperar esa información. La eliminación es la estrategia más sencilla y rápida, aunque tiene la desventaja de que los textos resultantes tendrán menos sentido debido a los huecos que

quedan al eliminar los datos sensibles. La última y tercera estrategia es la **agrupación** de los términos a anonimizar en conjuntos o agregadores amplios que te ayuden a utilizarlos sin entrar en problemas de re-identificación. Un ejemplo es usar el nombre de la categoría de información a la que pertenece, por ejemplo sustituyendo las direcciones postales por la palabra "DIRECCIÓN". Agrupar los términos es una estrategia intermedia entre las dos presentadas anteriormente que permite que el texto mantenga cierto sentido sin tener que dedicar tanto esfuerzo en crear sustitutos realistas. La desventaja es que señalan claramente que un texto ha sido deidentificado.

Por estos motivos, debido a la diversidad de estrategias, es muy relevante la elección y evaluación de cada uno de los métodos para verificar cual de las estrategias reduce más el riesgo de re-identificación del paciente.

5.2.4 Medidas de anonimización y pseudonimización en imagen médica

Podemos encontrar datos sensibles de los sujetos en las imágenes médicas a diferentes niveles. A nivel de los metadatos que contienen las imágenes y a nivel de la propia imagen.

5.2.4.1 Metadatos

El estándar DICOM proporciona una amplia definición sobre la eliminación de información sensible relacionada con los sujetos de estudio que se encuentra en los metadatos de las imágenes. Toda esta información se encuentra recogida dentro de la **Parte 15 de DICOM**: *Perfiles de seguridad y gestión del sistema*, en la ***Sección E**: Perfiles de confidencialidad de atributos* (14). El estándar DICOM proporciona unos perfiles para eliminar cualquier información que contenga o pueda contener información relacionada con el sujeto, permitiendo al mismo tiempo conservar la información necesaria para que los datos sigan siendo útiles para su propósito.

Por defecto se recomienda utilizar el *Perfil de Confidencialidad de Nivel de Aplicación Básico*, que define un enfoque extremadamente conservador que elimina o sustituye toda la información relacionada con:

- Identidad y características demográficas del paciente.
- Identidad de los posibles responsables o familiares.
- Identidad del personal implicado en el procedimiento.
- Identidad de las organizaciones implicadas en la petición o realización del procedimiento
- Información adicional que se podría utilizar para cotejar las instancias si se diera acceso a los originales, como los UID, fechas y horas.
- Atributos privados.

Para añadir flexibilidad a este enfoque por defecto, el estándar DICOM proporciona las siguientes opciones para conservar la información, que de otro modo se eliminaría, pero que es necesaria para usos específicos:

- Conservar la información temporal longitudinal con la opción de fechas completas o modificadas.
- Conservar las características del paciente.
- Conservar la opción de identidad del dispositivo.
- Conservar la opción de identidad de la institución.
- Conservar los Unified Identifier DICOM (UIDs) (15).
- Limpiar descriptores.
- Limpiar el contenido estructurado.
- Limpiar gráficos.
- Conservar la opción de seguridad privada.

La lista de atributos a procesar y su método de procesamiento para cada opción están listados en la Tabla E1-1 del estándar DICOM (Tabla 1). Esta tabla enumera un gran número de etiquetas DICOM que se eliminan o sustituyen por valores ficticios en el perfil básico.

D	replace with a non-zero length value that may be a dummy value and consistent with the VR
Z	replace with a zero length value, or a non-zero length value that may be a dummy value and consistent with the VR
X	remove
K	keep (unchanged for non-sequence attributes, cleaned for sequences)
C	clean, that is replace with values of similar meaning known not to contain identifying information and consistent with the VR
U	replace with a non-zero length UID that is internally consistent within a set of Instances
Z/D	Z unless D is required to maintain IOD conformance (Type 2 versus Type 1)
X/Z	X unless Z is required to maintain IOD conformance (Type 3 versus Type 2)
X/D	X unless D is required to maintain IOD conformance (Type 3 versus Type 1)
X/Z/D	X unless Z or D is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1)
X/Z/U*	X unless Z or replacement of contained instance UIDs (U) is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1 sequences containing UID references)

Attribute Name	Tag	Retd. (from PS3.6)	In Std. Comp. IOD (from PS3.3)	Basic Prof.	Rtn. Safe Priv. Opt.	Rtn. UIDs Opt.	Rtn. Dev. Id. Opt.	Rtn. Inst. Id. Opt.	Rtn. Pat. Chars. Opt.	Rtn. Long. Full Dates Opt.	Rtn. Long. Modif. Dates Opt.	Clean Desc. Opt.	Clean Struct. Cont. Opt.	Clean Graph. Opt.
Accession Number	-8,005	N	Y	Z										
Acquisition Comments	-18,4	Y	N	X								C		
Acquisition Context Sequence	-40,0555	N	Y	X/Z									C	
Acquisition Date	-8,0022	N	Y	X/Z						K	C			
Acquisition DateTime	(0008,002A)	N	Y	X/Z/D						K	C			
Acquisition Device Processing Description	-18,14	N	Y	X/D								C		
Acquisition Field Of View Label	(0018,1188)	N	Y	D								C		
Acquisition Protocol Description	-18,9424	N	Y	X								C		

Tabla 1.- Muestra del procesamiento de las etiquetas DICOM en función de la opción elegida

La RSNA ha trabajado de la mano de expertos DICOM para desarrollar una Guía y un Protocolo Deidentificación de Datos basada en estándares y buenas prácticas reconocidos. Dentro del ámbito de metainformación es necesario asegurar que los procedimientos de deidentificación sigan este protocolo. (Ver anexo B). El procedimiento requiere que se cree un identificador único (seudónimo) para cada sujeto.

5.2.4.2 Imagen

Otro punto importante en la anonimización/pseudonimización de la imagen médica es la revisión de la propia imagen para confirmar que no exista ningún elemento identificativo en las mismas.

La realización de una revisión visual de todas las imágenes clínicas se convierte en una tarea muy difícil de llevar a cabo cuando se trata de conjuntos de datos con un número elevado de imágenes. Hay que tener en cuenta también que cada modalidad de imagen tiene sus propias características y es muy difícil establecer un modelo que generalice la detección, edición o eliminación de los pixeles identificativos de las imágenes. Para facilitar esta tarea, se está implementando el uso de modelos de Inteligencia Artificial (IA) para la detección de información sensible de pacientes en las imágenes (16).

Una de las formas más sencillas de tratar que nos podemos encontrar es cuando la información sensible en forma de texto se encuentra en zonas concretas de la imagen, tales como los bordes superiores o inferiores, donde además no existe información relevante de la misma. En estos casos, una simple máscara que elimine la parte inferior y superior sería suficiente (Figura 3). Por desgracia, existen casos más complejos donde la información identificativa se encuentra embebida dentro de la imagen médica, en áreas con información médica relevantes. En estas situaciones, se recurre a modelos de IA como EAST (EAST: An Efficient and Accurate Scene Text Detector) (17). No obstante, sea cual sea la técnica automática utilizada, es conveniente hacer una revisión manual de las imágenes para comprobar que no aparece texto sensible en ellas.

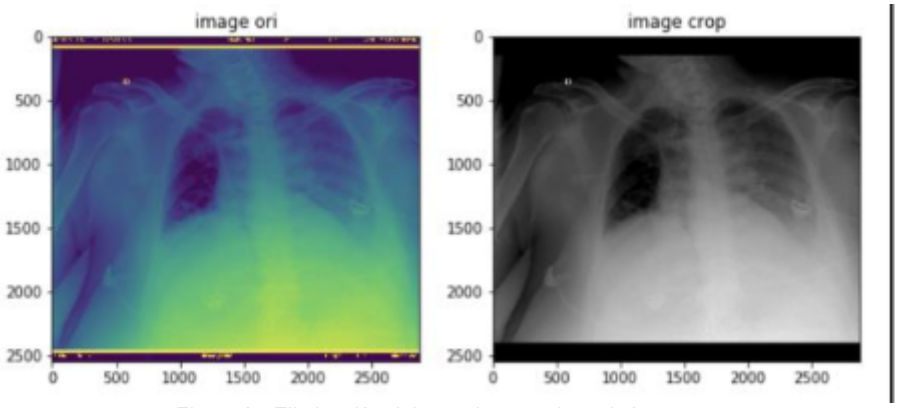

Figura 3.- Eliminación del texto incrustado en la imagen

Otro aspecto a tener en cuenta a la hora de proteger la privacidad de los pacientes es la capacidad que tienen algunas técnicas de realizar una **reconstrucción facial del sujeto** a partir de neuroimágenes (18). Para evitar esta reconstrucción, se pueden trabajar directamente eliminando la estructura ósea de las imágenes (19) pero se trata de una técnica agresiva que puede eliminar estructuras importantes que pueden ser relevantes para posteriores análisis.

Por este motivo, existen otras técnicas menos agresivas que se encargan de desfigurar la reconstrucción facial eliminando o difuminando determinadas partes de la imagen, como la nariz, la boca, los ojos, la barbilla y/o las orejas (Figura 4). Existen multitud de herramientas que se encargan de realizar este tipo de anonimización, como afni_refacer (20), deepdefacer

(21), mri_deface (22), mridefacer (23), pydeface (24), quickshear (25). Aunque la eliminación de las estructuras relevantes en la posible reconstrucción facial es más segura que la difuminación de las mismas, hay que tener en cuenta que estas técnicas no son del todo infalibles, pues existen otros métodos capaces de revertir la anonimización y volver a la imagen original (26).

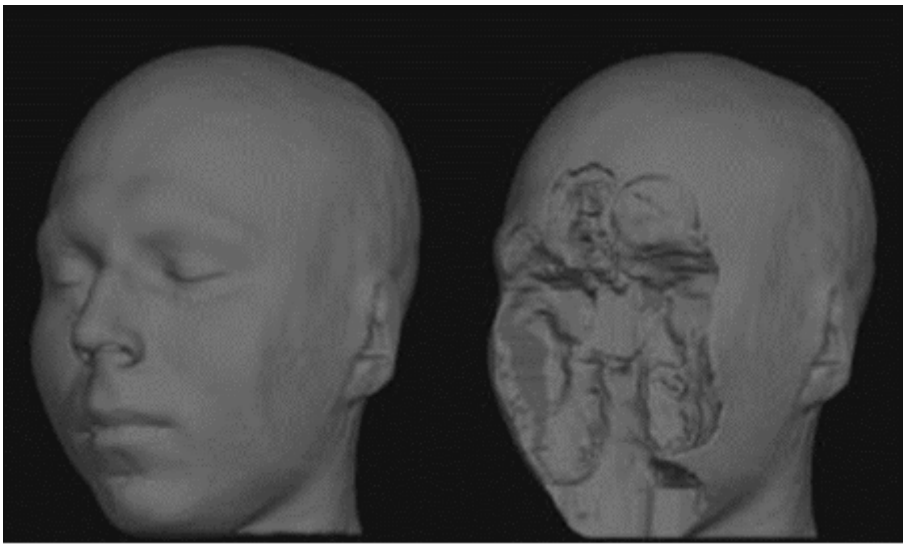

Figura 4.- Ejemplo de anonimización facial mediante la herramienta mri_deface

5.2.5 Medidas propuestas de anonimización y pseudonimización en información genómica

La naturaleza única de los datos genéticos obliga a extremar las precauciones en el proceso de anonimización y pseudonimización de esta información, no siendo suficiente la eliminación de la identidad del donante. Como se ha demostrado en varios estudios científicos, la combinación de recursos genéticos públicos y los metadatos de donantes de ADN (fecha de donación, edad, lugar de residencia) puede revelar la identidad de donantes anónimos, por ello es necesario la aplicación de técnicas sobre la información genómica que aseguren su privacidad.

Entre las técnicas más utilizadas para tal fin se encuentran:

- **Transformación de datos:** la técnica de k-anonimidad es comúnmente utilizada en datos genómicos, un ejemplo de su aplicación es la generalización de nucleótidos en tipos más amplios basados en sus propiedades para satisfacer el k-anonimato. Sin embargo, debido a la alta dimensionalidad de los datos genómicos, las estrategias basadas en generalización o aleatorización no tienden a mantener los datos a un nivel de detalle útil para su estudio.
- **Agregación de datos**: consisten en técnicas basadas en sistemas semi-confiables de consulta, que procesan las consultas internamente y devuelven sólo los resultados resumidos, evitando así el tratamiento de los datos a nivel individual. El sistema más destacado actualmente para este fin, son los servicios BEACON, popularizados por la Global Alliance for Genomics and Health (GA4GH), servicios que sólo permiten

realizar consultas de un tipo de información dentro de datos genómicos (presencia de alelos).

- A pesar, de estar considerado como uno de los métodos más seguros en la preservación de la seguridad de los datos genómicos, se ha demostrado que es susceptible a ataques de inferencia, sin embargo, los efectos de este ataque pueden minimizarse añadiendo ruido, añadiendo familiares o parientes o cambiando estratégicamente las respuestas a la consulta para un subconjunto de variantes genéticas.
- **Ocultación de datos:** técnicas basadas en la adicción de ruido, dentro de esta categoría se encuadra la técnica de la privacidad diferencial, la cual garantiza, mediante la incorporación de ruido aleatorio a la información original, que en el resultado del proceso de análisis de los datos no haya pérdida en la utilidad de los resultados obtenidos.
- Dicha técnica presenta limitaciones en la protección de los GWAS, ya que requiere de una gran cantidad de ruido para proporcionar protección, por ello es recomendable su aplicación junto con otras técnicas criptográficas modernas como el cifrado homomórfico, la cual permite realizar funciones predefinidas sobre datos cifrados sin necesidad de descifrarlos.
- **Generación de datos sintéticos:** el uso de conjuntos de datos genómicos sintéticos asegura la protección del anonimato, al mismo tiempo que mantienen la utilidad al replicar la mayoría de las características de los datos originales. Las técnicas de aprendizaje profundo más comunes para la generación de estos son las redes neuronales generativas adversarias (GAN)s o máquinas de Boltzmann restringidas

5.2.6 Armonización de la anonimización entre diferentes tipologías de datos en proyectos con usos comunes

Es cada vez más habitual que los proyectos de investigación incluyan datos clínicos procedentes de fuentes muy heterogéneas y con tipología y formatos muy diferenciados (datos numéricos, codificados, textuales, imágenes, genómicos, etc.).

Para poder trabajar con ellos resulta fundamental mantener la vinculación de estos distintos tipos de datos al sujeto al que corresponden. Esta condición fundamental implica una armonización de la anonimización de identificadores.

Cada organización, cada proyecto y cada consorcio de investigación tiene sus propios flujos de trabajo, lo cual puede determinar distintas necesidades. Por ello, es importante no limitar el proceso de generación de identificadores anonimizados a la tipología específica del dato. Dicho de otra forma, se debe asegurar que sea un identificador único anonimizado el que sustituya a los identificadores utilizados en el proceso asistencial del paciente sea cual sea el tipo de datos que vaya a ser extraído para el estudio analítico.

Debe, además, considerarse que los datos de un único sujeto pueden utilizarse para múltiples proyectos, ya sean los mismos datos o diferentes. De esta forma, los datos extraídos para cada proyecto de un mismo sujeto podrán tener distintos identificadores para la gestión de cada estudio.

Las organizaciones sanitarias que desarrollan un proyecto de uso secundario de los datos Clínicos para investigación, deben contar con un servicio de disociación corporativo, centralizado y sometido a un estricto modelo de gobernanza en el que se conserve la identidad real del paciente con la generada “ad hoc” para cada proyecto, incluyendo en esta vinculación a la identificación del proyecto en cuestión. De esta forma, en un modelo de cesión de datos a terceros por exportación de los mismos se registra documentalmente el destino de los datos de cada paciente a los correspondientes proyectos.

Esta traza permite satisfacer las demandas, cada vez mayores, de los propios pacientes que si bien autorizan la cesión de sus datos anonimizados para la investigación exigen saber para qué se han utilizado y cuáles han sido los resultados y beneficios esperados de los proyectos a los que de alguna forma ellos mismos han contribuido.

5.3 Medidas propuestas de accesibilidad y trazabilidad

La infraestructura desarrollada en IMPaCT-Data será utilizada por distintos equipos investigadores colaboradores que realizarán análisis sobre diversos conjuntos de datos, siendo imprescindible por ello mantener un registro de quién hizo qué, sobre qué conjuntos de datos, en qué fechas y con qué propósito. Este registro procedente de la ejecución de los flujos de trabajo permitirá la trazabilidad de los orígenes de los datos, de los procesos, y de su evolución entre las diferentes etapas de su uso, posibilitando consultar la información de los análisis, así como la detección de errores y de accesos no autorizados, y en última instancia la verificación del cumplimiento de ciertos requisitos legales.

5.3.1 Control de acceso

Debido a que diversos usuarios accederán a la plataforma IMPaCT- Data es necesario disponer de un sistema de control de acceso y seguridad con el fin de restringir o permitir el acceso a la información, detectar accesos no autorizados y registrar y revisar eventos críticos realizados por los usuarios en los sistemas.

El control de acceso y seguridad se sustenta en los principios de identificación, autenticación y autorización. En este caso, la red de autenticación de IMPaCT-Data se basará en la implementación de una federación de proveedores de identidad correspondientes a los distintos nodos proveedores de computación y datos. Las identidades de los usuarios de IMPaCT-Data serán validadas por un nodo central vía OIDC/oAUTH2, que actuará como intermediario, redirigiendo las peticiones a los distintos centros donde se delegará dicho proceso de identificación, obteniendo de esta manera un sistema de administración central de usuarios que permita implementar una estrategia de autorización compartida entre los distintos centros y un servidor de autentificación común para habilitar “Inicio de Sesión Unificado" o procedimientos. Este nodo central, para primer acceso, ofrece al usuario la posibilidad de registro y para accesos sucesivos gestiona la validación de la identidad,

aceptando identidades proporcionadas por las instituciones que se designen, además de proveedores reconocidos como Life Sciences Login, ORCID, etc.

La política de control de acceso se basará en roles (*Role Based Access Control*), estableciendo una serie de derechos y responsabilidades asociadas a una determinada actividad, actividad asociada a su vez a un rol especifico, y en base a la cual se gestionan los derechos y permisos de acceso. Los distintos roles que se contemplan en IMPaCT-Data, especificados en el entregable *E2.2.Diseño Inicial de la Infraestructura*, son los siguientes:

- **Investigador:** Usuario que requiere acceso a los datos para realizar nuevos análisis combinando herramientas y datos. Su acceso al sistema se realizará a través de un portal web, donde tras identificarse, podrá acceder a un espacio virtual de trabajo con los recursos específicos para dicho rol (catálogo de los datos disponibles en el sistema, un espacio personal de datos, un catálogo de herramientas adecuado a la naturaleza de los datos, y resultados y estadísticas de análisis ya realizados).
- **Desarrollador/a:** Usuario que requiere acceso al entorno de desarrollo, interesado en el desarrollo de herramientas y algoritmos de análisis o gestión de datos.
- **Data Steward:** Usuario encargado de la gestión de los datos alojados y producidos por el sistema y cumplimiento de los principios FAIR, el cual requiere acceso al catálogo de datos del sistema, para gestionar su integridad, metadatos, y realizar actuaciones de procesado o armonización de estos.
- **Proveedor de datos:** Usuario responsable de proyectos de recolección de los datos que se incorporarán al sistema y que debe acceder al catálogo de datos para llevar a cabo la correcta gestión de los datos.

Las solicitudes del acceso a datos realizados por los distintos usuarios de IMPaCT-Data, serán gestionadas por el **Comité de acceso a datos,** persona u organismo que ejercerá la responsabilidad de controlar el acceso a datos controlados, y proveer/revocar credenciales de acceso a los mismos. Dicho Comité accederá a un portal web específico en el que se detallarán las solicitudes de acceso y su estado (concedida, en progreso, denegada, revocada) y se facilite la gestión de estas.

5.3.2 Herramientas analíticas sobre logs de acceso y detección de operaciones

De entre las herramientas más populares para la gestión de logs se implementará la más apropiada que cumpla con los requisitos y las necesidades de la plataforma desarrollada en IMPaCT-Data:

- **ELK (Elasticsearch, Logstash & Kibana**) (27): plataforma open source implementada en Java que utiliza como backend varios paquetes populares de administración de registros como Graylog2, Logstash y Kibana. Integrada por 3 componentes principales: un agente de recogida de logs, una base de datos donde

almacenar, indexar y buscar los eventos de logs de las aplicaciones y una aplicación frontend donde consultar los eventos.
Presenta como ventajas su potencia, la escalabilidad que ofrece (los clusters de Elasticsearch pueden manejar terabytes de datos), su flexibilidad (disponer de una configuración abierta y flexible que se adapta a cualquier entorno y necesidad) y su apertura al contar con plugins y APIs para extender casi cualquier aspecto de la plataforma.

- **Graylog** (28): plataforma open-source y con planes para empresa, permite una fácil gestión de registros de datos estructurados y no estructurados junto con aplicaciones de depuración. Basada en Elasticsearch, MongoDB y Scala, se diferencia de ELK en que ofrece una solución de aplicación única en la recopilación, análisis y visualización de datos, sin necesidad de instalar varios componentes.
- **Fluentd** (29): herramienta open-source implementada en C que unifica la recogida de registros y datos de múltiples fuentes. Estructura los datos en un formato JSON, unificando así todas las etapas del registro de datos (recopilación, filtrado, análisis sintáctico y salida de registros en múltiples nodos). Fluentd ocupa poco espacio, consume pocos recursos y ofrece una amplia flexibilidad a través de sus más de 500 plugins desarrollados por la comunidad.
- **LOGalyze** (29): herramienta gratuita de monitorización y gestión de registros de red, para analizar los registros de servidores y aplicaciones, presentándolos en varios formatos de informe (PDF, CSV y HTML). Dispone al igual que las herramientas anteriores de una interfaz web intuitiva donde los usuarios tras iniciar sesión pueden monitorizar varias fuentes de datos y analizar los archivos de registro.
- **NXlog** (31)**:** herramienta potente y versátil con versión gratuita para la gestión de registros multiplataforma que permite identificar riesgos de seguridad y analizar problemas en los registros del sistema, las aplicaciones y los servidores. Permite realizar diversas tareas sobre los registros (rotación de registros, reescritura de registros, compresión de registros) y configurarse para envío de alertas.

6 Conclusiones

En conclusión, en este entregable se han repasado los aspectos legales, éticos, organizativos y técnicos que se deben cumplir para el correcto tratamiento seguro de datos sensibles en el uso de estos con fines de investigación, especificando la diferencia que conlleva el uso de conjuntos de datos anonimizados y pseudonimizados así como por la variada naturaleza y origen de los mismos. Se abarca la necesidad del consentimiento informado de los pacientes como base legal, así como las situaciones de excepción que pueden presentarse. Respecto a las responsabilidades del tratamiento se sigue las especificaciones marcadas por el esquema Nacional de Seguridad y la Guía CCN-STIC 801, indicando los distintos roles que deben estar presentes. En cuanto a los riesgos relacionados con la seguridad que conlleva el tratamiento de datos sensibles, se describen el acceso ilegitimo a los datos, la modificación no autorizada de los datos, la eliminación de los datos y la re-identificación a partir de los datos, así como las posibles acciones preventivas, correctivas y reductivas que minimizan el nivel de exposición a los distintos riesgos. Respecto a las implicaciones de seguridad relacionadas con la infraestructura, es fundamental el vínculo con el E2.2 de diseño de la infraestructura, porque a partir de este se enumeran los componentes identificados que integrarán la misma, la política de control de accesos basada en roles y el sistema de autenticación que implementará IMPaCT-Data, especificando la necesidad de realización de un acuerdo de tratamiento de datos cuando existan transferencias de datos entre distintos nodos participantes. A destacar que el contenido de dicho entregable se ha realizado teniendo en cuenta la tipología de la información con la que se va a trabajar en IMPaCT-Data, diferenciando los aspectos relevantes de seguridad a considerar en datos clínicos, imagen médica y datos genómicos, así como exponiendo las más avanzadas técnicas de anonimización y pseudonimización que están siendo aplicadas actualmente.

Referencias

1. Somolinos, R., Muñoz, A., Hernando, M. E., Pascual, M., Cáceres, J., Sánchez-de-Madariaga, R., ... & Salvador, C. H. (2014). Service for the pseudonymization of electronic healthcare records based on ISO/EN 13606 for the secondary use of information. IEEE journal of biomedical and health informatics, 19(6), 1937-1944.
2. Página web especificación de datos recogido por The Cancer Genomic Data (TCGA) https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/types
3. March, S., Andrich, S., Drepper, J., Horenkamp-Sonntag, D., Icks, A., Ihle, P., ... & Hoffmann, F. (2020). Good practice data linkage (GPD): a translation of the German version. International Journal of Environmental Research and Public Health, 17(21), 7852.
4. Página web portal Gen Expression Omnibus, del Nacional Center for Biotechnology Information, GEO NCBI https://www.ncbi.nlm.nih.gov/geo/info/submissionftp.html
5. Página web portal Genomic Data Commons (GDC) del National Cancer Institute NIH https://docs.gdc.cancer.gov/
6. Página web de European Network of Cancer Registries https://www.encr.eu/
7. International Classification of Diseases for Oncology (ICD-O) de la International association of Cancer Registries
8. Página web de la Red Española de Registros de Cáncer https://redecan.org/es
9. Nature Guidelines on scientific data https://www.nature.com/sdata/policies/repositories
10. Crossfield, S. S., Zucker, K., Baxter, P., Wright, P., Fistein, J., Markham, A. F., ... & Hall, G. (2022). A data flow process for confidential data and its application in a health research project. PloS one, 17(1), e0262609.
11. Real Decreto 311/2022, de 3 de mayo, por el que se regula el Esquema Nacional de Seguridad https://www.boe.es/eli/es/rd/2022/05/03/311/dof/spa/pdf
12. World Medical Association. (2016). WMA declaration of **Taipei** on ethical considerations regarding health databases and biobanks. Secondary WMA Declaration of taipei on ethical considerations regarding health databases and biobanks. https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/
13. Informed Consent Templates from WHO Research Ethics Review Committee (ERC) https://www.who.int/groups/research-ethics-review-committee/guidelines-on-submitting-research-proposals-for-ethics-review/templates-for-informed-consent-forms
14. Perfiles de confidencialidad de atributos. Estándar DICOM. https://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E
15. Unique Identifiers (UIDs). Estándar DICOM. https://dicom.nema.org/dicom/2013/output/chtml/part05/chapter_9.html

16. Monteiro, E., Costa, C., & Oliveira, J. L. (2015, August). A machine learning methodology for medical imaging anonymization. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1381-1384). IEEE
17. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). East: an efficient and accurate scene text detector. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 5551-5560).
18. Schwarz, C. G., Kremers, W. K., Therneau, T. M., Sharp, R. R., Gunter, J. L., Vemuri, P., ... & Jack Jr, C. R. (2019). Identification of anonymous MRI research participants with face-recognition software. New England Journal of Medicine, 381(17), 1684-1686.
19. Kalavathi, P., & Prasath, V. B. (2016). Methods on skull stripping of MRI head scan images—a review. Journal of digital imaging, 29(3), 365-379.
20. Proyecto afni_refacer en el repositorio GitHub https://github.com/PennLINC/afni_refacer
21. Descripción proyecto deepfacer https://pypi.org/project/deepdefacer/
22. Ejemplo uso de la herramienta mri_deface https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface
23. Proyecto mridefacer en el repositorio GitHub https://github.com/mih/mridefacer
24. Descripción proyecto pydeface https://pypi.org/project/pydeface/
25. Proyecto quickshear en el repositorio GitHub https://github.com/nipy/quickshear
26. Abramian, D., & Eklund, A. (2019, April). Refacing: reconstructing anonymized facial features using GANs. In 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019) (pp. 1104-1108). IEEE
27. Página web herramienta ELK Stack https://www.elastic.co/es/what-is/elk-stack
28. Página web herramienta Graylog: Industry Leading Log Management https://www.graylog.org/
29. Página web Fluentd https://www.fluentd.org/
30. Página web LOGalyze http://www.zuriel.hu/en/
31. Página web NXlog https://nxlog.co/

Acrónimos y Abreviaturas

ARCO	Acceso, rectificación, Cancelación y Oposición
CCN-STIC	Centro Criptológico Nacional – Seguridad de las tecnologías de la Información y las Comunicaciones
CVMFS	CernVM File System
DICOM	Digital Imaging and Communication In Medicine
DPD	Delegado de Protección de Datos
EGA	European Genome-Phenome Archive
ENS	Esquema Nacional de Seguridad
GA4GH	Global Alliance for Genomics and Health
GAN	Generative Adversarial Networks
GWAS	Genome-wide association study
IMPaCT	Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología
IMPaCT-Data	Programa de ciencia de datos de IMPaCT
LOPDGDD	Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales
NER	Name Entity Recognition
OICD	OpenID Connect
RGPD	Reglamento General de Protección de Datos
UE	Unión Europea

Anexo A. Tipos de tratamientos de datos que requieren evaluación de Impacto relativa a la Protección de Datos de la AEPD

Esta lista se basa en los criterios establecidas por el Grupo de Trabajo del Artículo 29 en la guía WP248 “Directrices sobre la evaluación de impacto relativa a la protección de datos (EIPD) y para determinar si el tratamiento «entraña probablemente un alto riesgo» a efectos del RGPD”, los complementa y debe entenderse como una lista no exhaustiva:

1. Tratamientos que impliquen perfilado o valoración de sujetos, incluida la recogida de datos del sujeto en múltiples ámbitos de su vida (desempeño en el trabajo, personalidad y comportamiento), que cubran varios aspectos de su personalidad o sobre sobre sus hábitos.
2. Tratamientos que impliquen la toma de decisiones automatizadas o que contribuyan en gran medida a la toma de tales decisiones, incluyendo cualquier tipo de decisión que impida a un interesado el ejercicio de un derecho o el acceso a un bien o un servicio o formar parte de un contrato.
3. Tratamientos que impliquen la observación, monitorización, supervisión, geolocalización o control del interesado de forma sistemática y exhaustiva, incluida la recogida de datos y metadatos a través de redes, aplicaciones o en zonas de acceso público, así como el procesamiento de identificadores únicos que permitan la identificación de usuarios de servicios de la sociedad de la información como pueden ser los servicios web, TV interactiva, aplicaciones móviles, etc.
4. Tratamientos que impliquen el uso de categorías especiales de datos a las que se refiere el artículo 9.1 del RGPD, datos relativos a condenas o infracciones penales a los que se refiere el artículo 10 del RGPD o datos que permitan determinar la situación financiera o de solvencia patrimonial o deducir información sobre las personas relacionada con categorías especiales de datos.
5. Tratamientos que impliquen el uso de datos biométricos con el propósito de identificar de manera única a una persona física.
6. Tratamientos que impliquen el uso de datos genéticos para cualquier fin.
7. Tratamientos que impliquen el uso de datos a gran escala. Para determinar si un tratamiento se puede considerar a gran escala se considerarán los criterios establecidos en la guía WP243 “Directrices sobre los delegados de protección de datos (DPD)” del Grupo de Trabajo del Artículo 29.
8. Tratamientos que impliquen la asociación, combinación o enlace de registros de bases de datos de dos o más tratamientos con finalidades diferentes o por responsables distintos.
9. Tratamientos de datos de sujetos vulnerables o en riesgo de exclusión social, incluyendo datos de menores de 14 años, mayores con algún grado de discapacidad, discapacitados,

personas que acceden a servicios sociales y víctimas de violencia de género, así como sus descendientes y personas que estén bajo su guardia y custodia.

Anexo B. Perfil de de-identificación de las cabeceras DICOM propuesta por RSNA

Data Element	Description	De-identification
(0008,0014)	InstanceCreatorUID	@hashuid(@UIDROOT,this)
(0008,0015)	InstanceCoercionDateTime	@hashdate(this,PatientID)
(0008,0018)	SOPInstanceUID	@hashuid(@UIDROOT,this)
(0008,0020)	StudyDate	@hashdate(this,PatientID)
(0008,0021)	SeriesDate	@hashdate(this,PatientID)
(0008,0022)	AcquisitionDate	@hashdate(this,PatientID)
(0008,0023)	ContentDate	@hashdate(this,PatientID)
(0008,0024)	OverlayDate	@hashdate(this,PatientID)
(0008,0025)	CurveDate	@hashdate(this,PatientID)
(0008,002A)	AcquisitionDatetime	@hashdate(this,PatientID)
(0008,0050)	AccessionNumber	@hash(this,8)
(0008,0058)	FailedSOPInstanceUIDList	@remove()
(0008,0080)	InstitutionName	@empty()
(0008,0081)	InstitutionAddress	@remove()
(0008,0082)	InstitutionCodeSeq	@remove()
(0008,0090)	ReferringPhysicianName	@empty()
(0008,0092)	ReferringPhysicianAddress	@remove()
(0008,0094)	ReferringPhysicianPhoneNumbers	@remove()
(0008,0096)	ReferringPhysicianIdentificationSeq	@remove()
(0008,009C)	ConsultingPhysicianName	@empty()
(0008,009D)	ConsultingPhysicianIdentificationSeq	@remove()
(0008,1010)	StationName	@empty()
(0008,1040)	InstitutionalDepartmentName	@remove()

(0008,1041)	InstitutionalDepartmentTypeCodeSeq	@remove()
(0008,1048)	PhysicianOfRecord	@remove()
(0008,1049)	PhysicianOfRecordIdentificationSeq	@remove()
(0008,1050)	PerformingPhysicianName	@remove()
(0008,1052)	PerformingPhysicianIdentificationSeq	@remove()
(0008,1060)	NameOfPhysicianReadingStudy	@remove()
(0008,1062)	PhysicianReadingStudyIdentificationSeq	@remove()
(0008,1070)	OperatorName	@empty()
(0008,1072)	OperatorIdentificationSeq	@remove()
(0008,1080)	AdmittingDiagnosisDescription	@remove()
(0008,1084)	AdmittingDiagnosisCodeSeq	@remove()
(0008,1110)	RefStudySeq	@remove()
(0008,1111)	RefPPSSeq	@remove()
(0008,1120)	RefPatientSeq	@remove()
(0008,1140)	RefImageSeq	@remove()
(0008,1155)	RefSOPInstanceUID	@remove()
(0008,1195)	TransactionUID	@remove()
(0008,2111)	DerivationDescription	@remove()
(0008,2112)	SourceImageSeq	@remove()
(0008,3010)	IrradiationEventUID	@remove()
(0008,4000)	IdentifyingCommentsRetired	@remove()
(0010,0010)	PatientName	@param(@SITEID)-@integer(PatientID,"ptid",6)
(0010,0020)	PatientID	
(0010,0021)	IssuerOfPatientID	@param(@SITEID)-@integer(PatientID,"ptid",6)
(0010,0030)	PatientBirthDate	@remove()
(0010,0050)	PatientInsurancePlanCodeSeq	@hashdate(this,PatientID)
(0010,0101)	PatientPrimaryLanguageCodeSeq	@remove()

(0010,0102)	PatientPrimaryLanguageModifierCodeSeq	@remove()
(0010,1000)	OtherPatientIDs	@remove()
(0010,1001)	OtherPatientNames	@remove()
(0010,1002)	OtherPatientIDSeq	@remove()
(0010,1005)	PatientBirthName	@remove()
(0010,1040)	PatientAddress	@remove()
(0010,1050)	InsurancePlanIdentificationRetired	@remove()
(0010,1060)	PatientMotherBirthName	@remove()
(0010,1080)	MilitaryRank	@remove()
(0010,1081)	BranchOfService	@remove()
(0010,1090)	MedicalRecordLocator	@remove()
(0010,1100)	RefPatientPhotoSeq	@remove()
(0010,2000)	MedicalAlerts	@remove()
(0010,2110)	ContrastAllergies	@remove()
(0010,2150)	CountryOfResidence	@remove()
(0010,2152)	RegionOfResidence	@remove()
(0010,2154)	PatientPhoneNumbers	@remove()
(0010,2155)	PatientTelecomInformation	@remove()
(0010,2180)	Occupation	@remove()
(0010,21B0)	AdditionalPatientHistory	@remove()
(0010,21D0)	LastMenstrualDate	@remove()
(0010,21F0)	PatientReligiousPreference	@hashdate(this,PatientID)
(0010,2297)	ResponsiblePerson	@remove()
(0010,2299)	ResponsibleOrganization	@remove()
(0010,4000)	PatientComments	@remove()
(0012,0010)	ClinicalTrialSponsorName	@remove()
(0012,0020)	ClinicalTrialProtocolID	@empty()

(0012,0021)	ClinicalTrialProtocolName	@empty()
(0012,0030)	ClinicalTrialSiteID	@empty()
(0012,0031)	ClinicalTrialSiteName	@empty()
(0012,0040)	ClinicalTrialSubjectID	@empty()
(0012,0042)	ClinicalTrialSubjectReadingID	@empty()
(0012,0050)	ClinicalTrialTimePointID	@empty()
(0012,0051)	ClinicalTrialTimePointDescription	@empty()
(0012,0060)	CoordinatingCenterName	@remove()
(0012,0062)	PatientIdentityRemoved	@empty()
(0012,0064)	DeIdentificationMethodCodeSeq (see table)	YES
(0012,0071)	ClinicalTrialSeriesID	113100/113101/113102/113103
(0012,0072)	ClinicalTrialSeriesDescription	@remove()
(0012,0081)	ClinicalTrialProtocolEthicsCommitteeName	@remove()
(0012,0082)	ClinicalTrialProtocolEthicsCommitteeApprovalNum	@empty()
(0016,002B)	MakerNote	@remove()
(0016,004B)	DeviceSettingDescription	@remove()
(0016,004D)	CameraOwnerName	@remove()
(0016,004E)	LensSpecification	@remove()
(0016,004F)	LensMake	@remove()
(0016,0050)	LensModel	@remove()
(0016,0051)	LensSerialNumber	@remove()
(0016,0070)	GPSVersionID	@remove()
(0016,0071)	GPSLatitudeRef	@remove()
(0016,0072)	GPSLatitude	@remove()
(0016,0073)	GPSLongitudeRef	@remove()
(0016,0074)	GPSLongitude	@remove()
(0016,0075)	GPSAltitudeRef	@remove()

(0016,0076)	GPSAltitude	@remove()
(0016,0077)	GPSTimeStamp	@remove()
(0016,0078)	GPSSatellites	@remove()
(0016,0079)	GPSStatus	@remove()
(0016,007A)	GPSMeasureMode	@remove()
(0016,007B)	GPSDOP	@remove()
(0016,007C)	GPSSpeedRef	@remove()
(0016,007D)	GPSSpeed	@remove()
(0016,007E)	GPSTrackRef	@remove()
(0016,007F)	GPSTrack	@remove()
(0016,0080)	GPSImgDirectionRef	@remove()
(0016,0081)	GPSImgDirection	@remove()
(0016,0082)	GPSMapDatum	@remove()
(0016,0083)	GPSDestLatitudeRef	@remove()
(0016,0084)	GPSDestLatitude	@remove()
(0016,0085)	GPSDestLongitudeRef	@remove()
(0016,0086)	GPSDestLongitude	@remove()
(0016,0087)	GPSDestBearingRef	@remove()
(0016,0088)	GPSDestBearing	@remove()
(0016,0089)	GPSDestDistanceRef	@remove()
(0016,008A)	GPSDestDistance	@remove()
(0016,008B)	GPSProcessingMethod	@remove()
(0016,008C)	GPSAreaInformation	@remove()
(0016,008D)	GPSDateStamp	@remove()
(0016,008E)	GPSDifferential	@remove()
(0018,1000)	DeviceSerialNumber	@remove()
(0018,1002)	DeviceUID	@empty()

(0018,1004)	PlateID	@hashuid(@UIDROOT,this)
(0018,1005)	GeneratorID	@remove()
(0018,1007)	CassetteID	@remove()
(0018,1008)	GantyID	@remove()
(0018,1009)	UniqueDeviceIdentifier	@remove()
(0018,100A)	UDISeq	@remove()
(0018,100B)	ManufacturerDeviceClassUID	@remove()
(0018,1030)	ProtocolName	@hashuid(@UIDROOT,this)
(0018,1400)	AcquisitionDeviceProcessingDescription	@remove()
(0018,2042)	TargetUID	@remove()
(0018,4000)	SeriesCommentsRetired	@hashuid(@UIDROOT,this)
(0018,700A)	DetectorID	@remove()
(0018,9185)	RespiratoryMotionCompensationTechniqueDescrip	@remove()
(0018,9367)	XRaySourceID	@remove()
(0018,9369)	SourceStartDateTime	@empty()
(0018,936A)	SourceEndDateTime	@hashdate(this,PatientID
(0018,9371)	XRayDetectorID	@hashdate(this,PatientID)
(0018,9373)	XRayDetectorLabel	@empty()
(0018,937B)	MultienergyAcquisitionDescription	@remove()
(0018,937F)	DecompositionDescription	@remove()
(0018,9424)	AcquisitionProtocolDescription	@remove()
(0018,9516)	StartAcquisitionDateTime	@remove()
(0018,9517)	EndAcquisitionDateTime	@hashdate(this,PatientID)
(0018,A003)	ContributionDescription	@hashdate(this,PatientID)
(0020,000D)	StudyInstanceUID	@remove()
(0020,000E)	SeriesInstanceUID	@hashuid(@UIDROOT,this)
(0020,0010)	StudyID	@hashuid(@UIDROOT,this)

(0020,0052)	FrameOfReferenceUID	@empty()
(0020,0200)	SynchronizationFrameOfReferenceUID	@hashuid(@UIDROOT,this)
(0020,3401)	ModifyingDeviceIDRetired	@hashuid(@UIDROOT,this)
(0020,3406)	ModifiedImageDescriptionRetired	@remove()
(0020,4000)	ImageComments	@remove()
(0020,9158)	FrameComments	@remove()
(0020,9161)	ConcatenationUID	@remove()
(0020,9164)	DimensionOrganizationUID	@hashuid(@UIDROOT,this)
(0028,1199)	PaletteColorLUTUID	@hashuid(@UIDROOT,this)
(0028,1214)	LargePaletteColorLUTUID	@hashuid(@UIDROOT,this)
(0028,4000)	PixelCommentsRetired	@hashuid(@UIDROOT,this)
(0032,0012)	StudyIDIssuer	@remove()
(0032,1020)	ScheduledStudyLocation	@remove()
(0032,1021)	ScheduledStudyLocationAET	@remove()
(0032,1030)	ReasonforStudy	@remove()
(0032,1032)	RequestingPhysician	@remove()
(0032,1033)	RequestingService	@remove()
(0032,1060)	RequestedProcedureDescription	@remove()
(0032,1066)	ReasonForVisit	@empty()
(0032,1067)	ReasonForVisitCodeSeq	@remove()
(0032,1070)	RequestedContrastAgent	@remove()
(0032,4000)	StudyComments	@remove()
(0034,0001)	FlowIdentifierSeq	@remove()
(0034,0002)	FlowIdentifier	@remove()
(0034,0005)	SourceIdentifier	@empty()
(0034,0007)	FrameOriginTimestamp	@empty()
(0038,0004)	RefPatientAliasSeq	@remove()

(0038,0010)	AdmissionID	@remove()
(0038,0011)	IssuerOfAdmissionIDRetired	@remove()
(0038,0014)	IssuerOfAdmissionIDSeq	@remove()
(0038,001E)	ScheduledPatientInstitutionResidence	@remove()
(0038,0020)	AdmittingDate	@remove()
(0038,0021)	AdmittingTime	@remove()
(0038,0040)	DischargeDiagnosisDescription	@remove()
(0038,0050)	SpecialNeeds	@remove()
(0038,0060)	ServiceEpisodeID	@remove()
(0038,0061)	IssuerOfServiceEpisodeIDRET	@remove()
(0038,0062)	ServiceEpisodeDescription	@remove()
(0038,0064)	IssuerOfServiceEpisodeIDSeq	@remove()
(0038,0300)	CurrentPatientLocation	@remove()
(0038,0400)	PatientInstitutionResidence	@remove()
(0038,0500)	PatientState	@remove()
(0038,4000)	VisitComments	@remove()
(0040,0001)	ScheduledStationAET	@remove()
(0040,0002)	SPSStartDate	@remove()
(0040,0003)	SPSStartTime	@remove()
(0040,0004)	SPSEndDate	@remove()
(0040,0005)	SPSEndTime	@remove()
(0040,0006)	ScheduledPerformingPhysicianName	@remove()
(0040,0007)	SPSDescription	@remove()
(0040,000B)	ScheduledPerformingPhysicianIdentificationSeq	@remove()
(0040,0010)	ScheduledStationName	@remove()
(0040,0011)	SPSLocation	@remove()
(0040,0012)	PreMedication	@remove()

(0040,0241)	PerformedStationAET	@remove()
(0040,0242)	PerformedStationName	@remove()
(0040,0243)	PerformedLocation	@remove()
(0040,0244)	PPSStartDate	@remove()
(0040,0245)	PPSStartTime	@remove()
(0040,0250)	PPSEndDate	@remove()
(0040,0251)	PPSEndTime	@remove()
(0040,0253)	PPSID	@remove()
(0040,0254)	PPSDescription	@remove()
(0040,0275)	RequestAttributesSeq	@remove()
(0040,0280)	PPSComments	@remove()
(0040,050A)	SpecimenAccessionNumber	@remove()
(0040,0512)	ContainerIdentifier	@remove()
(0040,0513)	IssuerOfTheContainerIdentifierSeq	@empty()
(0040,051A)	ContainerDescription	@remove()
(0040,0551)	SpecimenIdentifier	@remove()
(0040,0554)	SpecimenUID	@empty()
(0040,0555)	AcquisitionContextSeq	@remove()
(0040,0562)	IssuerOfTheSpecimenIdentifierSeq	@remove()
(0040,0600)	SpecimenShortDescription	@remove()
(0040,0602)	SpecimenDetailedDescription	@remove()
(0040,0610)	SpecimenPreparationSeq	@remove()
(0040,06FA)	SlideIdentifier	@remove()
(0040,1001)	RequestedProcedureID	@remove()
(0040,1002)	ReasonForTheRequestedProcedure	@remove()
(0040,1004)	PatientTransportArrangements	@remove()
(0040,1005)	RequestedProcedureLocation	@remove()

(0040,100A)	ReasonforRequestedProcedureCodeSeq	@remove()
(0040,1010)	NamesOfIntendedRecipientsOfResults	@remove()
(0040,1011)	IntendedRecipientsOfResultsIdentificationSeq	@remove()
(0040,1101)	PersonIdentificationCodeSeq	@remove()
(0040,1102)	PersonAddress	@remove()
(0040,1103)	PersonTelephoneNumbers	@remove()
(0040,1104)	PersonTelecomInformation	@remove()
(0040,1400)	RequestedProcedureComments	@remove()
(0040,2001)	ReasonForTheImagingServiceRequest	@remove()
(0040,2008)	OrderEnteredBy	@remove()
(0040,2009)	OrderEntererLocation	@remove()
(0040,2010)	OrderCallbackPhoneNumber	@remove()
(0040,2011)	OrderCallbackTelecomInformation	@remove()
(0040,2016)	PlacerOrderNumber	@remove()
(0040,2017)	FillerOrderNumber	@empty()
(0040,2400)	ImagingServiceRequestComments	@empty()
(0040,3001)	ConfidentialityPatientData	@remove()
(0040,4005)	SPSStartDateAndTime	@remove()
(0040,4008)	ScheduledProcedureStepExpirationDateTime	@remove()
(0040,4010)	SPSModificationDateandTime	@remove()
(0040,4011)	ExpectedCompletionDateAndTime	@remove()
(0040,4023)	RefGPSPSTransactionUID	@remove()
(0040,4025)	ScheduledStationNameCodeSeq	@remove()
(0040,4027)	ScheduledStationGeographicLocationCodeSeq	@remove()
(0040,4028)	PerformedStationNameCodeSeq	@remove()
(0040,4030)	PerformedStationGeographicLocationCodeSeq	@remove()
(0040,4034)	ScheduledHumanPerformersSeq	@remove()

(0040,4035)	ActualHumanPerformersSeq	@remove()
(0040,4036)	HumanPerformerOrganization	@remove()
(0040,4037)	HumanPerformerName	@remove()
(0040,4050)	PerformedProcedureStepStartDateTime	@remove()
(0040,4051)	PerformedProcedureStepEndDateTime	@remove()
(0040,4052)	ProcedureStepCancellationDateTime	@remove()
(0040,A027)	VerifyingOrganization	@remove()
(0040,A073)	VerifyingObserverSeq	@empty()
(0040,A075)	VerifyingObserverName	@remove()
(0040,A078)	AuthorObserverSeq	@empty()
(0040,A07A)	ParticipantSeq	@remove()
(0040,A07C)	CustodialOrganizationSeq	@remove()
(0040,A088)	VerifyingObserverIdentificationCodeSeq	@remove()
(0040,A123)	PersonName	@remove()
(0040,A124)	UID	@empty()
(0040,A171)	ObservationUID	@remove()
(0040,A172)	RefObservationUIDTrial	@remove()
(0040,A192)	ObservationDateTrial	@remove()
(0040,A193)	ObservationTimeTrial	@remove()
(0040,A307)	CurrentObserverTrial	@remove()
(0040,A352)	VerbalSourceTrial	@remove()
(0040,A353)	AddressTrial	@remove()
(0040,A354)	TelephoneNumberTrial	@remove()
(0040,A358)	VerbalSourceIdentifierCodeSequenceTrial	@remove()
(0040,A402)	ObservationSubjectUIDTrial	@remove()
(0040,A730)	ContentSeq	@remove()
(0040,DB0C)	TemplateExtensionOrganizationUID	@remove()

(0040,DB0D)	TemplateExtensionCreatorUID	@remove()
(0050,001B)	ContainerComponentID	@remove()
(0050,0020)	DeviceDescription	@remove()
(0050,0021)	LongDeviceDescription	@remove()
(0062,0021)	TrackingUID	@remove()
(0070,0001)	GraphicAnnotationSeq	@remove()
(0070,0084)	PresentationCreatorName	@remove()
(0070,0086)	ContentCreatorIdentificationSeq	@empty()
(0070,031A)	FiducialUID	@remove()
(0070,1101)	PresentationDisplayCollectionUID	@remove()
(0070,1102)	PresentationSequenceCollectionUID	@remove()
(0088,0140)	StorageMediaFileSetUID	@remove()
(0088,0200)	IconImageSeq	@remove()
(0088,0904)	TopicTitle	@remove()
(0088,0906)	TopicSubject	@remove()
(0088,0910)	TopicAuthor	@remove()
(0088,0912)	TopicKeyWords	@remove()
(0400,0100)	DigitalSignatureUID	@remove()
(0400,0402)	RefDigitalSignatureSeq	@remove()
(0400,0403)	RefSOPInstanceMACSeq	@remove()
(0400,0404)	MAC	@remove()
(0400,0550)	ModifiedAttributesSeq	@remove()
(0400,0561)	OriginalAttributesSeq	@remove()
(0400,0600)	InstanceOriginStatus	@remove()
(2030,0020)	TextString	@remove()
(2200,0002)	LabelText	@remove()
(2200,0005)	BarcodeValue	@empty()

(3006,0024)	RefFrameOfReferenceUID	@empty()
(3006,00C2)	RelatedFrameOfReferenceUID	@remove()
(3008,0054)	FirstTreatmentDate	@remove()
(3008,0056)	MostRecentTreatmentDate	@remove()
(3008,0105)	SourceSerialNumber	@remove()
(3008,0250)	TreatmentDate	@remove()
(3008,0251)	TreatmentTime	@remove()
(300A,0002)	RTPlanLabel	@remove()
(300A,0003)	RTPlanName	@remove()
(300A,0004)	RTPlanDescription	@remove()
(300A,0006)	RTPlanDate	@remove()
(300A,0007)	RTPlanTime	@remove()
(300A,000E)	PrescriptionDescription	@remove()
(300A,0013)	DoseReferenceUID	@remove()
(300A,0016)	DoseReferenceDescription	@remove()
(300A,0072)	FractionGroupDescription	@remove()
(300A,0083)	RefDoseReferenceUID	@remove()
(300A,00B2)	TreatmentMachineName	@remove()
(300A,00C3)	BeamDescription	@remove()
(300A,00DD)	BolusDescription	@remove()
(300A,0196)	FixationDeviceDescription	@remove()
(300A,01A6)	ShieldingDeviceDescription	@remove()
(300A,01B2)	SetupTechniqueDescription	@remove()
(300A,0216)	SourceManufacturer	@remove()
(300A,02EB)	CompensatorDescription	@remove()
(300A,0608)	TreatmentPositionGroupLabel	@remove()
(300A,0609)	TreatmentPositionGroupUID	@remove()

(300A,0611)	RTAccessoryHolderSlotID	@remove()
(300A,0615)	RTAccessoryDeviceSlotID	@remove()
(300A,0619)	RadiationDoseIdentificationLabel	@remove()
(300A,0623)	RadiationDoseInVivoMeasurementLabel	@remove()
(300A,062A)	RTToleranceSetLabel	@remove()
(300A,0650)	PatientSetupUID	@remove()
(300A,0676)	EquipmentFrameOfReferenceDescription	@remove()
(300A,067C)	RadiationGenerationModeLabel	@remove()
(300A,067D)	RadiationGenerationModeDescription	@remove()
(300C,0113)	ReasonForOmissionDescription	@remove()
(300E,0008)	ReviewerName	@remove()
(3010,0006)	ConceptualVolumeUID	@remove()
(3010,000B)	RefConceptualVolumeUID	@remove()
(3010,000F)	ConceptualVolumeCombinationDescription	@remove()
(3010,0013)	ConstituentConceptualVolumeUID	@remove()
(3010,0015)	SourceConceptualVolumeUID	@remove()
(3010,0017)	ConceptualVolumeDescription	@remove()
(3010,001B)	DeviceAlternateIdentifier	@remove()
(3010,002D)	DeviceLabel	@remove()
(3010,0031)	RefFiducialsUID	@remove()
(3010,0033)	UserContentLabel	@remove()
(3010,0034)	UserContentLongLabel	@remove()
(3010,0035)	EntityLabel	@remove()
(3010,0036)	EntityName	@remove()
(3010,0037)	EntityDescription	@remove()
(3010,0038)	EntityLongLabel	@remove()
(3010,003B)	RTTreatmentPhaseUID	@remove()

(3010,0043)	ManufacturerDeviceIdentifier	@remove()
(3010,004C)	IntendedPhaseStartDate	@remove()
(3010,004D)	IntendedPhaseEndDate	@remove()
(3010,0054)	RTPrescriptionLabel	@remove()
(3010,0056)	RTTreatmentApproachLabel	@remove()
(3010,005A)	RTPhysicianIntentNarrative	@remove()
(3010,005C)	ReasonForSuperseding	@remove()
(3010,0061)	PriorTreatmentDoseDescription	@remove()
(3010,006E)	DosimetricObjectiveUID	@remove()
(3010,006F)	RefDosimetricObjectiveUID	@remove()
(3010,0077)	TreatmentSite	@remove()
(3010,007A)	TreatmentTechniqueNotes	@remove()
(3010,007B)	PrescriptionNotes	@remove()
(3010,007F)	FractionationNotes	@remove()
(3010,0081)	PrescriptionNotesSeq	@remove()
(4000,0010)	ArbitraryRetired	@remove()
(4000,4000)	ArbitraryCommentsRetired	@remove()
(4008,0042)	ResultsIDIssuer	@remove()
(4008,0102)	InterpretationRecorder	@remove()
(4008,010A)	InterpretationTranscriber	@remove()
(4008,010B)	InterpretationText	@remove()
(4008,010C)	InterpretationAuthor	@remove()
(4008,0111)	InterpretationApproverSeq	@remove()
(4008,0114)	PhysicianApprovingInterpretation	@remove()
(4008,0115)	InterpretationDiagnosisDescription	@remove()
(4008,0118)	ResultsDistributionListSeq	@remove()
(4008,0119)	DistributionName	@remove()

(4008,011A)	DistributionAddress	@remove()
(4008,0202)	InterpretationIDIssuer	@remove()
(4008,0300)	Impressions	@remove()
(4008,4000)	ResultsComments	@remove()
(FFFA,FFFA)	DigitalSignaturesSeq	@remove()
	DataSetTrailingPadding	@remove()
		@remove()