

GUÍA PARA PROTOCOLO DE DATOS

Un **protocolo de datos** es un documento que describe cómo se gestionarán los datos a lo largo de un proyecto de investigación. Recuerde que el protocolo se debe usar durante todo el ciclo de vida del proyecto y ajustar según las necesidades del proyecto. Si el protocolo tiene cambios debe notificar al Comité de Ética.

El propósito del protocolo de datos es garantizar que los datos se recolecten, almacenen, procesen, usen y preserven de manera organizada y segura. Además, deben cumplir con principios éticos y normativos para asegurar su adecuado ciclo de vida.

El siguiente documento es una **guía** diseñada para ayudar a estudiantes e investigadores a estructurar su protocolo de datos. Puede ser adaptado según las necesidades específicas de cada proyecto, añadiendo o modificando secciones según corresponda.

Tenga en cuenta las siguientes definiciones:

- **Autorización:** Consentimiento previo, expreso e informado del Titular para llevar a cabo el Tratamiento de datos personales.
- **Dato personal:** Cualquier información vinculada o que pueda asociarse a una o varias personas naturales determinadas o determinables.
- **Dato sensible:** aquellos que afectan la intimidad del titular o cuyo uso indebido puede generar su discriminación, tales como aquellos que revelen el origen racial o étnico, la orientación política, las convicciones religiosas o filosóficas, la pertenencia a sindicatos, organizaciones sociales, de derechos humanos o que promueva intereses de cualquier partido político o que garanticen los derechos y garantías de partidos políticos de oposición así como los datos relativos a la salud, a la vida sexual y los datos biométricos, entre otros, la captura de imagen fija o en movimiento, huellas digitales, fotografías, iris, reconocimiento de voz, facial o de palma de mano, etc.
- **Nivel de confidencialidad:** Clasificación que determina el grado de acceso y protección requerido para un conjunto de datos. Puede incluir categorías como "Público", "Confidencial", "Uso Interno", "Restringido", entre otras.
- **Dato público:** Información que puede ser accedida y utilizada con pocas restricciones de uso o legales, y sin compromisos de confidencialidad. Por ejemplo, datos publicados en fuentes oficiales, directorios públicos o investigaciones con acceso abierto. Los datos públicos suelen tener restricciones de uso, es importante revisar.
- **Dato confidencial:** Información que debe protegerse debido a su naturaleza estratégica, legal o contractual. Solo puede ser accedida por personas autorizadas y bajo condiciones específicas. Puede incluir acuerdos comerciales, resultados de investigación no publicados y datos sujetos a acuerdos de no divulgación.
- **Dato anonimizado:** Información que ha sido procesada para eliminar cualquier referencia que permita identificar a la persona a la que pertenece. Un dato anonimizado no se considera un dato personal, siempre que no pueda ser reidentificado.
- **Dato seudonimizado:** Datos personales que han sido modificados para ocultar la identidad del titular, pero que aún pueden ser vinculados a este mediante información adicional (almacenada por separado).
- **Dato agregado:** Información combinada de múltiples fuentes de manera que no sea posible identificar a individuos específicos. Suele utilizarse en análisis estadísticos y estudios poblacionales.

Estructura esperada del protocolo de datos:

1. Información General del Proyecto

Esta sección proporciona un panorama general del proyecto de investigación. Incluye información clave como el título del estudio, los investigadores responsables, la institución que respalda la investigación y las fuentes de financiamiento (si aplica). Además, se definen los objetivos generales y específicos del estudio, estableciendo el propósito y alcance del trabajo.

Ejemplo:

Título del proyecto: Análisis de biomarcadores inflamatorios en pacientes con enfermedad renal crónica

Investigador(es) responsable(s):

Nombre Completo	Cargo/perfil
Maria Botero	Profesor Uniandes
Juan Perez	Estudiante pregrado Ingeniería Biomédica

Objetivo general: Evaluar la relación entre biomarcadores inflamatorios y la progresión de la enfermedad renal crónica en pacientes en estadio 3 y 4.

Objetivos específicos (se recomiendan máximo 3):

1. Identificar los niveles séricos de IL-6, TNF- α y PCR en pacientes con enfermedad renal crónica.
2. Comparar los biomarcadores inflamatorios entre pacientes con diferentes estadios de la enfermedad.
3. Determinar la correlación entre los niveles de biomarcadores y la tasa de filtración glomerular.

2. Propósito y Descripción de los Datos

En esta sección se especifican las características de los datos que se utilizarán en la investigación. Es importante definir el tipo de datos, su volumen estimado, el formato en el que serán almacenados y la fuente de donde provienen. Esto permite una planificación adecuada para el manejo, análisis y almacenamiento de los datos, asegurando su calidad y organización. Adicionalmente, justifique la necesidad de recolectar cada tipo de dato en función de los objetivos del estudio. Especialmente pregúntese la relevancia de solicitar datos personales y sensibles y, en la medida de lo posible, no solicite este tipo de datos.

Ejemplo (recuerde agregar columnas según la necesidad de su proyecto):

Nombre de la variable	Tipo de variable o archivo	Descripción	Formato / Unidad	Rango / Valores posibles	Número de registros	Fuente / Método de obtención	Notas adicionales
ID_paciente	Numérico	Identificador único de cada paciente	Entero	1 - 1000	1000	Registro clínico	Seudonimizado
Edad	Numérico	Edad del paciente	Años	0 - 100	1000	Encuesta	-
Diagnóstico	Catórica	Enfermedad diagnosticada	Texto	Diabetes, Hipertensión, etc.	1000	Historia clínica	Clasificación ICD-10
Archivo_ECG	Archivo (CSV)	Señal de electrocardiograma	CSV	-	500	Dispositivo médico	10 segundos de registro
Fecha_muestra	Fecha	Fecha de toma de la muestra	YYYY-MM-DD	2023-01-01 - 2024-12-31	1000	Laboratorio	-

3. Ciclo de vida de los datos: creación/recolección, procesamiento, almacenamiento, intercambio, uso y análisis de datos, eliminación/preservación, y reutilización

I. Creación y/o recolección

En esta sección, se debe detallar el proceso mediante el cual se obtendrán los datos utilizados en la investigación. Es fundamental describir cómo se capturarán los datos, qué herramientas se utilizarán y qué medidas se implementarán para asegurar su calidad y confiabilidad.

Se recomienda incluir la siguiente información (recuerde que debe adaptar el formato según las características de su proyecto):

- **Métodos y herramientas de recolección:** *Dispositivos (sensores, monitores fisiológicos, wearables, etc.), software (software de imagenología, LabChart, etc.) o instrumentos (cuestionarios físicos, cuestionarios OneDrive, entrevistas semiestructuradas, pruebas de laboratorio, exámenes clínicos, etc.).*
- **Procedimientos de verificación y control de calidad:** *Métodos y estrategias para minimizar errores en la recolección de los datos, tales como: uso de formatos estandarizados, protocolo experimental, calibración de equipos, pruebas de consistencia en bases de datos y análisis exploratorio de datos.*
- **Aspectos Éticos y de Consentimiento:** *Indique si los participantes darán su consentimiento para la recolección de datos y cómo se documentará este consentimiento (formato físico, digital, grabación). En caso de datos sensibles o personales, describa qué medidas se tomarán para garantizar la privacidad y confidencialidad desde la recolección.*

Ejemplo:

Métodos y herramientas de recolección:

- **Sensores:** Los datos fisiológicos se registrarán mediante sensores biomédicos conectados a una plataforma LabChart, con exportación en formato CSV para su posterior procesamiento. Los sensores serán calibrados antes de cada sesión y se aplicará un protocolo estandarizado para su uso.
- **Cuestionarios:** Se administrarán digitalmente mediante Microsoft Forms, con exportación en CSV. Las preguntas incluirán validaciones automáticas para minimizar respuestas erróneas o inconsistentes.

Procedimientos de verificación y control de calidad:

- **Validación automática:** Se implementarán reglas de validación en los cuestionarios digitales (campos obligatorios, valores esperados).
- **Evaluación de registros atípicos:** Se generarán resúmenes estadísticos y visualizaciones para detectar valores extremos o inconsistentes antes del análisis.

Aspectos éticos y consentimiento: Todos los participantes firmarán un consentimiento informado digital o físico antes de la recolección de datos. Se informará a los participantes sobre el propósito de la investigación, aspectos claves del ciclo de vida de los datos recolectados.

II. Tratamiento de datos personales, sensibles y anonimización

Si los datos incluyen información personal o sensible, debe detallar cómo serán protegidos, tratados y anonimizados antes de su almacenamiento, procesamiento y difusión. La falta de medidas adecuadas puede comprometer la privacidad de los participantes y generar riesgos legales y éticos.

Se recomienda incluir la siguiente información (recuerde que debe adaptar el formato según las características de su proyecto):

- **Identificación y Clasificación de los Datos:** Se debe identificar y clasificar los datos de acuerdo con su nivel de sensibilidad y confidencialidad.
- **Métodos de Protección y Anonimización:** Se deben aplicar técnicas de protección adecuadas según el tipo de datos recolectados.

Ejemplo:

Identificación y Clasificación de los Datos:

Nombre de la variable	Descripción	Dato sensible? (Sí/No)	Dato personal? (Sí/No)	Nivel de confidencialidad (Público, Confidencial, etc.)	Fuente Método / de obtención	Metodos de protección y anonimización
ID_paciente	Identificador único de cada paciente	No	Sí	Confidencial	Registro clínico	Seudonimizado
Edad	Edad del paciente	No	Sí	Público	Encuesta	-
Diagnóstico	Enfermedad diagnosticada	No	No	Confidencial	Historia clínica	Eliminación de Identificadores Directos
Archivo_ECG	Señal de electrocardiograma	No	No	Confidencial	Dispositivo médico	Eliminación de Identificadores Directos
Fecha_muestra	Fecha de toma de la muestra	No	No	Público	Laboratorio	-

Métodos de Protección y Anonimización (ejemplos):

- **Seudonimización:** Sustituir identificadores directos por códigos únicos sin relación aparente con la identidad real de la persona. Reemplazar nombres por códigos alfanuméricos generados aleatoriamente. Se va a llevar un archivo de correspondencia entre códigos y sujetos, protegido con cifrado y acceso restringido.
- **Eliminación de Identificadores Directos:** Remover nombres, cédulas, direcciones y cualquier otro dato que pueda identificar a un individuo.
- **Agregación y Generalización:** Reemplazar valores específicos por categorías amplias para evitar la identificación: en lugar de almacenar la edad exacta, registrar rangos de edad (20-30 años, 31-40 años). En la difusión de resultados no se comparte la base de datos completa, solo los resúmenes estadísticos y datos agregados.
- **Enmascaramiento y Cifrado:** Aplicable a datos altamente sensibles como registros médicos y datos financieros: utilizar cifrado AES-256 para proteger bases de datos almacenadas en servidores.
- **Anonimización de Datos Audiovisuales:** Distorsión de voz en grabaciones para evitar reconocimiento de los participantes. Desenfoque y pixelado de rostros en videos para eliminar características identificables.

III. Procesamiento y análisis

Durante el procesamiento y análisis de los datos, siempre se deben considerar las implicaciones éticas del proyecto, asegurando que el análisis sea transparente, estadísticamente robusto y libre de sesgos. Es fundamental minimizar los riesgos éticos asociados al manejo de la información y garantizar que los resultados reflejen con precisión la realidad de los datos sin inducir interpretaciones sesgadas.

Se recomienda incluir la siguiente información (recuerde que debe adaptar el formato según las características de su proyecto):

- **Estrategias de Limpieza de Datos:** Los datos recopilados pueden contener errores, valores atípicos o información incompleta. Describa los procedimientos específicos para manejar datos incompletos, duplicados o atípicos, indicando qué técnicas o herramientas se emplearán.
- **Descripción del análisis:** Describir las técnicas estadísticas, computacionales o cualitativas que se aplicarán a los datos, así como la justificación para su selección.
- **Documentación proceso de análisis:** Cómo se documentará el proceso de análisis para asegurar su reproducibilidad.

Ejemplo:

Los datos fisiológicos recopilados mediante sensores biomédicos se procesarán utilizando Python y bibliotecas como NumPy, Pandas y SciPy. El procedimiento de limpieza incluirá la detección y eliminación de registros incompletos o duplicados, así como la imputación de valores faltantes mediante interpolación lineal en registros de señales continuas. Para los valores atípicos, se utilizará el método de Tukey (rango intercuartil) y se documentará cualquier exclusión con justificación estadística.

Posteriormente, se aplicarán análisis descriptivos y análisis de series temporales (por ejemplo, transformada rápida de Fourier y análisis de componentes principales), con el fin de extraer patrones relevantes en las señales. La selección de estos métodos responde a la naturaleza temporal y multidimensional de los datos.

El proceso será documentado en un cuaderno digital reproducible (Jupyter Notebook), incluyendo scripts, parámetros utilizados y comentarios explicativos. Esto garantizará la trazabilidad del análisis y permitirá su verificación por otros investigadores. Las visualizaciones generadas se compartirán con el equipo para su validación y discusión, y se incluirán advertencias explícitas sobre los límites de inferencia de los resultados, evitando interpretaciones erróneas o sesgadas.

IV. Intercambio, publicación y usos de los datos

Durante el desarrollo del proyecto, los datos pueden ser compartidos o utilizados por distintas partes interesadas (instituciones, colaboradores, empresas, etc.). Esta sección debe describir con precisión las condiciones bajo las cuales los datos serán intercambiados, usados o publicados, diferenciando claramente entre datos sin anonimizar, datos anonimizados y datos agregados, cuando sea relevante.

Es importante tener en cuenta: (1) usos académicos, (2) usos no académicos (si aplica), (3) Restricciones de uso y difusión.

Ejemplo:

Uso académico y producción científica: Solo se utilizarán, compartirán y publicarán los datos una vez hayan sido anonimizados según los métodos y estrategias descritos en el presente documento. Los datos serán utilizados exclusivamente con fines académicos y de investigación. Servirán como base para la elaboración de publicaciones científicas, tesis de grado y otros trabajos académicos. Los hallazgos obtenidos a partir del análisis de los datos podrán presentarse en congresos.

Restricciones de uso y difusión: Los datos solo podrán ser compartidos con fines académicos y de investigación y solo bajo la firma de un **Acuerdo de Confidencialidad y Uso Adecuado de la Información**.

V. Almacenamiento y Seguridad

Esta sección detalla dónde y cómo se almacenarán los datos para garantizar su integridad y seguridad. Es importante definir la ubicación de almacenamiento, las medidas de seguridad implementadas para proteger los datos contra accesos no autorizados o pérdidas, las políticas de acceso y control de versiones para gestionar actualizaciones o modificaciones en los datos y la eliminación de los datos al final del proyecto.

Ejemplo:

Almacenamiento: Los datos se almacenarán en servidores institucionales de la Universidad, con acceso restringido a los investigadores del proyecto.

Medidas de seguridad: Todos los archivos que contengan información sensible estarán encriptados utilizando AES-256. Se establecerán permisos de acceso según roles: los investigadores podrán leer y modificar los datos, mientras que los asistentes solo tendrán acceso de lectura.

VI. Eliminación, Preservación y Reutilización

Se deben especificar los métodos o estrategias para la eliminación o conservación de los datos una vez finalizado el proyecto: (1) Criterios para la eliminación segura de datos sensibles (al cuanto tiempo después de finalizado el proyecto, el investigador responsable). (2) Condiciones para la preservación y reutilización de los datos en estudios futuros.

Ejemplo:

Una vez finalizado el proyecto, los datos sensibles (incluyendo datos personales y clínicos) serán conservados únicamente por el investigador principal durante un periodo máximo de 12 meses, con el fin de realizar ajustes o verificaciones posteriores al análisis. Pasado este periodo, los archivos que contengan información identificable serán eliminados de forma segura mediante técnicas de sobrescritura múltiple (borrado seguro) y eliminación definitiva de copias de respaldo, conforme a las políticas institucionales.

Los datos procesados y completamente anonimizados serán preservados en un repositorio. Estos datos podrán ser reutilizados exclusivamente con fines académicos o científicos, siempre que los nuevos estudios cuenten con la aprobación del comité de ética correspondiente.