

Norma técnica de anonimización para la publicación bases de datos como datos abiertos

Departamento de Estadísticas e Información de Salud

División de Planificación Sanitaria

Subsecretaría de Salud Pública

Ministerio de Salud



Responsables técnicos

Jorge Pacheco Jara

Jefe del Departamento de Estadísticas e Información de Salud (DEIS), Ministerio de Salud

Pamela Suárez Ojeda

Jefa de Oficina de Análisis Estadístico, Departamento de Estadísticas e Información de Salud (DEIS), Ministerio de Salud

José Luis Toro Peñailillo

Profesional de Oficina de Gestión de Datos, Departamento de Estadísticas e Información de Salud (DEIS), Ministerio de Salud

Tomás Bralic Muñoz

Profesional de Oficina de Estudios y Análisis Estadísticos Avanzados, Departamento de Epidemiología, Ministerio de Salud

Revisores

José Villa Catalán

Encargado de Ciberseguridad y Seguridad de la Información, Departamento de Tecnologías de la Información y Comunicación, Ministerio de Salud

Lorena Donoso Abarca

Abogada de División Jurídica, Ministerio de Salud

Ninoska Kroff Cortez

Analista de Gobernanza de Datos e IA, Gobierno Digital, Ministerio de Hacienda

René Lagos Barrios

Estudiante Doctorado de Salud Pública



Tabla de contenidos

Respo	onsables técnicos2
Revis	ores2
Tabla	de contenidos3
Int	roducción4
Ma	rco normativo5
De	finiciones6
Alc	ances8
Pro	cedimiento de anonimización de MINSAL8
1.	Roles y Responsabilidades8
2.	Planificación y diseño:8
3.	Implementación:9
4.	Verificación y Validación11
5.	Medidas de Seguridad11
Bib	liografía12
Ane	exo I: Formulario de anonimización13
Ane	exo II: Herramientas para apoyar el procedimiento de anonimización14
1)	Anonimización utilizando software ARX14
2)	Anonimización utilizando software R32



Introducción

Los organismos del Estado están facultados para tratar datos personales, sin el consentimiento de sus titulares, respecto de las materias de su competencia. En el caso del Ministerio de Salud, gran parte de los datos personales que son tratados, dicen relación con la salud de las personas, lo que corresponde, además, a un dato sensible que cuenta con la máxima protección en el marco regulatorio vigente. Para proteger los datos personales y sensibles se requiere de procesos institucionales de tratamiento de información donde se garantice un adecuado resguardo a la privacidad de las personas. Esto involucra que los equipos que tratan datos personales estén capacitados en los aspectos regulatorios, éticos y técnicos de esta tarea.

Asimismo, la información producida por las instituciones públicas, o que exista en poder de éstas, tiene interés para la ciudadanía y debe cumplir con los principios de publicidad y transparencia. Los datos abiertos corresponden a una estrategia de transparencia activa donde los datos se ponen a disposición con las características técnicas y jurídicas necesarias para que puedan ser usados, reutilizados y redistribuidos libremente por cualquier persona, en cualquier momento y en cualquier lugar (Open Knowledge, 2015). Sin embargo, para cumplir con el mandato de transparentar la información institucional, el Ministerio de Salud debe previamente aplicar técnicas de anonimización que permitan poner a disposición los datos de salud, resguardando la privacidad de las personas. Los datos abiertos contribuyen a la innovación, la toma de decisiones informadas, la colaboración intersectorial y el avance del conocimiento científico y tecnológico.

Durante los últimos años, se han desarrollado nuevos modelos y softwares destinados a esta tarea. Por ejemplo, la criptografía ha desarrollado nuevos algoritmos más seguros para cifrar la información personal. Asimismo, se han generado nuevas técnicas de anonimización como la k-anonimidad, l-diversidad, privacidad diferencial, entre otras. Asociado a estos desarrollos técnicos, los marcos regulatorios han ido evolucionando para hacer frente al incremento de datos disponibles producto de la transformación digital, con énfasis en la ciberseguridad y la protección de datos personales.

Este documento busca regular la aplicación de técnicas de **anonimización en la divulgación de datos abiertos**. En este caso de uso particular, la libre utilización de los datos abiertos determina la necesidad de aplicar protocolos estrictos que garanticen la privacidad de las personas asumiendo un elevado riesgo de re-identificación (Information Commissioner's Office, 2012). Este riesgo de identificación considera: (1) la singularización, definida como el riesgo de identificar a los individuos mediante registros específicos o combinaciones de atributos en un conjunto de datos, (2) la vinculabilidad, definida como la posibilidad de relacionar dos o más registros con una misma persona, ya sea dentro del mismo conjunto de datos o entre diferentes conjuntos de datos y (3) la inferencia, definida como la posibilidad de deducir detalles específicos sobre una persona utilizando los datos disponibles.

Para este propósito se revisará el marco normativo vigente y las definiciones técnicas utilizadas en los procesos de seudoanonimización y anonimización. Posteriormente, se presentarán los algoritmos de k-anonimidad y l-diversidad con dos ejemplos de su utilización. En ambos casos se utilizarán softwares libres: ARX – Data anonymization Tool (Prasser et al, 2020) y R.

Marco normativo

A partir de la modificación del año 2018, la **Constitución Política de la República** consagra el **derecho a la protección de datos personales**. En el artículo 19 N°4, se establece que *"La Constitución asegura a todas las personas: [...]* 4°. El respeto y protección a la vida privada y a la honra de la persona y su familia, y asimismo, la protección de sus datos personales. El tratamiento y protección de estos datos se efectuará en la forma y condiciones que determine la ley".

Por su parte, la **ley 19.628 sobre protección de la vida privada** es la que regula el tratamiento de los datos personales. En esta ley se definen como datos personales *"los relativos a cualquier información concernientes a personas naturales, identificadas o identificables"*. Una persona puede identificarse de **manera directa**, mediante el uso identificadores como el RUN o la biometría, o de **manera indirecta**, mediante la combinación de cuasi-identificadores como el sexo, edad, lugar y fecha de nacimiento. Se considera que una persona es identificable cuando **el esfuerzo de determinación no resulta excesivo o desproporcionado**. Cuando un dato, desde su origen o a consecuencia de su tratamiento, no puede ser asociado a un titular identificado o identificable, se llama **dato estadístico**.

Esta misma ley define como **datos sensibles** a un subgrupo de datos personales que se refieren a las características físicas o morales de las personas o a hechos o circunstancias de su vida privada o intimidad, indicando, a modo de ejemplo, el estado de salud físico o psíquico como un dato sensible de una persona. En el artículo 10 de la ley 19.628 se establece **una prohibición general de tratamiento de datos personales sensibles, salvo cuando exista una disposición legal que lo autorice, exista consentimiento del titular o sean datos necesarios para la determinación u otorgamiento de beneficios de salud que correspondan a sus titulares**. Esta restricción rige tanto para organismos públicos como para privados. Siendo así, siempre que se pretenda realizar un tratamiento de datos sensibles, deberá analizarse previamente si existe una ley que autorice su tratamiento, o, en su defecto, si se cuenta con el consentimiento del titular o si son necesarios para la determinación o el otorgamiento de un beneficio de salud que corresponda al titular de los datos de que se trate.

En el caso de los organismos públicos, la ley N° 19.628, prevé que éstos podrán efectuar tratamientos de datos personales dentro de la órbita de sus competencias y sujetándose a las reglas previstas en esta ley, y en estas condiciones no necesitará el consentimiento del titular. En este sentido, y para el caso del Ministerio de Salud, el artículo 4 numeral 5 del DFL 1, de 2005, del Ministerio de Salud,¹ lo faculta para tratar datos personales y datos sensibles, con el fin de proteger la salud de la población o para la determinación y otorgamiento de beneficios de salud, como también se le faculta para tratar datos con fines estadísticos y mantener registros o bancos de datos respecto de las materias de su competencia.

A su vez, se debe tener presente lo dispuesto por la Ley N° 20.285 sobre acceso a la información pública, que establece que "toda persona tiene derecho a solicitar y recibir información de cualquier órgano de la Administración del Estado, en la forma y condiciones que establece la Ley". Esto significa que las personas pueden acceder a todos los antecedentes contenidos en "actos, resoluciones, actas, expedientes, contratos y acuerdos, así como a toda información elaborada con presupuesto público, cualquiera sea el formato o soporte en que se contenga, salvo las excepciones

¹ Artículo 4, numeral 5.- "Tratar datos con fines estadísticos y mantener registros o bancos de datos respecto de las materias de su competencia. Tratar datos personales o sensibles con el fin de proteger la salud de la población o para la determinación y otorgamiento de beneficios de salud. Para los efectos previstos en este número, podrá requerir de las personas naturales o jurídicas, públicas o privadas, la información que fuere necesaria. Todo ello conforme a las normas de la ley N° 19.628 y sobre secreto profesional."

legales". Este acceso puede ser realizado a través de una solicitud hacia la institución (transparencia pasiva) o de manera abierta y voluntaria (transparencia activa).

Sin perjuicio de esto, y en atención a que los registros de salud contienen datos sensibles, el cumplimiento de las obligaciones de transparencia se debe compatibilizar con el derecho fundamental de protección de datos personales. Es por esto que, el Consejo para la Transparencia en su Oficio N° E7986, del 10 de mayo del 2022, recomendó a la Subsecretaría de Salud Pública que **la publicación proactiva de registros de salud se realizara utilizando un procedimiento de anonimización o disociación de datos**. Es decir que, para su publicación, se aplicara en estos registros un *"procedimiento irreversible en virtud del cual un dato personal no pueda vincularse o asociarse a una persona determinada, ni permitir su identificación, por haberse destruido o eliminado el nexo con la información que vincula, asocia o identifica a esa persona"*. La aplicación de esta técnica permitiría obtener datos estadísticos y como tales tendrían el carácter de información pública².

A lo anterior se suma lo establecido en la Resolución Exenta N° 1465, del 3 de noviembre del 2023, que **Aprueba Política General de Seguridad de la Información y Ciberseguridad del Ministerio de Salud** donde se establece como principio adherido por la máxima autoridad de *"Proteger la privacidad y confidencialidad de toda información sensible o personal, independiente de su formato o medio de almacenamiento, a fin de respetar los derechos individuales y la integridad de los datos". Asimismo, esto se ve reforzando en las instrucciones impartidas a través del Ordinario A22/N°3681 del 17.11.2021 sobre Directrices de Seguridad de la Información y Ciberseguridad para el Sector y la Resolución Exenta N°785 /2021 que aprueba dicho instructivo. En específico en el <i>"Lineamiento sobre Tratamiento de datos sensibles para uso en nube y contratos relacionados Marco normativo de la nube"* y medidas técnicas sobre seudonimización de los datos. Asimismo, refuerza lo indicado en el Ordinario A22/N°2385 de 07.07.2023 y Resolución Exenta N°698/2023 que instruye la incorporación de **Cláusulas de Seguridad para Contratos de Tecnologías de la Información y Comunicación del Sector Salud** que suma como requerimiento la seudonimización en los ambientes de desarrollo y prueba, así como en el tratamiento de datos en nube.

Definiciones

- Datos abiertos son datos que pueden usarse, reutilizarse y redistribuirse libremente por cualquier persona, y que se encuentran sujetos al requerimiento de atribución y de compartirse igual que aparecen (Open Knowledge, 2015).
- Datos personales: datos relativos a cualquier información concerniente a personas naturales, identificadas o identificables.
- Datos sensibles: aquellos datos personales que se refieren a las características físicas o
 morales de las personas o a hechos o circunstancias de su vida privada o intimidad, tales como
 los hábitos personales, el origen racial, las ideologías y opiniones políticas, las creencias o
 convicciones religiosas, los estados de salud físicos o psíquicos y la vida sexual (Ley. 19.628).
- Técnicas de Anonimización:

² Disponible en:

https://repositoriodeis.minsal.cl/ContenidoSitioWeb2020/EstandaresNormativa/Oficio%20E7986,%2010.05.20 22.%20Emite%20pronunciamiento.%20Subs.%20Salud%20Pu%CC%81blica.%20Transparencia%20Proactiva.pd f

- Anonimización: Procedimiento en virtud del cual un dato personal no puede vincularse o asociarse a una persona determinada, ni permitir su identificación, por haberse destruido o eliminado el nexo con la información que vincula, asocia o identifica a esa persona. Un dato anonimizado deja de ser un dato personal.
- Seudoanonimización: es el tratamiento de datos personales de manera tal que ya no puede atribuirse a un titular sin utilizar información adicional, siempre que dicha información adicional figure por separado y esté sujeta a medidas técnicas y organizativas destinadas a garantizar que los datos personales no se atribuyan a una persona física identificada o identificable. Este proceso es reversible, en cuanto que, al juntar información adicional con los datos personales seudoanonimizados, se podrá volver a atribuir ese dato a una persona identificada o identificable. Este proceso se denomina reidentificación.
- Ofuscamiento o enmascaramiento: Corresponde a la modificación de los valores en un conjunto de datos. Estos valores pueden ser suprimidos o cambiados por información similar.
- **Generalización:** Reducción de la precisión de los datos (por ejemplo, agrupar edades en rangos).
- Supresión: Eliminación de datos directamente identificables.
- **Perturbación de Datos:** Introducción de ruido en los datos para dificultar la reidentificación
- Tipos de categorías de variables para procedimiento de anonimización
 - Identificador explícito (IdE): Todos aquellos atributos/características que identifican de manera directa a una persona. Por ejemplo: Run, Nombre, Teléfono, Correo electrónico, entre otros. Estos atributos no deben exponerse en el conjunto de datos.
 - Cuasi identificador (QId): Todos aquellos atributos/características que en su conjunto pueden identificar de manera única a una persona. Ej: Fecha de nacimiento, sexo, región, edad, etc. En este punto es necesario considerar no sólo la información que se presenta directamente en la base en tratamiento, sino también considerar todas aquellas bases de información que puede ser encontradas en la web, como información filtrada y expuesta o información publicada por otros organismos y que no se les ha aplicado algún tratamiento de anonimización.
 - Atributos Sensible (As): Información que su valor se desea proteger y que corresponden a características físicas o morales de una persona, como también a hechos o circunstancia de la vida privada o intimidad. Por ejemplo: etnia, salud física o psicológica, hábitos, ideologías o creencias. Estos atributos deben ser tratados para llevar a la l-diversidad >= 2 o bien ser ofuscados.
- **K-anonimidad**: Corresponde a cuantas veces (k) es lo mínimo que está repetido un conjunto de atributos identificados como cuasi-identificadores en el conjunto de datos. Un set de datos k-anónimos debería tener un k>1 e idealmente >3. Si el set de datos tiene k=1 en anonimidad, se requiere aplicar agrupaciones o transformaciones a algunas características para aumentar la misma.
- L-diversidad: Corresponde a cuantos valores distintos (I) existen en un atributo/característica sensible de una misma tupla de cuasi-identificadores. Si no se puede alcanzar al menos la 2-diversidad, por protección se sugiere ofuscar con un asterisco el atributo.
- **Tupla:** En el contexto de las bases de datos relacionales, se refiere a un único registro o fila dentro de una tabla que contiene un conjunto específico de valores para cada atributo definido por el esquema de la tabla. Si el conjunto de valores en la fila sólo se presenta en una ocasión, se denomina tupla única.

Alcances

Esta norma técnica se refiere al proceso de anonimización de datos estructurados en bases de datos para su publicación como datos abiertos en el sector salud y abarca de manera integral y transversal todos los procesos de anonimización, de la Subsecretarías de Salud Pública y Redes Asistenciales, Servicios de Salud, Secretarías Regionales Ministeriales de Salud y Establecimientos relacionados. En esta versión de esta norma técnica no se aborda el procedimiento de anonimización de datos no estructurados como son los textos libres, imágenes o videos contenidos en los registros clínicos. Asimismo, esta norma no aborda el uso de datos personales en otros ámbitos institucionales como la atención clínica directa, la gestión de programas de salud o la vigilancia epidemiológica ni los mecanismos para el intercambio de estos datos. Para estos propósitos se deben seguir las directrices vigentes establecidas en la Política General de Seguridad de la Información y Ciberseguridad del Ministerio de Salud.

Procedimiento de anonimización de MINSAL

1. Roles y Responsabilidades

- Responsables del procedimiento de anonimización: es responsable de la aplicación del procedimiento de anonimización el responsable del registro o banco de datos, la persona natural o jurídica privada, o el respectivo organismo público, a quien compete las decisiones relacionadas con el tratamiento de los datos de carácter personal. Corresponde a las jefaturas de unidades, departamentos o establecimientos donde se realiza el procedimiento y tendrán el rol de supervisar la anonimización y su resultado.
- Equipos o Áreas resolutoras: Corresponde a los estadísticos o ingenieros de datos de departamentos de estadísticas, estudios o análisis de datos de las Unidades, Departamentos o Establecimientos que realicen el tratamiento de registros o bancos de datos que contengan información de carácter personal. Tendrán el rol operativo de realizar el procedimiento y documentarlo, según lo mencionado en esta norma técnica.

2. Planificación y diseño:

Para planificar el proceso de anonimización, primero se debe reconocer la existencia de una base de datos que contenga información relevante para ser publicada como datos abiertos y que contenga datos personales. Esta base de datos debe ser de interés público. Se recomienda que la base de datos a tratar se haya validado previamente en su estructura y contenido y que, si corresponde a una serie histórica, haya sido homologada para todo el período. Asimismo, la base de datos debe contar con un diccionario de datos completo que describa todas las variables y su significado.

En segundo lugar, se debe definir la unidad y los funcionarios de la institución responsables del procedimiento de anonimización. Los responsables del procedimiento deben conocer cabalmente la información contenida en la base de datos. Esta unidad debe especificar la información a publicar,

la fuente de los datos y el período considerado. En esta etapa se debe identificar cuáles variables se publicarán según el propósito de la publicación. Para esto se debe valorar si los datos disponibles son relevantes para el caso de uso particular. Aquellos campos donde no se reconozca su relevancia, deben eliminarse para minimizar la información disponible (minimización de datos).

En tercer lugar, se debe establecer las técnicas de anonimización que serán aplicadas. Este equipo deberá documentar el proceso realizado, verificar que el procedimiento de anonimización se aplicó de manera correcta y que el riesgo de identificación es mínimo. Asimismo, se deberá revisar periódicamente la aplicación de los procesos de anonimización con la actualización de los registros.

3. Implementación:

Una vez que se definió la información a publicar, la fuente de los datos y el período considerado, la unidad responsable de la anonimización debe identificar la naturaleza de las variables que se deben tratan reconociendo si son identificadores explícitos (IdE), cuasi-identificadores (QId), atributos sensibles (As) y atributos no sensibles (Ans). Hay que reiterar que no es necesario publicar todas las variables de la base de datos, sino solo las que cumpla el propósito definido. Es decir, se debe realizar una minimización de datos.

A continuación, se ejemplificarán cada una de las categorías mencionadas anteriormente:

- Identificadores explícitos. Son datos que permiten identificar de forma inequívoca a una persona como el nombre, número de identificación nacional (por ejemplo, RUN o DNI), número de pasaporte u otro³, correo electrónico, número de teléfono móvil.
- Cuasi-identificadores. Son datos que no permiten una identificación directa del individuo, pero que en conjunto con otros datos pueden llegar a señalar a la persona como sexo, género, fecha de nacimiento, edad, ocupación, estado civil, nacionalidad, lugar de atención, comuna, región, previsión, fecha de egreso, entre otros.
- Atributos sensibles: Son datos que revelan características físicas o morales y que pueden comprometer la privacidad de los individuos como los diagnósticos, procedimientos clínicos, estado de vacunación, entre otros.
- Atributos no sensibles: Son datos que no comprometen la privacidad de los individuos como el rubro de trabajo. Esta información habitualmente no está presente en los registros de salud.

Una vez realizada esta tarea, se aplican técnicas para cada variable. La primera técnica que se aplica a la base de datos es la desidentificación que consiste en la eliminación de los identificadores explícitos de la base de datos. Muchas veces se confunde la desidentificación con la anonimización, pero la desidentificación es sólo una técnica de un conjunto de procedimientos a aplicar a la base de datos y, aplicada por si sola, genera un conjunto de datos que puede ser identificable al combinarlos con otros datos de acceso público. Este proceso se denomina reidentificación y es lo que se busca evitar con la anonimización. Una base desidentificada no debe contener ninguna de las siguientes variables: RUN u otro número de identificador personal, nombre y nombre social, teléfono, dirección particular, dirección laboral, correo electrónico, usuarios o contraseñas, usuarios de redes sociales o

³ Corresponden a números de identificación personal: RUN, N° de pasaporte, Cédula o DNI del país de origen, NIP (Número de identificación provisorio asignado por FONASA), IPE/IPA (Número de identificación provisorio asignado por MINEDUC), NIC (Número de identificación asignado por AFP para cotizar).

páginas web personales o cualquier otra información que permita identificar directamente a una persona.

Si la base de datos contiene variables registradas en texto libre pueden existir identificadores explícitos contenidos en estos campos que no hayan sido reconocidos. Esta posibilidad aumenta cuando el volumen de información es masivo, por lo que estas variables también deben eliminarse previo al proceso de publicación como datos abiertos, mientras no se pueda asegurar la anonimización del campo.

Una vez desidentificada la base de datos se procede a transformar los cuasi-identificadores para lograr que no existan tuplas únicas. Es decir, que no exista una combinación única de cuasiidentificadores entre todas las filas de la base de datos. El objetivo de este procedimiento es obtener una k-anonimidad mayor a 1, donde el valor K corresponde a cuantas veces es lo mínimo que está repetido una tupla de cuasi-identificadores en el conjunto de datos. Para lograr esto se realiza una técnica llamada generalización que consiste en limitar la precisión de los datos a través del establecimiento de una jerarquía en la que ciertos atributos del mismo grupo comparten valores.

A modo de ejemplo, se presenta la tabla 1 donde las variables sexo, edad, país de origen, comuna y previsión corresponde a cuasi-identificadores. Para obtener una K-anonimidad mayor a 2 se aplicó una generalización a las variables edad y país de origen. En el caso de la edad se transformó a rangos etarios y en el caso del país de origen se trató de manera dicotómica (chileno y no chileno).

Sexo	Edad	País de origen	Comuna	Previsión	Diagnóstico 1	Días de estada	Intervención principal
Hombre	30 a 39	Chileno	Puerto Montt	FONASA	G402	2	
Hombre	30 a 39	Chileno	Puerto Montt	FONASA	1213	8	
Hombre	30 a 39	Chileno	Puerto Montt	FONASA	K810	7	Colecistectomía por videolaparoscopía, proc. completo
Hombre	30 a 39	Chileno	Puerto Montt	FONASA	K810	2	Colecistectomía por videolaparoscopía, proc. completo
Hombre	30 a 39	Chileno	Puerto Montt	FONASA	S128	9	Estenosis laringotraqueales y/o faríngeas, trat. quir.

Tabla 1.- Tupla con K-anonimidad con K=5 y L-Diversidad con L=4

Otra opción es enmascarar las variables para lograr una K-anonimidad mayor a 2. A modo de ejemplo, se puede construir una agrupación territorial a nivel de comuna con 5 dígitos donde los primeros 2 dígitos corresponden a la región, el tercer dígito corresponde a provincia y los últimos dos dígitos corresponden a comuna. En este caso, 05302 corresponde a la Región de Valparaíso (05), la Provincia de Los Andes (3) y la comuna de Calle Larga (02). Si existe una tupla única para la comuna de Calle larga se puede enmascarar el código reemplazando esta agregación territorial por un asterisco, obteniendo el siguiente código 053**. En este caso, se desconoce la comuna, pero se cuenta con información de la provincia y la región.

Una vez obtenida una K-anonimidad igual o mayor a 2, se debe evaluar la L- diversidad que corresponde al número de valores distintos de los atributos sensibles que existen en una misma tupla única de cuasi-identificadores. Esto se debe a que si una tupla repetida K veces tiene el mismo atributo sensible se puede identificar en la base de datos. Al igual que en la K-anonimidad, el valor de la L-diversidad debe ser mayor a 1. Como ejemplo, en la tabla 1 se observa una L-diversidad para el atributo sensible Diagnóstico de 4, ya que hay 4 diagnósticos diferentes en los 5 casos presentados, y una L-diversidad de 3 para la Intervención principal.

El procedimiento anteriormente realizado se denomina anonimización utilizando la técnica de k-anonimidad y l-diversidad. Tal como se puede observar durante el proceso existe una reducción de la utilidad de la base de datos publicada debido a que, al transformar o enmascarar las variables, se

reduce el detalle de la información afectando los análisis que se realicen. Las personas responsables de la anonimización deben evaluar el equilibrio entre utilidad y privacidad identificando el número de registros ofuscados y el porcentaje de pérdida de información en la base de datos resultante.

Es importante señalar que la efectividad de los métodos de anonimización depende de la información disponible. Si esta información aumenta en el tiempo, lo que hoy no se considera un cuasi-identificador, podría convertirse en uno en el futuro. En este sentido, las técnicas de anonimización, en general, no garantizan sus propiedades formales de manera permanente en el tiempo y deben ser reevaluadas periódicamente de acuerdo con la información que se encuentra públicamente disponible.

4. Verificación y Validación

- Para verificar que la base de datos ha sido correctamente anonimizada, un funcionario del mismo departamento u otro departamento que tenga competencias en análisis de datos revisará el procedimiento y la base de datos resultante.
- Para el propósito de esta norma técnica, se definirá una base de datos como correctamente anonimizada si no cuenta con identificadores explícitos ni textos libres y presenta una Kanonimidad mayor o igual a 2 y una L-diversidad mayor o igual a 2 para los cuasiidentificadores.
- En caso de información que requiera protección adicional por su contenido particularmente sensible, se debe utilizar una K-anonimidad superior a 2.
- Para validar que el riesgo de re-identificación es mínimo, se realizará la "prueba del intruso motivado". Esta prueba la realiza una persona sin conocimientos previos, pero interesada en identificar a los individuos de la base de datos anonimizada. Esta aproximación considera que la persona es razonablemente competente y tiene acceso a internet, bibliotecas u otros documentos públicos para el propósito de identificar personas, incluidas otras bases de datos publicadas por instituciones públicas. Asimismo, esta prueba no asume que la persona tenga conocimientos especializados en informática o investigación criminal.
- Previo a la publicación se debe consultar al Departamento de Estadísticas e Información de Salud al correo: <u>deis@minsal.cl</u>

5. Medidas de Seguridad

- Realizar auditorías y revisiones periódicas de la aplicación de procesos de anonimización, registros y resultados.
- Documentar el proceso de anonimización y mantener registros de las decisiones tomadas según formulario de Anexo I.
- Capacitar al personal en técnicas de anonimización y en la importancia de la protección de datos personales.



Bibliografía

- 1. Constitución Política de la República de Chile. Disponible en: https://www.bcn.cl/leychile/navegar?idNorma=242302
- 2. Ley 19.628 sobre protección de la vida privada. Disponible: https://www.bcn.cl/leychile/navegar?idNorma=141599&idVersion=2023-05-09&idParte=
- 3. Ley 20.285 sobre acceso a la información pública. Disponible en: https://www.bcn.cl/leychile/navegar?idNorma=276363
- 4. Open Knowledge (2015). The open data handbook. Disponible en: https://opendatahandbook.org/
- 5. Tratamiento y Protección de Datos UC. Cuidando el tratamiento de los datos al interior de la UC. Disponible en: https://protecciondedatos.uc.cl/politica/definiciones
- 6. Agencia Española de Protección de Datos (2022). Guía básica de anonimización. Disponible en: https://www.aepd.es/documento/guia-basica-anonimizacion.pdf
- 7. Information Commissioner's Office (2012). Anonymisation: managing data protection risk code of practice. Disponible en: https://ico.org.uk/media/1061/anonymisation-code.pdf
- 8. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Disponible en: https://www.https//www.https://www.https://www.https//www
- Prasser F, Eicher J, Spengler H, Bild R, Kuhn KA (2020). Flexible data anonymization using ARX— Current status and challenges ahead. Software: Practice and Experience. 50: 1277–1304. https://doi.org/10.1002/spe.2812



Anexo I: Formulario de anonimización⁴

Fecha:	:
Departamento responsable de la base de datos	t
Encargados/as del procedimiento de anonimización	3

La base de datos cuenta con (responder si/no):

Acto administrativo que aprueba registro	 Validación de datos	
Diccionario de variables	Homologación de datos	

Propósito de publicación como datos abiertos

Listado de variables que serán publicados y su naturaleza

Variable	Identificador	Cuasi-	Atributo	Atributo no
	explícito	identificador	sensible	sensible
	_			

Describa brevemente la técnica de anonimización aplicada:

Realizó prueba del intruso motivado y hallazgos	
Firma del jefe de Depto responsable de la base de datos	i
Firma del Encargados/as del procedimiento de anonimización	:
Firma del Encargados/as de verificar correcta anonimización	
Fecha de publicación como datos abiertos	
Periodicidad de actualización	·

Anexo II: Herramientas para apoyar el procedimiento de anonimización.

Existe múltiples herramientas para realizar los procedimientos de anonimización. Una opción es utilizar el software libre ARX. Esta herramienta es flexible ya que soporta una amplia variedad de modelos de privacidad y permite aplicar diversos métodos para transformar los datos y analizar la privacidad y utilidad del resultado. Este software puede ser descargado en la página web: https://arx.deidentifier.org/.

Otra opción para realizar el procedimiento de anonimización es utilizar el software libre R. En este caso no se utilizó un paquete específico, sino que se creó un código para la desidentificación de la base de datos y el enmascaramiento de distintas variables para lograr una K-Anonimidad y L-Diversidad mayor o igual a 2. Este software puede ser descargado en la página web: <u>https://cran.r-project.org/bin/windows/base/.</u>

1) Anonimización utilizando software ARX

1. Creación de proyecto

Se busca en la barra de menú "File" y se hace clic en "New Project".



Se asigna nombre al proyecto, en este caso EH_2020.

Se crea el proyecto mostrando el nombre en la barra principal de la ventana.

2. Importación de datos

Se importan los datos a utilizar a través del menú "File", opción "Import data".



Open project Save project Save project as Import data Export data Create certificate Import hierarchy Export hierarchy	1-0	New project	1 . Č
Save project Save project as Import data Export data Create certificate Import hierarchy Export hierarchy	4	Open project	. B
Save project as Import data Export data Create certificate Import hierarchy Export hierarchy	4	Save project	
Import data Export data Create certificate Import hierarchy Export hierarchy	d	Save project as	
Export data Create certificate Import hierarchy Export hierarchy	*	Import data	
Create certificate Import hierarchy Export hierarchy		Export data	
Import hierarchy Export hierarchy		Create certificate	
Export hierarchy		Import hierarchy	
		Export hierarchy	

Se despliega una ventana donde se selecciona el tipo de origen de los datos, en este caso es un archivo CSV.

💱 Import data		D X
Source Select the source you want to impo	ort data from	+
 ● ESY ○ Excel (XLS, XLSX) ○ Database (JDBC) 		
	= Saus: Ne	xt > Cancel

En la siguiente pantalla se busca el archivo en la ruta respectiva haciendo clic en "Browse..."

♥ Import data CSV			
Please provide the information requeste	d below		Browse
	(Parts	Navi s	Cancel

Se selecciona el archivo correspondiente.



1 ** Pruel	oa ARX 🔸 Archivos Anuales	👻 🙋 🔗 Buscar en Arc	hivos Anueles
rganizar 🔻 Nueva carp	eta		- 🔟 🕜
OneDrive	Nombre	Fecha de modificación	Тіро
Cata and in a	EH_2015	01-12-2022 11:26	Archivo de vali
Este equipo	EH_2016	01-12-2022 11:23	Archivo de valo
🔶 Descargas	EH_2017	01-12-2022 11:18	Arcinizo de valo
Documentos	EH_2018	01-12-2022 11:13	Archivo de val
Escritoric	EH_2019	01-12-2022 11:09	Archivo de valo
nigenes	EH_2020	01-12-2022 11:06	Archivo de velo
Música	EH_2021	01-12-2022 10:58	Archivo de val
Dbjetos 3D			
Videos			
E Disco local (C:)			
- DATOS (D:)			
- Disseitation (111 V 6			13
Nombre	EH_2020	✓ *iC5V	×
	have been a second seco		

En la siguiente ventana, se verifican las características del archivo y, sobre todo, del conjunto de caracteres a utilizar, en este caso UTF-8

Import SV	data	Ē
Please pro	vide the information requested below	1
Location	D:\2022\Disociacion Datos\Prueba ARX\Archivos Anuales\EH_2020.csv	~ Brows
Charset	UTF-8	~
Delimiter		
Quete	и	~
Quote		

Se despliega una previsualización de los datos donde se puede verificar que cada columna viene con su correspondiente nombre.

data						
					19	
vide the informat	tion requested below					
D:\2022\Disociac	tion Datos\Prueba AR)	(\Archivos Anuales\EH_20	20.csv	Y	Browse	
UTF-8				~		
;				v		
8						
				Y		
Windows				VI		
First row conta	ains column names					
NCIA SEXO	EDAD_AÑ	OS ETNIA	GLOSA_PAIS_O	. COMUNA_R	ES G ^	
necien MUJER	2	No se identific	Chileno	05109	v	
entes HOMBF	RE 2	No se identific	Chileno	05502	C	
entes MUJER	0	No se identific	Chileno	05101	V.	and the second second second
entes MUJER	0	No se identific	Chileno	05101	V:	1000
entes MUJER	0	No se identific	Chileno	05101	V,	10
entes MUJER	1	No se identific	Chileno	05101	V /	02.0
necien MILIER	75	No se identific	Chileno	12122	>	
		< <u>B</u> ack <u>N</u> ext	> Em	a) (Cancel	
	data vide the informal D:\2022\Disociac UTF-& ; " Windows ☑ First row conta MCIA SEXO necien MUJER entes MUJER entes MUJER entes MUJER entes MUJER entes MUJER entes MUJER entes MUJER	data vide the information requested below D:\2022\Disociacion Datos\Prueba AR3 UTF-8 ; " Windows ☑ First row contains column names NCIA SEXO EDAD_AÑu tecien MUJER 2 entes MUJER 2 entes MUJER 0 entes MUJER 0 entes MUJER 0 entes MUJER 0 entes MUJER 1 berien MIIIFR 75	data vide the information requested below D:\2022\Disociacion Datos\Prueba ARX\Archivos Anuales\EH_20 UTF-8 ; " Windows ☑ First row contains column names NCIA SEXO EDAD_AÑOS ETNIA necien MUJER 2 No se identific entes MUJER 0 No se identific entes MUJER 0 No se identific entes MUJER 1 No se identific	data vide the information requested below D\2022\Disociacion Datos\Prueba ARX\Archivos Anuales\EH_2020.csv UTF-8 ; " Windows ☑ First row contains column names NCIA SEXO EDAD_AÑOS ETNIA GLOSA_PAIS_O tecien MUJER 2 No se identific Chileno entes MUJER 0 No se identific Chileno entes MUJER 0 No se identific Chileno entes MUJER 0 No se identific Chileno entes MUJER 1	data vide the information requested below D\2022\Disociacion Datos\Prueba ARX\Archivos Anuales\EH_2020.csv UTF-8 UTF-8 VUITF-8 V Windows V Windows V PFirst row contains column names NCIA SEXO EDAD_AÑOS ETNIA GLOSA_PAIS_O COMUNA_R eceien MUJER 2 No se identific Chileno 05109 entes MUJER 0 No se identific Chileno 05101 entes MUJER 0 No se identific Chileno 05101 entes MUJER 1 No se identific Chileno 05101 entes MULE 1 No se identific Chileno 05101 entes MULE 1 No se identific Chileno 05101 entes MULE 1 No 0 Entes MULE 1 No 0 Entes MULE 1 No 0 Entes NO 0	data vide the information requested below D\2022\Disociacion Datos\Prueba ARX\Archivos Anuales\EH_2020.csv V Browse UTF-8 V Windows V Windows V First row contains column names NCIA SEXO EDAD_AÑOS ETNIA GLOSA_PAIS_O COMUNA_RES G V entes MUJER V No se identific Chileno S109 V entes MUJER No se identific Chileno S101 V entes MUJER S S S S S S S S S S S S S S S S S S S

Hacemos clic en Next >, y se muestran los nombres de las columnas de los datos en el archivo con los cuales vamos a trabajar.

lumns lick right to	edit			÷
,				-
Selected	Name	Data type	Format	
¥	PERTENENCIA_ESTABLECIMIENTO_SALUD	String		
w	SEXO	String		
Se	EDAD_AÑOS	Integer		
V	ETNIA	String		
¥	GLOSA_PAIS_ORIGEN	String		
٣	COMUNA_RESIDENCIA	Integer		
v	GLOSA_COMUNA_RESIDENCIA	String		
1997 - C.	REGION_RESIDENCIA	Integer		
Correct Correction	GLOSA_REGION_RESIDENCIA	String		
w	PREVISION	Integer		
Y	GLOSA_PREVISION	String		
Carton Carton	MES_EGRESO	Integer		
٣	AÑO_EGRESO	Integer		
٠	DIAG1	String		
~	DIAG2	String		
	Move up	Ma	e down	
Perform da	ita cleansing			

Como la variante que vamos a utilizar no va a contener el mes de egreso, se quita de la carga, haciendo clic derecho sobre el dato "MES_EGRESO", con lo que se despliega un menú sobre los datos y se hace clic en "Select / deselect", como se ve en la siguiente imagen.

lick right to	edit		12	
Selected	Name	Data type	Format	
¥	PERTENENCIA_ESTABLECIMIENTO_SALUD	String		
¥	SEXO	String		
V	EDAD_AÑOS	Integer		
¥	ETNIA	String		
¥	GLOSA_PAIS_ORIGEN	String		
V	COMUNA_RESIDENCIA	Integer		
¥	GLOSA_COMUNA_RESIDENCIA	String		
¥	REGION_RESIDENCIA	Integer		
¥	GLOSA_REGION_RESIDENCIA	String		
¥	PREVISION	Integer		
v	GLOSA_PREVISION	String		
V	MES_EGRESO	. Inkawan		
¥	AÑO_EGRESO	Select / deselect		
Se	DIAG1	Change name		
9	DIAG2			
		Change data type >		
		Select all		9
	👔 Move up	Deselect all	wn	
Perform da	ta cleansing		1	

El dato MES_EGRESO queda deseleccionado, lo cual se verifica que ya no aparece el tic verde en su izquierda.

17

Import dat	a		
lumns			
lick right to	edit		
Selected	Name	Data type	Format
V	PERTENENCIA_ESTABLECIMIENTO_SALUD	String	
¥	SEXO	String	
1	EDAD_AÑOS	Integer	
V	ETNIA	String	
¥	GLOSA_PAIS_ORIGEN	String	
¥	COMUNA_RESIDENCIA	Integer	
¥	GLOSA_COMUNA_RESIDENCIA	String	
v	REGION_RESIDENCIA	Integer	
w/	GLOSA_REGION_RESIDENCIA	String	
V	PREVISION	Integer	
Y	GLOSA_PREVISION	String	
	MES_EGRESO	Integer	
¥	AÑO_EGRESO	Integer	
V	DIAG1	String	
Se	DIAG2	String	
	Move up	👃 Move o	down
Perform dat	a cleansing		
	< Back	Next > Fin	Cance

Se hace clic en "Next >", y se muestra una vista preliminar de la carga de datos seleccionados.

Y	Import data						י נ ר
re	view						
Ple	ease check whethe	r everything is rig	jht				Ľ
	GLOSA_REGIO	PREVISION	GLOSA_PREVIS	AÑO_EGRESO	DIAG1	DIAG2	
	De Valparaíso	3	CAPREDENA	2020	D763	NULL	
	De Valparaiso	1	FONASA	2020	C910	NULL	
	De Valparaiso	1	FONASA	2020	C910	NULL	
	De Valparaíso	1	FONASA	2020	C910	NULL	
	De Valparaisc	1	FONASA	2020	C910	NULL	
	De Valparaiso	1	FONASA	2020	C910	NULL	
	Metropolitana	2	ISAPRE	2020	T848	Y831	
	Ignorada	2	ISAPRE	2020	M169	NULL	
	Metropolitana	2	ISAPRE	2020	M169	NULL	
	Metropolitana	2	ISAPRE	2020	D469	NULL	
	Metropolitana	2	ISAPRE	2020	T848	Y831	
	Metropolitana	2	ISAPRE	2020	\$720	W019	
	Metropolitana	2	ISAPRE	2020	M248	NULL	
	Metropolitana	1	FONASA	2020	\$720	W180	
	Metropolitana	z	ISAPRE	2020	M248	NULL	
	Ignorada	96	NINGUNA	2020	\$720	W189	
	Metropolitana	2	ISAPRE	2020	M241	NULL	
	De Valparaiso	1	FONASA	2020	C910	NULL	
	De Valparaiso	1	FONASA	2020	C910	NULL	
¢							>
			< Back	C Next >	Fir	ish C	ancel

Se hace clic en "Finish" y se procede a la carga de datos en la pantalla principal.

3. Clasificando el tipo de dato

Los datos cargados en la pantalla principal tienen un color asignado en cada columna, el cual representa la clasificación del dato.

En la carga inicial son todos verdes, lo que representa que son insensibles.

					141
SEXO	EDAD_AÑOS	ETNIA	CLOSA_PAIS_ORIGEN	GOMUNA_RESIDENCIA	CLC
MUJER	2	No se identifica	Chileno	05109	Viña

Como se indica en la imagen siguiente, al costado derecho de la pantalla se encuentra un set de colores, los que se utilizan para cambiar el total de las columnas al tipo de dato que representan.

. Configure transformation 🔍 Explore results part Analyze unity 🍕	Analyze nik								
resst data 1					1 * * * *	Ostalizzed arrest	a Attaliate metadete		
100	-IDVD, ARIOS	ETrada	CLIEGA PARE CRA	iete connutur leiboie	ICH CLUSA, COMUNA, *	5.95 =w	and a second sec	- bardengen Gangilisten	
vi ha Perenecentes al Satema Neuronal d., MAPER	6	THE AS 120-1744	Childran .	\$5.02	V-Sa del Utar				I set all attribute types to goess identify of
Perferon ander al homo ka than sond de Se. NCA248			C HOLES	10000	1.800	STREET AND THE		- Marchart All	
The sector where so is the sector where is the first sector is the sector of the sector sector is the sector secto	5	NO CONTRACTO	10.001	1000	A CONTRACTOR OF THE OWNER OWNE				
a set of the sector of factors a list of a list of a	3 U	The last starting of		and here					
i afterter merter afterten effertren der an forfatte		the in marking	Colorado de Colora		Control in				
a die Catener was al Catena Verman di Milli	à.	East monthly	2	40429	and the second				
3 vitto faster a metat à States à far de a d' fittiff	2 · · · · · · · · · · · · · · · · · · ·	To is hard a	(minute	ofeen a	10102				
No Performes includes all Estimate tions at HIC 1981 6 51	1	No se dertifica -	Liberarp	12118	Tes Sar See				
st who Peter wantes a Suteria bacava di M2U C	n	No se identifice	Chileno	*Cos	Su-taliz-a				
as white Temeneser tea of Subernatilements on an Allaffi m	1	No se Identifica	Chiero	10114	s.esCondiss				
12 - Tro Picterscontes & Step no fuotions in INC15151	10	No se identifice .	Chileno	10122	Provensia.				
13 Villo Remonecionies al Schema flations st., BOLER 18	8	No se identifica	Chiena	+1325	Exists.				
14 With Representation Stream and Martin 20	7	here workships.	Chiefe .		Presidence.				
15 V No Veneraperter a Simona Naciona 4 (ACalifia) al	0	Name and Address	Logo .	10.04	54552-04				
PA y too Fe - electric al Seventa hack - 1 HCV#FE	£	Sale and Les	\$	60036	Number				
17 - No Person arts at Science baciena du KOLESI	đ	No widentifica .	Istan,ero	12/18	o Secures				
12 - Petersconnet ar Griema Nanceal ne ta Holdiffi	e	* 0.24 mm#cz	Ketoro.	0000	large man				
19 v Pertenes enter al Saleira Navional de Se. HONERE		Name Until the -	Children .	1000	(Maxwe) ii				
22 v: Protemacionales al Esterna Marchaella Hickdeff		Parketine Phone.	Criteral	X6700	largest 14				
21 - Partenec-entes-el erarma faac-enal de Se. HOLIE46		Note interview.	Dutte	44/03	(Walnut H				
 dio Potensi prese Strenship – u – MJ.El 		No se identifice	Chileno	27561	(Qum2)				
13 N. Farteretierten N Seller's fub" childe Se. 181/45		THE REPORT OF	Tellare.	10104	Trida yan Dian				
14 Personec entes al Sistema Racional da Se 1802ER	5 U	fiolds Centrics	Chileno	04-045	VIABOUS Nev				
B «Peteresettes sisters Necdandele Politik	\$ Q	No sendo stan	Childre	02-200	10 Ca 201 354				
29		The party of the local data	C. Margaret		-,194				
· ····································			A COMPANY OF A COMPANY	10.00	1000				
2 Vestoreards # Sins Alla doe or te Hutting		THE PROPERTY A	Contraction of the local distribution of the	10.95	Cardinana (
2) epiteratur siter siter site and a private		No. of Concession, Name		100.000	Courses .				
A Determined to all temp Marries of a Volders		14 18 99 11 14		200.001	24.284				
provide and a substance of the second		10.14	A PROPERTY.	44 M T	1214 TRue	PHLACE MEDRIN	PRE-lesion 1.csti and benefits		+ - 1 4
the Colorest start of the Servery of the hCtfld		Any on the latter	Theorem 1	CONT.	Courses of the second sec	(ppe	Marian	ALC: BURNER	
The Challed a series of the series of a stability of the series of the s		State of Conception of Concept	Comment	NU TON	ALC: NO				
15 - Performanie tes at Suitaves Mariter at de Se MAU		for an exercise	Sec.	1000	Law Col				
M Pederecenter al Summa Damenia nella 18216		Party months of	2000	citul	Latina				
27 Visio Salener enters Sitematia one d., NO1978		Bright Street State	PERMIT	0710	the cost Mart				
to who Persenances a Screena hap a di MCI-166		An as includes	ATT OF THE OWNER OF	10.50	Villa diel Sta-	General settings	Mility measure. Coding model		
10 will Pamprociantes a Screen filer mar a NOSHEE		The section erest	Orlants	10.10	Vita del Mar				
40 With the second a Science the contribute MESS		his and the	X + 3+10	00163	Velpeneta	Supprenico Frat	Gf.		
at - Ferneronmonat Supera Nacio-N de Se 813.62		firmieren.	Chiata	EGHER.	Villa Alemana V		-		
					,	Appreximate	Assume practical monofonicity		
ath glie en tráctich					() () () () () () () () () () () () () (
	- Trees.		Laboratory and	de l'étane		Precompletetion	Ensible Thresholds St.		
Sang (1120077 a 117047)	a) 1996.		24 autor pro	de faire		Precomputation	Chable Theeholds		

Los colores representan los tipos de datos Cuasi-identificador, identificador e insensible, respectivamente. Como la mayoría de los datos están clasificados como cuasi-identificadores, se hace clic en el circulo amarillo. Todos los colores de las columnas cambiarán a amarillo.

SEXO	EDAD_AÑOS	ETNIA	CLOSA_PAIS_ORIGEN	COMUNA_RESIDENCIA	CLC
MUJER	2	No se identifica	Chileno	05109	Viña

Ahora, para cambiar la clasificación de las columnas de diagnóstico, se hace clic en la columna respectiva del dato y se modifica el "Type" en la pantalla de la derecha.

	4 1 🚍 🗄 😏	Data transfo	ormation Attribute metadata	
DIAG1	DIAG2	Type:	Quasi-identifying	~
D763	NULL			
C910	NULL	Minimum:	All	~
C910	NULL			
C910	NULL			

Se despliegan los tipos de datos disponibles, y se selecciona el tipo Sensitive.

	1 1 🖹 🖻 😼	Data transf	ormation Attribute metadata	
- 1-61	DIAG2	Туре:	Quasi-identifying	~
D763	NULL		Insensitive	
C910	NULL	Minimum:	Sensitive	
C910	NULL		Quasi-identifying	
C910	NULL		Identifying	
0010	KIR M F			

La columna queda con un círculo violeta el cual identifica al dato sensible.

	다. 🕇 🔜 🖬 🥹	Data transfo	ormation Attribute metadata	
0 00161 D763	DIAG2	Туре:	Sensitive	ر √
C910	NULL	Minimum:	All	~ 1
C910	NULL			

Se hace lo mismo con el DIAG2

	(1) (1) 🗮	E 👳	Data transfo	ormation	Attribute	e metadata	
o DIAG1	o DIAG2	^	Type:	Sensitive			
D763	NULL		1			***************	
C910	NULL		Minimum:	All			~
C910	NULL		1				
Privacy models P	opulation Costs and bene	fits					+ - 1
Туре М	odel				Attribute		Add privacy model
Seneral settings	Utility measure Coding mo	odel Attribut	e weights				1
suppression limit;	0%						
Approximate: [Assume practical monot	onicity					

4. Incluir el tipo de tratamiento de anonimidad

Posterior a clasificar los datos sensibles, se debe indicar el tipo de tratamiento que se hará durante el proceso. Para esto se debe hacer clic en el + que se muestra en la imagen anterior y que se nombra como "Add privacy model".

Se despliega una nueva ventana donde se selecciona el nombre del dato y el tipo de tratamiento. En este caso L-diversidad.



Please	e select a privacy model which will be applied	to the data set				
Type (6) (67)	Model 8-Presence Profitability	Attribute				^
(1)	f-Diversity	DIAG1				
(t)	t-Closeness	DIAG1	*******			
(5)	δ-Disclosure privacy	DIAG1				
B	β-Likeness	DIAG1				
	l-Diversity	DIAG2				
1	t-Closeness	DIAG2				
6)	δ-Disclosure privacy	DIAG2				~
Configu	iration					v 🌚
L: 2	Variant: Distinct-I-diversity		~	C:	0.001	
	Note: you can also enter values by d	louble-clicking	the con	troll	knobs	

Hacemos clic en OK.

Privacy models	Population Costs and benefits		+ - / 4
Туре	Model	Attribute	
(I)	Distinct-2-diversity	DIAG1	
General settings	Utility measure Coding model Attribute weights		9
Suppression limi	t: 0%		
Approximate:	Assume practical monotonicity		
Precomputation	: Eneble. Threshold: 0%		

Se repite lo mismo para el DIAG2.

También, se debe configurar el tratamiento que tendrán los datos cuasi-identificadores, por lo que haremos clic nuevamente en el +.



2				×
Add a	new privacy model			
Please	select a privacy model which will b	e applied to the data set		
Туре	Model	Attribute		
DP	(ε, δ)-Differential privacy			
(k)	k-Anonymity			
R	k-Map			
(8)	δ-Presence			
ଟ	Profitability			
Ð	t-Closeness	DIAG'I		
(5)	δ-Disclosure privacy	DIAG1		
B	β-Likeness	DIAG1		
Ð	t-Closeness	DIAG2		 `
Configu	ration			~ •Y
К: 2				 10
	Note: you can also enter va	lues by double-clicking the contro	l knobs	
-	OK Cancel			

Seleccionamos la K-anonymity, el cual no está asignado a ningún campo en especial.

También desplazaremos la barra de "Suppression limit" totalmente a la derecha, lo que significa que, si hay registros únicos, se deben ofuscar el 100% de ellos. Como se muestra en la figura siguiente.

Privacy models	Population Costs and benefits		+ - 🖋 = = 🥹
Туре	Model	Attribute	
®	2-Anonymity		
0	Distinct-2-diversity	DIAG1	
0	Distinct-2-diversity	DIAG2	
General settings	Utility measure Coding model Attribute weights		W
Suppression limit	t 100%		
Approximate:	Assume practical monotonicity		
Precomputation	Enable, Threshold: 0%		

5. Creando jerarquías

Cuando se tienen cuasi-identificadores que se pueden agregar en jerarquías o en un orden alfabético, se tiene la opción de crear jerarquías que permiten visibilizar y verificar los valores presentes. Además, se puede seleccionar la forma en que aparecen en el conjunto de resultado.

6. Jerarquía de orden

Por ejemplo, si hacemos clic en la columna de pertenecia_establecimiento_salud, y luego en el icono mostrado en el recuadro en la figura siguiente, podremos crear la jerarquía de estos datos.

npu	data					1 t = E 🐭
	SEXO	EDAD_ANOS	> ETNIA	CLOSA_PAIS_ORIGI	EN COMUNA_RESIDENC	A CLOSA_COMUNA_ A
1	No Pertenecientes al Sistema Nacional d MUJER	2	No se identifica	Chileno	05109	Viña del Mar
2	Pertenecientes al Sistema Nacional de Se HOMBRE	2	No se identifica	Chileno	05502	Calera
3	Pertenecientes al Sistema Nacional de Se., MUJER	0	No se identifica	Chileno	05101	Valparaiso
4	Pertenecientes al Sistema Nacional de SeMUJER	C	No se identifica	Chileno	05101	Valparaíso
5	Pertenecientes al Sistema Nacional de Se., MUJER	C	No se identifica	Chileno	05101	Valparaiso
6	Pertenecientes al Sistema Nacional de Se MUJER	1	No se identifica	Chileno	05101	Valparaiso
7	W No Pertenecientes al Sistema Nacional d., MUJER	75	No se identifica	Chileno	13132	Vitacura
8	VI No Pertenecientes al Sistema Nacional d MUJER	40	No se identifica	Chileno	99999	Ignorada
9	Tho Pertenecientes al Sistema Nacional d., HOMBRE	51	No se identifica	Extranjero	13114	Las Condes

Se abrirá la ventana donde se debe seleccionar el tipo de datos que se va a utilizar y ordenar. En el ejemplo para pertenencia es "Use ordering".

tte a generalization hierarchy cify the type of hierarchy loc dates if or dates; tse intervate (for saniables with ratio scale) (e.g., for variables with ordinal scale) lse masking (e.g., for alphanumeric strings)	
crífy the type of hierarchy Ise dates (for dates) Ise intervale (for sariablas with ratio scale) Ise ordering (e.g., for alphanumeric strings) Ise masking (e.g., for alphanumeric strings)	~
lse Intarvat; (for tariable; with ratio scale) Ise intarvat; (for tariable; with ordinal scale) Ise masking (e.g., for alphanumeric strings)	~
fseintenate (for satiablae with ratio scate) ke ordering (e.g., for variables with ordinal scate) Ise masking (e.g., for alphanumeric strings)	
ise ordering (e.g., for alphanumeric strings) Ise masking (e.g., for alphanumeric strings)	
lse masking (e.g., for alphanumeric strings)	
Help I rad	Cancel

Clic en "Next >".

En la siguiente ventana se ve que se generó un set de datos con 2 valores.

Hierarchy wizard Create a hierarchy by on Specify the parameters	rdering and grouping items	×
Order Values No Pertenecientes al Pertenecientes al Sist	Groups	
Move up Move down Order: Custom ~	General Group Aggregate function: Set of values Function Parameter:	v
H	telp Søve < <u>B</u> øck <u>Next</u> >	Cancet

Hacemos clic en "Next >".

Se presentan los dos valores del set con distintos niveles de despliegue.

Non 23

Groups	Table						
221	Level-0	Level-1	Level-2				
	No Pertenecientes al Sistema Nacional de (No Pertenecientes al Sistem						
	Pertenecientes al Sistema Nacional de Ser	(Pertenecientes al Sistema N *					
1.1							

Se puede verificar que los valores se encuentren correctos y se hace clic en "Finish".

Se muestran nuevamente los niveles y el Minimum y Maximum nivel a utilizar en el valor por defecto All.

	2003/ Identifying	 Transformation: 	Generalization			
Minimum: A	: All	✓ Maxim	Maximum:	All		
	Level-0	Level-1			Level-2	

Al desplegar las alternativas de valores en el campo Minimum, se muestran los valores 0 al 2, correspondiente al valor del nivel (Level-O al Level-2). Y All correspondiente al uso de todos los niveles. Esto se repite en el campo Maximum.

En la imagen siguiente se ve valor del campo asociado al nivel 2.

Туре:	Quasi-identifying	v	Transformation:	Generalizatio	n	
Minimum:	All	~	Maximum:	All		
	A					-
	0	1000		and strength on the local division of the	-	Level-2
No Pertene	1	the second s	ervicios de Salud	SNSS}		

En este caso se utilizarán los valores asociados al Level-0, por lo que se selecciona el valor cero.

Type:	Quasi-identifying	✓ Transformatio	n: Generalization			
Minimum:	0	Y Maviguum 0				
	Level-0	Level-1	Level-2			

24 JUL

7. Jerarquía de intervalos

En el caso de las edades u otros valores numéricos se pueden agrupar.

Lo haremos con el campo EDAD_AÑOS, el cual seleccionaremos y haremos clic en el botón de jerarquía.

ジ AR File しっ し	X Anonymization Tool - EH_2020 Edit View Help 6 🐼 🐼 🕼 😂 🍰 👘 👘 🛩 🗶 🚓 📔 🖉 🗛	8 # 1 9.								
cas Co	nfigure transformation 🛛 -> Explore results 🕬 Analyze utility	y 🤑 Analyze risk								
Inpu	data					÷	Ť	-	目	10
	PERTENENCIA_ESTABLECIMIENTO_SALUD SEXO	- Dimitration	ETNIA	CLOSA_PAIS_ORIGEN	COMUNA_RESIDENCIA	+ CL	OSA	CON	MUN	A A
1	No Pertenecientes al Sistema Nacional d MUJER	2	No se identifica	Chileno	05109	Via	a del	Mar		
2	Pertenecientes al Sistema Nacional de Se HOMBRE	2	No se identifica	Chileno	05502	Cal	êra			
3	Pertenecientes al Sistema Nacional de Se., MUJER	0	No se identifica	Chileno	05101	Vals	ara!	so		
4	Pertenecientes al Sistema Nacional de Sa., MUJER	0	No se identifica	Chileno	05101	Valg	barai	iso		
5	Pertenecientes al Sistema Nacional de Se., MUJER	0	No se identifica	Chileno	05101	Vals	barai	so		
	177		Man and Informations	Children	DEADS	Section	1122	100		

Seleccionaremos en la ventana siguiente "Use intervals".

💱 Hierarchy wizard	o x
Create a generalization hierarchy	-
Specify the type of hierarchy	*
Use dates (for dates)	
Use intervals (for variables with ratio scale)	
Ouse ordering (e.g., for variables with ordinal scale)	
OUse masking (e.g., for alphanumeric strings)	

En la ventana siguiente se muestra un único set de valores desde 0 hasta 118 que es el mayor valor en la columna de datos.

Se deben setear en la pestaña "Range" los valores lower y upper según los valores mínimos y máximos de años, o los máximos valores esperados.

y defining intervals	
	· · · · · · · · · · · · · · · · · · ·
5	
erval Group	
erval Group	Upper bound
erval] Group)	Upper bound Snap from: 118
erval: Group)	Upper bound Snap from: 118 Top coding from: 112

Para crear el primer intervalo de valores de datos, se hace clic en el rectángulo del set hasta que esté en color amarillo, y nos movemos a la pestaña "Interval".

	W1281G	
reate a hiera	archy by defining intervals	×
Specify the pa	rameters	×
[0, 118[[0,	118[
General Rann	ne (lotenal - Group)	
General Rang	je (Interval ~ Group) nction: Default	v
General Rang Aggregate fun	je (Interval - Group) nction: Default	×
General Rang Aggregate fur Function Para Min:	je Interval - Group: nction: Default anieter:	×
General Rang Aggregate fur Function Para Min: Max:	ie (Interval ~ Group) nction: Default amieter: 0 118	•
General Rang Aggregate fur Function Para Min: Max:	je Interval - Group: Inclion: Default anteter 0 118	×

Actualmente el valor máximo del intervalo es 118, pero se quiere crear intervalos de 10 valores, por lo cual se establece el valor Max en 10.

	y denning inte	T VALS		
pecify the parameter	5			~
<mark>10, 101</mark> (0, 101				
	- Court			
eneral Range <mark>Inter</mark>	rat Group			
eneral Range <mark> Inter</mark> Aggregate function:	rat Group Default			·
eneral Range <mark>Inter</mark> Aggregate function: unction Parameter: Vin:	al Group Default			v
eneral Range <mark>Inter</mark> Aggregate function: uunction Parameter: Viin: Var:	al Group Default			v

Para crear el siguiente nivel, se hace clic derecho en el rectángulo del primer intervalo, y seleccionamos "Add new level".

aate a hierarchy	by defining intervals	~
pecify the parameter	irs —	~
[0, 10] [0, 10i		
	Nemove	
	Add before Add after	
	dema avon	
	Merge op	
	Mingk op Add new level	
	Minge up Add newlevel	
	Menga up Add new level	
General Range Inte	Menga op Add new level	
General Range Inte	Merge op Add newlevel	*
General Range / Inte Aggregate function Function Parameter	Merge op Add new level	
General Range / Inte Aggregate function Function Parameter Min:	Minga op Add new level	



Se crea un nuevo nivel con las mismas características del anterior.

eate a hierard	thy by defining intervals			<
pecify the para	meters			
<mark>(0, 10)</mark> (0, 10)	[0, 10] [0, 10]			
eneral i Banos	Interval Group			
eneral (Range)	Interval Group			
eneral [Range] ggregate func unction Peram	Interval Group; tion: Default eter:			v
eneral [Range] Aggregate func Function Param Min:	Interval Group tion: Default eter: 0		 	

Se hace clic sobre el nuevo nivel hasta que quede en color amarillo.

					×	
reste a hierarchy by defining intervais Specify the parameters						
[0, 10] [0, 10]	<mark>[0, 10]</mark> [0, 10]					
General (Range IIn	rer vali Group					
General (Range In Aggregate functio	terval Group				~	
General (Range In Aggregate functio Function Paramete	xe vali Group n: Default an	*			~	
General (Range In Aggregate functio Function Paramete Size:	ve vali Group n: Default s:	*			2	
General (Range In Aggregate functio Function Parametr Size:	rei val. Group n: Default an 1				~	

En la pestaña "Group", se modifica el valor "Size" a 2.

Specify the parameter	y defining intervals 3	~
[0, 10] (0, 10]	<mark>(0, 20)</mark> (0, 20)	
General Range (Inte	Val Group	
deneror i nonge ; mie	Default	~
Aggregate function:		
Aggregate functions Function Parameters		
Aggregate function: Function Parameter Size:	3	

El nuevo set quedará con el doble de valores del set anterior. Para crear el siguiente nivel, se selecciona el ultimo nivel creado, se hace clic derecho sobre el set y se selecciona "Add new level".

Hierarchy wizard Create a hierarchy by Specify the parameter:	y defining i	intervals					•	×
0, 10	<mark>10.201</mark> .0,2	Remove Add before Add after Merge down Merge up						
General Range Interv Aggregate function: Function Parameter:	ral Group Default	Add new level						4
Size:	d Help) 1	Save	< Back	Next >	Taish"	Cancel	

Nuevamente, se fija en 2 el valor de Size en la pestaña "Group".

Se continua así hasta que el primer nivel tenga el valor máximo de edad contenido en él.

🌍 Hierarchy wizard	10270103	نمية(10) . ا	Auf A.S.aur		-		n x
Create a hierarchy I Specify the paramete	by defining in	tervals					- C
50,000, 500,000 100,100, 100,00 100,100, 100,00 (20,100, 100,00 000,010, 100,000		1000 a					•
line and lines							*
Aggregate function: Function Parameter:	Default						~
Size:	2						
	Help	Vest.	Save	< Back	Next >	Filles (Cancel

Hacemos clic en "Next >".

Se mostrarán distintos niveles de intervalos.

broups	lable	1 tools	1 10012	1	1	1	
12	Level-0	Level-1	Lever-2	Level-3	Level-4	Level-5	
6	1	10, 101	10, 20	10, 401	10, 801	10, 118	
3	2	10, 10	10, 20	10, 40[10,80	10, 119	
1	2	10, 10	10, 201	10, 401	10, 801	[0, 118]	
		0, 10	10, 201	10, 40	10, 001	[0, 118]	
		0, 10	10, 201	10, 40	10, 801	[0, 110]	
- D		0, 10	10, 20	10, 40	[0, 00]	[0, 110]	
	7	10, 10	10, 20	10, 40	10, 301	[0, 118]	
	0	10, 10	10, 20	10, 40	10, 801	[0, 118]	
	8	10, 101	10, 201	[0, 40]	10, 801	[0, 118]	
	3	10, 101	10, 201	[0, 40]	10, 801	[0, 118]	
	10	110, 201	10, 20	10, 401	10, 801	10, 118[
	12	10, 20	10, 201	[0, 40]	10, 201	[0, 118]	
	12	[10, 20]	10, 201	10, 401	[0, 80]	10, 118[
	13	110, 201	10, 201	10, 401	10, 80	0, 118	
	14	10,201	10, 201	10, 40	10, 201	[0, 118]	

Se presiona "Finish".

Data trans	sformation Attr	ibute metadata							1 4 10
Type:	Quasi-identifying		~	Transformation:	Generalization				54
Minimum: All			··· Maximum:		All				~
	Level-0	1	Level-1			Level-2	Level-3	Level-4	1 ^
0		[0, 10]				[0, 20]	[0, 40]	[0, \$0]	[0,
1		[0, 10]				[0, 20]	[0, 40]	[0. 80]	[D,
2		[0, 10[[0, 20]	[0, 40]	[0, 80]	[0,
3		[0, 10]				[0, 20]	(0, 40[[0, 80]	[0,
4		[0, 10[[0, 20]	[0, 40[[0, 80[[0,
5		[0, 10[[0, 20]	[0, 40]	[0, 90[[0,
6		[0, 10]				(0, 20)	[0, 40[(0, 80[[0,
7		10. 10[[0, 20]	[0, 40]	(0, 80([0,
8		[0, 10[[G. 20]	10, 40[[0, 80[[0,
9		[0, 10[[0. 20]	[0, 40[[0, 80]	(O, '
10		[10, 20]				[0, 20]	[0, 45]	[0, 80]	(C,
11		[10, 20[[0, 20]	[0, 40]	[0, 90]	[0,
12		[10, 20]				10, 201	[0, 40[[0, 80]	(0,
13		[10, 20]				[0. 20]	[0, 40]	[0, 80]	[0,
14		[10, 20]				[0, 20]	[0, 40]	[0, 80[[0,
15		[10, 20]				[0, 20]	[0, 40]	(0, 80[[O,
16		[10, 20]				[0, 20]	[0, 40]	[0, 80]	[0,
17		[10, 20]				[0. 20]	[0, 40]	[0, 80[(0,
18		[10, 20]				[0, 20]	[0. 40]	[0, 80]	[0,
19		[10, 20]				[0, 20]	[0, 40]	10, 80[[0,
20		[20, 30]				[20, -10]	[0, 40]	(0, 30)	(O,
21		[20, 30]				(20, 40)	[0, 40[[0, 90[[0,
22		[20, 30]				[20, 40]	[0, 40]	[0, 80[[C,
23		[20, 30]				[20, 40]	[0, 40]	108 (0)	[0,
24		[20, 30]				[20, 40]	[0, 40]	[0, 80[[0, *
6									5.

Se selecciona el valor mínimum y máximo del nivel que queremos mostrar en el proceso. En este caso el intervalo de 10 en 10 que está en el nivel 1.

Data transformation Attr	ribute metadata							1
Type: Quasi-identifyi	ng	~	Transformation:	Generalization				
Minimum: 1		v	Maximum:	1				
Level-0		Level-1	×		Level-2	Level-3	Level-4	1
0	[0, 10]	Contraction of the local data	and the second se		[0, 20]	[0, 40[[0, 80]	[0,
1	[0, 10]				[0, 20]	[0, 40]	10, 80[[0, 1
2	[0, 10]				[0, 20]	[0. 40]	[0, 80]	10,
3	[0, 10]				[0, 20]	[0, 40]	[0, 80[[0,
4	[0, 10]				[0, 20]	[0, 40]	[0, 80[[0,
5	[0, 10]				[0, 20]	[0, 40]	10, 90[[0,
6	[0, 10]				[0, 20]	[0, 40]	[0, 85]	[0,
7	[0, 10]				[0, 20]	[0, 40]	[0, 80[[0,
8	[0, 10]				[0, 20]	[0, 40]	[0, 80[10,
9	[0, 10]				[0, 20]	[0, 40]	[0, 80[[C,
10	[10, 20]				[0, 20]	[0, 40]	[0, 80]	[0,
11	[10, 20]				[0, 20]	[0, 40]	[0, 80]	[0,
22					1000	4. 160		

Para generar el proceso de anonimización no es necesario crear jerarquías para cada columna, por lo que se usa principalmente cuando se quiere agrupar valores o crear intervalos. Ahí si es necesario crear la jerarquía respectiva.

Finalizada, la clasificación y jerarquización de los datos, se genera el proceso de anonimización, haciendo clic en el tic que se destaca en la siguiente imagen.

228 LC	nfigure transformatio	n & Explore	results 🖋 Analyze ul	tility 🦊 Analyze i	risk		
Input	data						1
	REGION_RESIDENCIA	PREVISION	CLOSA_PREVISION	ANO_EGRESO	· DIAG1	0 D(((d)	1.4
1	ıcá	1	FONASA	2020	\$523	W189	
2	icá	1	FONASA	2020	Q532	NULL	
3	icá	1	FONASA	2020	N47X	NULL	
4	:Cá	1	FONASA	2020	Z412	NULL	
5	ıcá	1	FONASA	2020	Q531	NULL	
6	ĸá	1	FONASA	2020	N47X	NULL	
7	icá	1	FONASA	2020	Z412	NULL	
8	ıca	1	FONASA	2020	N47X	NULL	
9	icà	1	FONASA	2020	N47X	NULL	
10	:cá	3	FONASA	2020	N47X	NULL	
11	ĸá	1	FONASA	2020	N47X	NULL	
12	icà	1	FONASA	2020	N47X	NULL	

Se mostrará la siguiente ventana en la cual haremos clic en OK.

9				×
Anonymization options				
Please enter the required para	ameters			
Search strategy				
Optimal	Best-effort, binary	Best-effort, bottom up		
O Best-effort, top down	Best-effort, genetic			
Please note: the optimal and I This threshold can be configu	binary search strate red in the project s	gies are not available, becau ettings.	ise the solution	space is too large.
Limits				
O Limited number of steps:	1000			
Limited time [s]:	30.0			
Transformation model				
Global transformation				
O Local transformation using	iterations: 100			
	Lange			
		_		

Comenzará el proceso de anonimidad.

Progress Information	
Operation in progress	
•	
	Const

El proceso finaliza cuando aparece el resultado como en el recuadro de la imagen siguiente.

File	RX Anonymization Tool - EH_2020 Edit View Help	×	14 m m 14	0			Attr	ibute: DIAG2 Tr.	ansformati	ons: 1 Sele	cted: [0, 0, 1, 0	0, 0, 0, 0, 0, 0, 0, 0, 0]	Applied: [0, 0,	1, 0, 0, 0	0	×
- i c	onfigure transformation Explo	re results : >= Ana	ilyze utility 🔶 Analy	ze risk)					-		_		-	-		~
Inpu	rt data				J T =	E .	Data transforma	tion Attribute	metadata						4.0	-
	INCIA CLOSA_REGION_RESIDENCI	PREVISION	CLOSA_PREVISION	ANO EGRESO	o DIAG1	-	Time: Sen	ritina		-	and constitutes	Concestration				
1	De Tarapacá	1	FONASA	2020	\$523	WN	(if the set	alove		- 21.5	insidentiasion.	Cemeraturation				
2	De Tarapacá	1	FONASA	2020	Q532	NUI	Minimum: 0			∼! M	aximum:	0				~1
1	De Tarapacá	1	FONASA	2020	NATX	NUE										_
12	DeTeranacé	4	FONASA	2020	Z412	NUN .	Level-0	Level-1		Level =2	1				_	•
124	De Tarapaçã	1	FONASA	2020	Q531	NUI	NULL	{NULL}	10							1
1	Da Tarapaca	1	FONASA	2020	15478	NUI	V010	[1010]	×.							1
1	De Tarapaçá	1	FONASA	3020	2412	NU	· V011	(V011)								- 1
							P CORE AND	and the second second								

Para analizar el resultado hacemos clic en la pestaña de "Analyze utility". Ahí podremos ver el conjunto resultante en la pestaña "Output data". Además, en la parte inferior en la pestaña "Class sizes",

podemos ver en la fila "Suppressed records", la cantidad de filas con datos ofuscados y el porcentaje que representan esos datos en el total de filas analizadas.

En la imagen siguiente se muestra un resultado de 7,63% de registros ofuscados.

ie Cont	figure ransformation	sults (per Analy)	e oblity . 4 Analy	zerisk									1
Input da	at a . Classification performance. C	walky models			1 7 =	₽	Output da	ta Classification performance	Quality models			1 1 3	E 5
	NCIA CLOSA_REGION_RESIDENCIA	PREVISION	CLOSA PREVISION	ANO_EGRESO	a DIAG1		INC	IA CLOSA_REGION_RESIDENCIA	PREVISION	CLOSA_PREVISION	ANO_EGRESO	· DIAG1	Q
4	De Tarapaca	1	FONASA	2020	5523	WIL	1	De Terepacé	1	FONASA	2020	\$523	W3ł
2	Ca Tarapacá		FONASA	2020	Q532	NUI	2	De Tarapaca	1	FONASA	2020	Q532	NUL
3	De Tarapaca	6.	FONASA	2020	N47X	NUI	3	De Tarapaca	3	FOMASA	2020	N47X	NUL
4	De Terapaca	F	FONASA	2020	Z412	NUI	4	De Terapacé	1	FONASA	2020	Z112	NUL
5	De Tarapacá	1	FONASA	2020	Q531	NUI	5	De Tarapacá	1	FONASA	2020	Q531	NUE
£	De Tarapacá	Ú	FONASA	2020	N47X	NUL	ð	De Tarapaca	1	FONASA	2020	N47X	NU
1	De Tarapacá	t.	FONASA	2020	Z412	NUI		De Tarapacá	1	FONASA	2020	Z412	NUL
	De Tarapaca	É.	FONASA	2020	N47X	NA	8	De Tarapacé	1	FONASA	2020	N47X	NUL
5	De Tarapacá	ii -	FONASA	2020	N47X	NUI		De Tarapacá	1	FONASA	2020	N-57X	NUL
10	Da Tarapacá	6	FONASA	2620	N47X	NUI	10	De Tarapaca	1	FONASA	2020	N47X	NUL
11	De Tarapaca	P	FONASA	2020	N47X	NU	11	De Tarapacá		FONASA	2020	N47X	NU
12	Ca Tarapaca	Č.	FONASA	2020	N47X	NUT	12	De Tarapacé	1	FONASA	2020	N47X	NUE
13	De Tarapacá	Ê	FONASA	2020	Z412	NUH	13	De Tarapaca	1	FONASA	2020	7412	NU
14	De Terapaca		FONASA	2020	N47X	RUI	14	De Tarapaca	1	FONASA	2020	N47X	NUR
15	De Tarapaca 1	É E	FONASA	2020	N457X	NUI	15	De Texaneré	3	FONASA	2020	N47X	NUE
16	Da Tarapaca		FONASA	2020	NGTX.	NIB	16	De Taranacá	1	FONASA	2020	21478	NIE
17	De Tarapaca		FONASA	2020	N47X	NUR	17	De Tarapaca	1	FONASA	2020	N4TX	NUE
18	De Tarapaca	P. C.	FONASA	2020	N478	NUL	18	De Tarapaca	1	FONASA	2020	NATE:	NER
19	De Tarapaca		FONASA	2020	N4TX	NUL	19	De Terapace	I	FONASA	2020	N47X	NR
20	De Tarapacá		FONASA	2026	7412	NUI	20	De Tertencá	1	FONASA	2020	7412	AS II
												2.012	
Summa	ry statistics Distribution Contingen	CY Class sizes	Properties Classifi	cation models		10	Summary	statistics Distribution Continge	ncy i Class sizes	ProPerties Classific	ation models		10 10
Measur		value (incl. suppo	essed)	Value (excl. su	opressed	^	Measure		Valuet (incl. support	essed)	Value feyel, su	onnessed)	
Averag Maxim Minima Suppre	e class size 7 al class size 8 al class size 9 al class size 9 seedre cords 0	7.40606 (0.00056 731 (0.0549.3%) 1 (0.00008%) 0 (0%)	(c) (c)	7.40606 (0.090 731 (0.05494% 1 (0.00008%)	55%))		Average of Maximal Minimal of Suppress	class size class size class size ed records	70.92983 (0.00533) 3417 (0.25683%) 2 (0.09015%) 103249.0 (7.7603%)	5a)	70.92983 (0.00) 3417 (0.278435 2 (0.0001655) 0	578%£) 6)	
A		112.11		*****			He -under and	et	1111		11101		

Los datos anonimizados se pueden exportar haciendo clic en "Export data" del menú "File".

🍟 A	RX Anonymization Tool	- EH_2020								
File	Edit View Help									
4	New project	P.3 11	sal .		1				Attribute: DIAG2 Transforma	tions: 1
16	Open project	Explore res	ults Analyz	e utility 🛛 🌵 Analyz	se risk					
4	Save project	formance Q	ality models			ITE	E 🐭	Output da	ta Classification performance	e Qual
	Save project as	ESIDENCIA	PREVISION	CLOSA_PREVISION	ANO_EGRESO	· DIAG1	0 4	INC	A CLOSA_REGION_RESIDENCE	A PI
	Import date	3		FONASA	2020	\$525	Whi	23	De Tarapacá	3
1228	Export data	1		FONASA	2020	Q532	NUI	2	De Tarapacá	1
	Careta antifanta			FONASA	2020	N47X	NUR	3	De Tarapacá	3
	Create centricate	1		FONASA	2020	Z412	NU	4	De Tarapacá	1
	Import hierarchy	1 1		FONASA	2020	Q531	NU	5	De Tarapacá	3
	Export merarchy	1		FONASA	2020	N47X	NUI	5	De Tarapacá	8
		1		FONASA	2020	Z412	NUI	7	De Tarapacá	.1
٠	Exit	1 3		FONASA	2020	N47X	NU	8	De Tarapacá	1
9	De Tarapacá	1		FONASA	2020	N47X	NU!	9	De Tarapacá	8
10	De Tarapacá	1		FONASA	2020	N47X	NUT	10	De Tarapacá	T
11	De Tarapaca	1		FONASA	2020	N478	NU	11	De Tarapacă	1
		1			****		and the second			100 m

Se debe buscar una ruta de destino y un nombre para el archivo con los resultados.

Ahí damos por finalizado el proceso de anonimización.

- Τ	« Disociacion Datos » Prueba ARX »	🗸 👌 🖉 Buscar en Pre	ueba ARX
Organizar 👻 Ni	ueva carpeta		(55 🔸 🔞
Este equipo	 Nombre 	Fecha de modificación	Tipo
Descargas	AnalisisAlternativa3	01-12-2022 19:51	Carpeta de ar
	Archivos Anuales	01-12-2022 16:30	Carpeta de ar
Excritorio	i edad_jerarquia	16-11-2022 23:18	Andrivo de va
Eschieno	EH_alt2_95_41%perdida	15-11-2022 11:22	Archivo de va
Imagenes	🛍 EH_alt2b_63_97%perdida	15-11-2022 11:38	Archivo de va
Música	EH_alt2c_43_75%perdida	15-11-2022 11:44	Archivo de va
🄰 Objetos 3D	ierComunaresi	25-11-2022 16:27	Archive de va
Vídeos	jerEdad	25-11-2022 15:37	Archivo de va
L Disco local (C	a) 🗐 jerEtnia	25-11-2022 15:37	Archivo de va
DATOS (D:)	ierGlosacomuna	25-11-2022 16:28	Archivo de va
Nombre:	EH_2020_con_comuna		×
<u>Т</u> іро:	*.csv		2
Ocultar carnetas		Guardar	Cancelar

2) Anonimización utilizando software R

Usaremos, como ejemplo, la anonimización de las bases ENO (Enfermedades de Notificación Obligatoria). Llamemos "base" al objeto con el data frame que contiene todos los casos sin anonimizar.

1. Variables identificadoras.

El primer paso será quitar las variables identificadoras de nuestro objeto "base". Por ejemplo, si las variables "identificacion_paciente", "nombre_paciente" y "direccion_paciente" son nuestras variables identificadoras, debemos correr el siguiente código:

```
base = base %>%
    select(-identificacion_paciente, -nombre_paciente, -direccion_paciente)
```

2. Variables cuasi-identificadoras

A modo de ejemplo, trabajaremos con tres variables cuasi-identificadoras: "sexo", "grupo_edad" y "codigo_comuna". El código de la comuna son los cinco dígitos del código único territorial Región-Provincia-Comuna, por ejemplo, 05302 será la Región de Valparaíso, Provincia Los Andes, Comuna Calle Larga. La manera de anonimizar el código de la comuna es 053** si se desea anonimizar la comuna y 05*** si se desea anonimizar la comuna y provincia. Se anonimizará priorizando "sexo" y "grupo_edad" por sobre "codigo_comuna". Por último, la variable sensible de nuestra base será "ENO", la enfermedad con la cual fue notificado el caso.

El primer paso será crear los tres niveles posibles de anonimización del código de comuna:

Ahora crearemos las variables K y L resultantes de cada nivel de anonimización del código de comuna:

```
base = base %>%
group_by(sexo, grupo_edad, cod_comuna_primer_nivel) %>%
mutate(K_primer_nivel = n(),
    L_primer_nivel = n_distinct(ENO)) %>%
ungroup() %>%
group_by(sexo, grupo_edad, cod_comuna_segundo_nivel) %>%
mutate(K_segundo_nivel = n(),
    L_segundo_nivel = n_distinct(ENO)) %>%
ungroup() %>%
group_by(sexo, grupo_edad, cod_comuna_tercer_nivel) %>%
mutate(K_tercer_nivel = n(),
    L_tercer_nivel = n_distinct(ENO)) %>%
ungroup()
```

Crearemos la variable "cod_comuna_final" con el menor nivel de anonimización que cumpla con que K y L sean mayores o iguales a dos (se debe cambiar el 2 en el código si se desea trabajar con un K o L mayor):

```
base = base %>%
mutate(cod_comuna_final = case_when(
    K_primer_nivel >= 2 & L_primer_nivel >= 2 ~ cod_comuna_primer_nivel,
    K_segundo_nivel >= 2 & L_segundo_nivel >= 2 ~ cod_comuna_segundo_nivel,
    K_tercer_nivel >= 2 & L_tercer_nivel >= 2 ~ cod_comuna_tercer_nivel,
    TRUE ~ NA_character_
))
```

Notemos que si se cumple con que "cod_comuna_final" no tenga elementos vacíos, la anonimización estará lista ya que cada grupo tendrá un K y un L mayor o igual a 2. Para verificar esto, podemos correr el siguiente código:

```
table(is.na(base(cod_comuna_final))
```

Si existe algún elemento vacío, tendremos que anonimizar la variable "sexo" como sigue:

```
base = base %>%
mutate(sexo =ifelse(is.na(cod_comuna_final), "***", sexo))
```

Y repetir el proceso anterior. Si al repetir el proceso volvemos a tener elementos vacíos en la variable "cod_comuna_final", debemos anonimizar la variable "grupo_edad" como sigue:

```
base = base %>%
    mutate(grupo_edad =ifelse(is.na(cod_comuna_final), "***", grupo_edad)))
```

Y repetir el proceso anterior. Si siguen existiendo elementos vacíos en la variable "cod_comuna_final", deberemos usar el código de comuna completamente anonimizado como sigue:

```
base = base %>%
mutate(cod_comuna_final = ifelse(is.na(cod_comuna_final), "*****", cod_comuna_final))
```

3. Quitar variables originales y auxiliares

Debemos remover la variable "codigo_comuna" original y las variables creadas con la información de K y L:

4. Exportar la base

El objeto "base" está listo para ser exportado con funciones comunes de escritura de R.

