Ministerio de Salud

Gobierno de Chile

- 241

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

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 pública[2].

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 "*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,%2010.05.2022.%20Emite%20pronunciamiento.%20Subs.%20Salud%20Pu%CC%81blica.%20Transparencia%20Proactiva.pdf

- **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 re-identificación

- **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 (l) 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 otro[3], 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 cuasi-identificadores 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).

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

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	I213	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.

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 K-anonimidad mayor o igual a 2 y una L-diversidad mayor o igual a 2 para los cuasi-identificadores.
- 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: deis@minsal.cl

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.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html
9. 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 anonimización[4]

Fecha: : _____

Departamento responsable de la base de datos : _____

Encargados/as del procedimiento de anonimización : _____

_____

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 explícito	Cuasi-identificador	Atributo sensible	Atributo no sensible

Describa brevemente la técnica de anonimización aplicada:

_____

_____

_____

_____

Encargados/as de verificar correcta anonimización : _____

Realizó prueba del intruso motivado y hallazgos : _____

_____

_____

Firma del jefe de Depto responsable de la base de datos : _____

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 : _____

[4] Formulario disponible en el siguiente enlace: https://forms.office.com/r/K6GEdNiRkh?origin=lprLink

Anexo II: Herramientas para apoyar el procedimiento de anonimización.

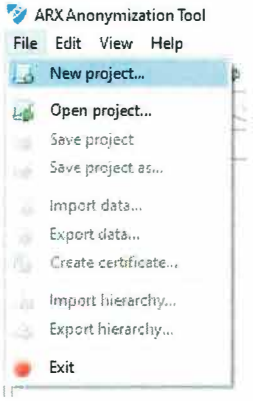
Existe múltiples herramientas para realizar los procedimientos de anonimización. Una opción es utilizar el software libre ARX. Esta herramienta es flexible ya que soporta una amplia variedad de modelos de privacidad y permite aplicar diversos métodos para transformar los datos y analizar la privacidad 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: https://cran.r-project.org/bin/windows/base/.

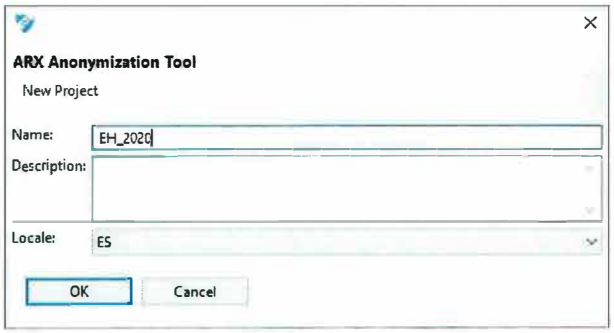
1) Anonimización utilizando software ARX

1. Creación de proyecto

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



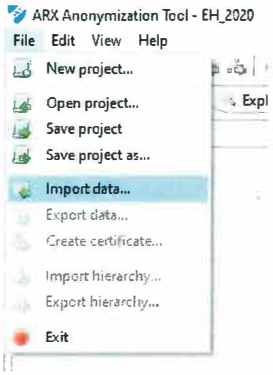
Se asigna nombre al proyecto, en este caso EH_2020.



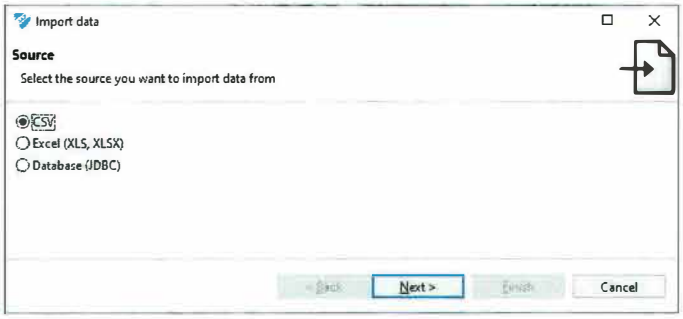
Se crea el proyecto mostrando el nombre en la barra principal de la ventana.

2. Importación de datos

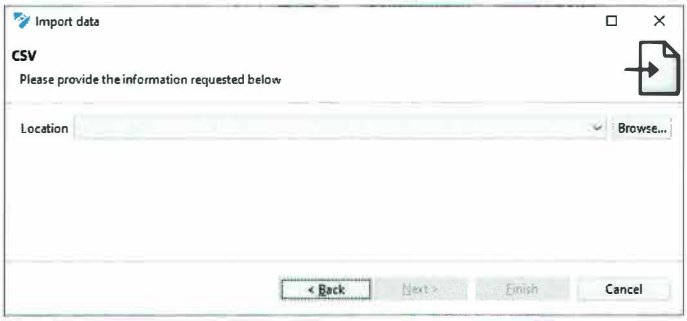
Se importan los datos a utilizar a través del menú "File", opción "Import data".



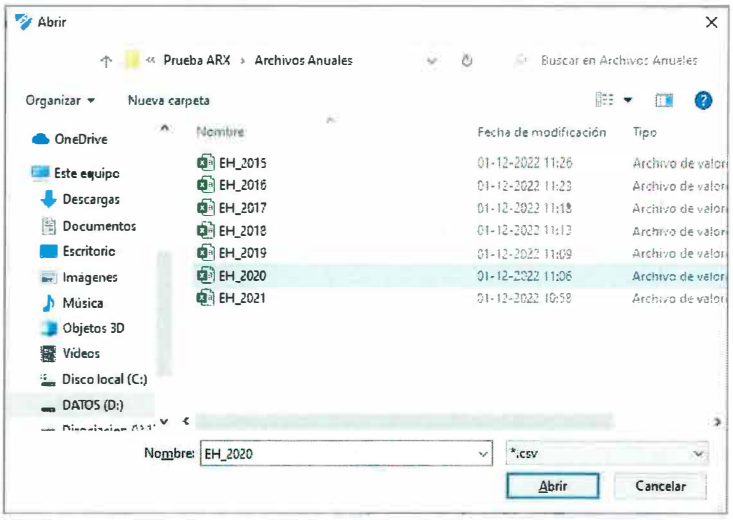
Se despliega una ventana donde se selecciona el tipo de origen de los datos, en este caso es un archivo CSV.



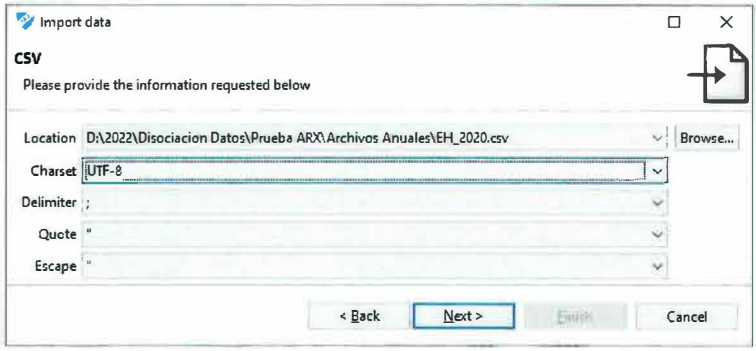
En la siguiente pantalla se busca el archivo en la ruta respectiva haciendo clic en "Browse..."



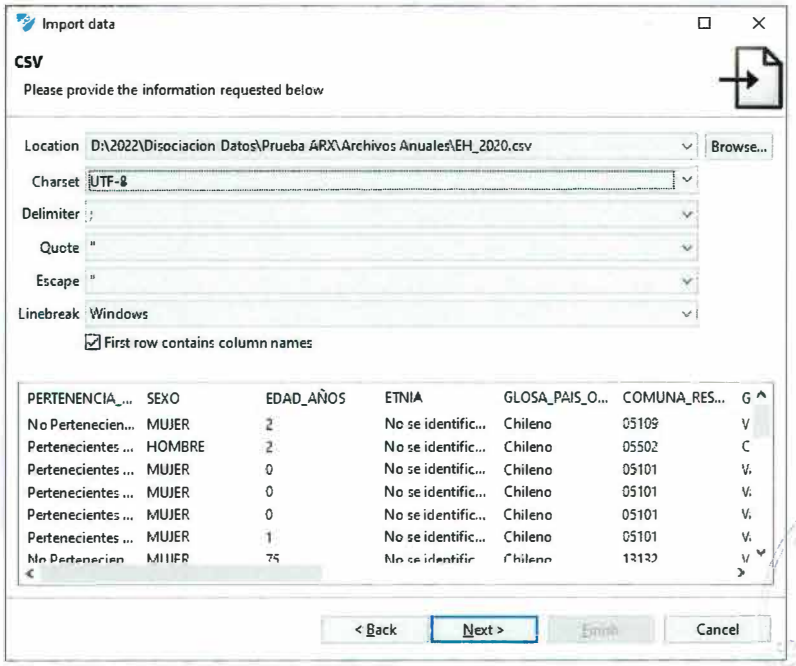
Se selecciona el archivo correspondiente.



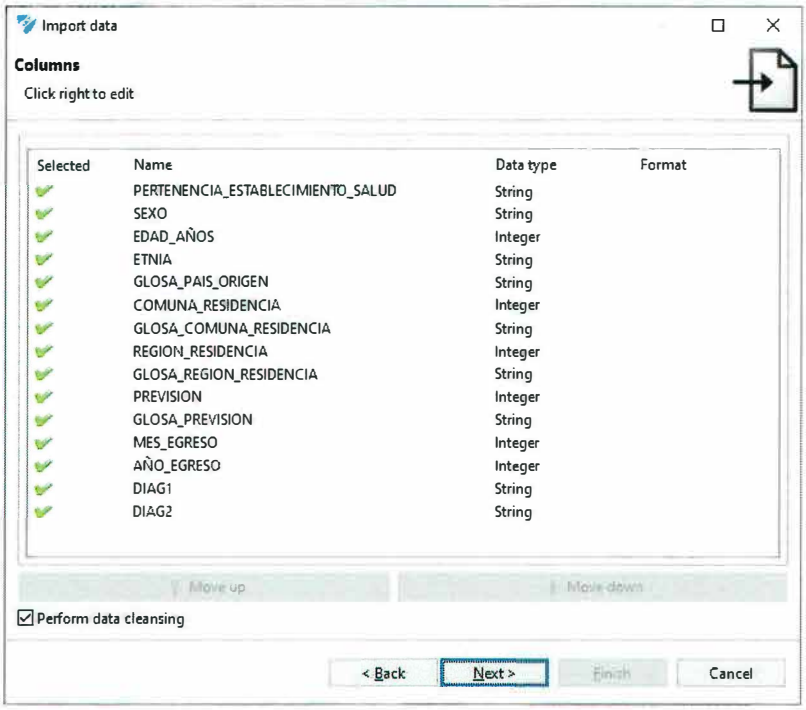
En la siguiente ventana, se verifican las características del archivo y, sobre todo, del conjunto de caracteres a utilizar, en este caso UTF-8



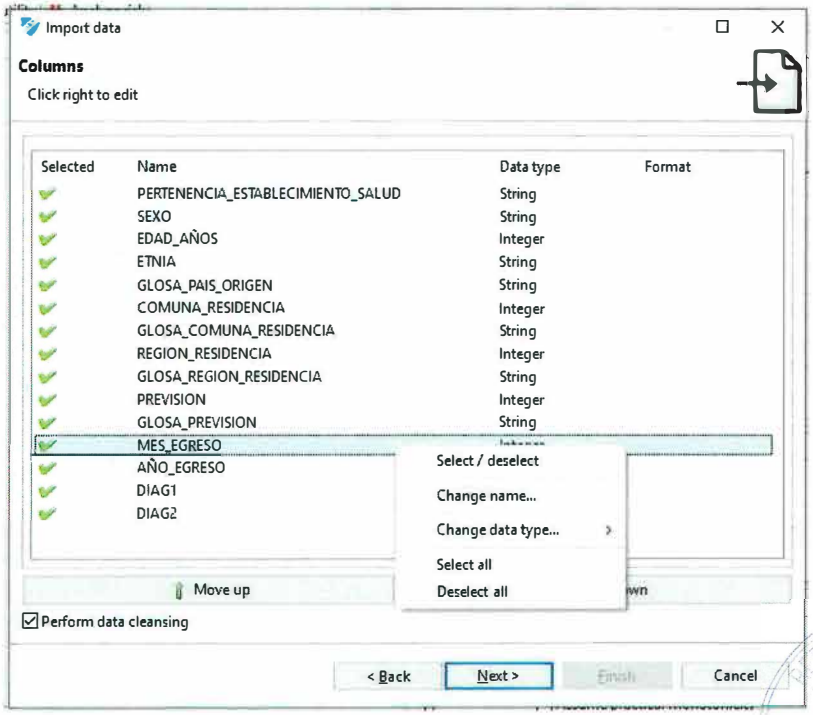
Se despliega una previsualización de los datos donde se puede verificar que cada columna viene con su correspondiente nombre.



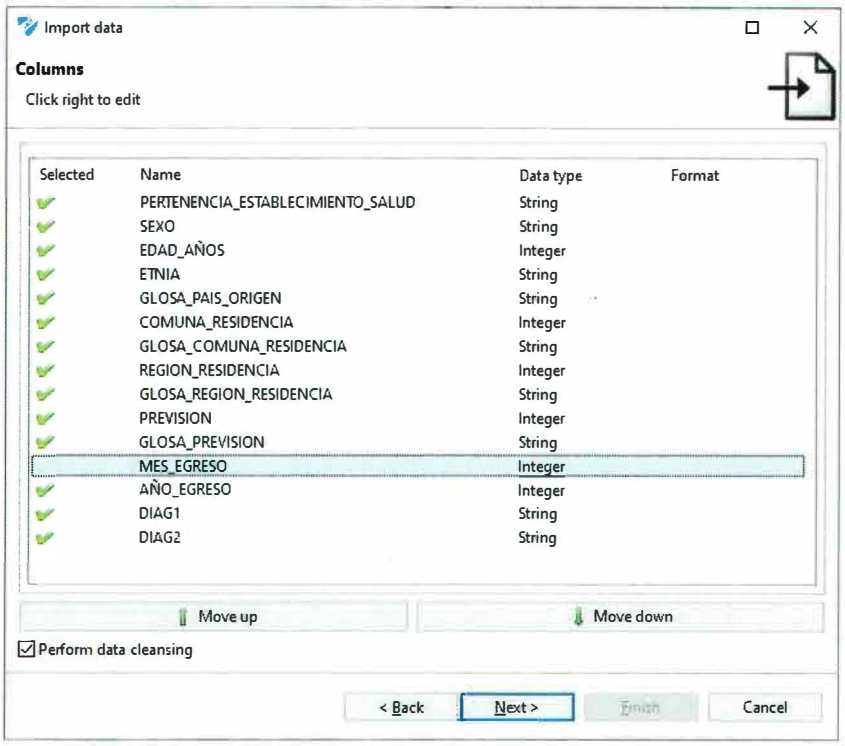
Hacemos clic en Next >, y se muestran los nombres de las columnas de los datos en el archivo con los cuales vamos a trabajar.



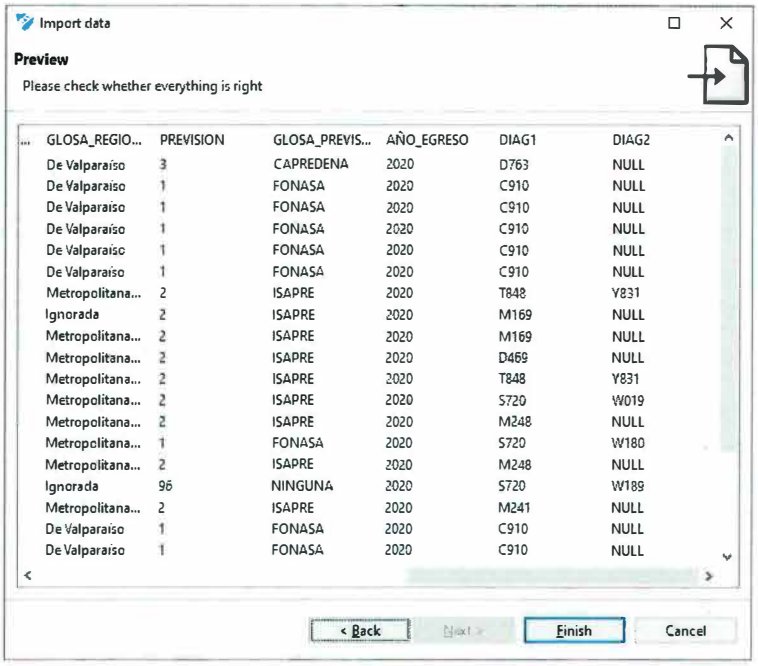
Como la variante que vamos a utilizar no va a contener el mes de egreso, se quita de la carga, haciendo clic derecho sobre el dato "MES_EGRESO", con lo que se despliega un menú sobre los datos y se hace clic en "Select / deselect", como se ve en la siguiente imagen.



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



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



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

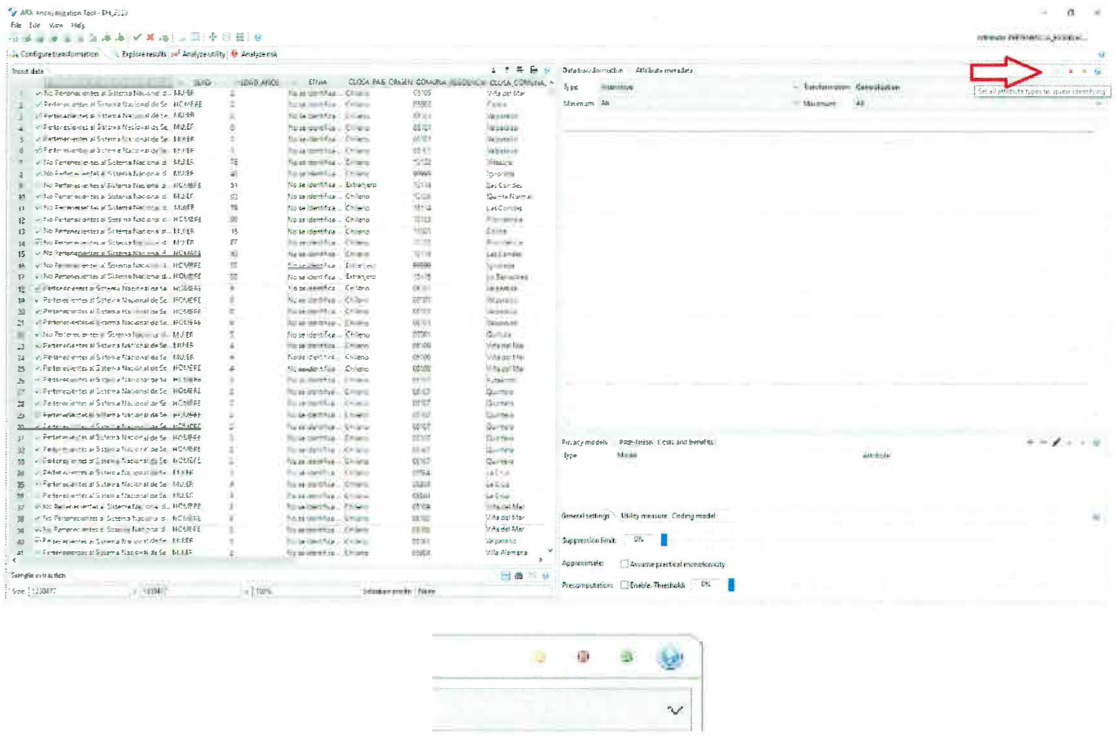
3. Clasificando el tipo de dato

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.

SEXO	EDAD_AÑOS	ETNIA	GLOSA_PAIS_ORIGEN	COMUNA_RESIDENCIA	GLO
MUJER	2	No se identifica ...	Chileno	05109	Viña

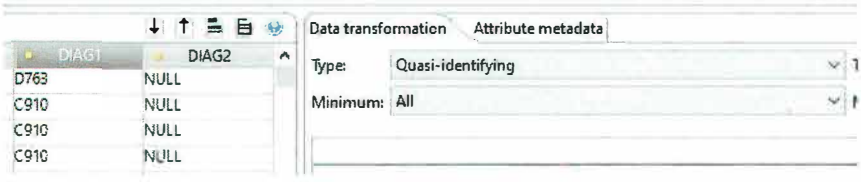
Como se indica en la imagen siguiente, al costado derecho de la pantalla se encuentra un set de colores, los que se utilizan para cambiar el total de las columnas al tipo de dato que representan.



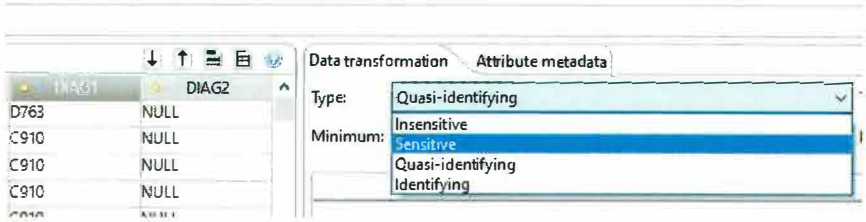
Los colores representan los tipos de datos Cuasi-identificador, identificador e insensible, respectivamente. Como la mayoría de los datos están clasificados como cuasi-identificadores, se hace clic en el circulo amarillo. Todos los colores de las columnas cambiarán a amarillo.

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

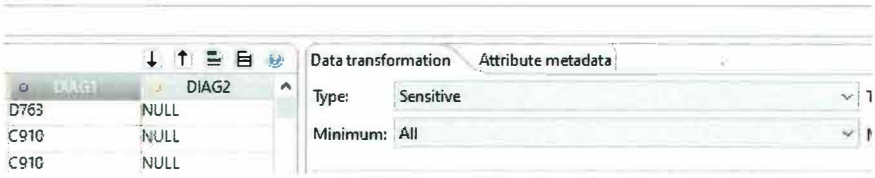
Ahora, para cambiar la clasificación de las columnas de diagnóstico, se hace clic en la columna respectiva del dato y se modifica el "Type" en la pantalla de la derecha.



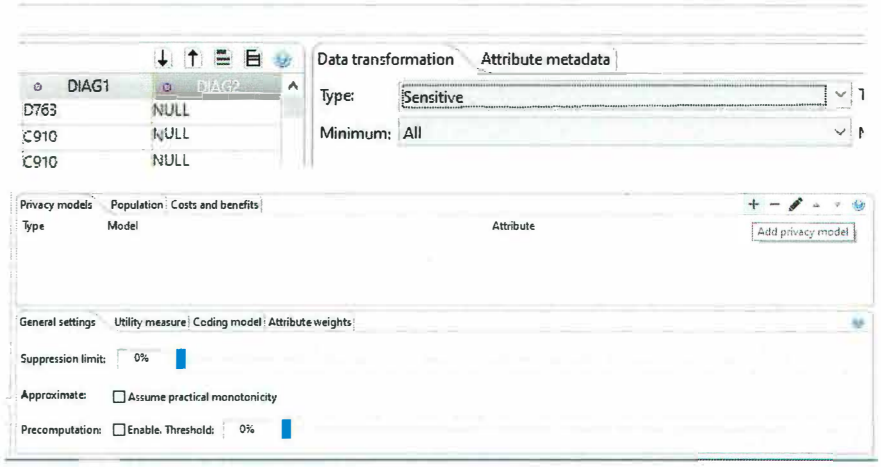
Se despliegan los tipos de datos disponibles, y se selecciona el tipo Sensitive.



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



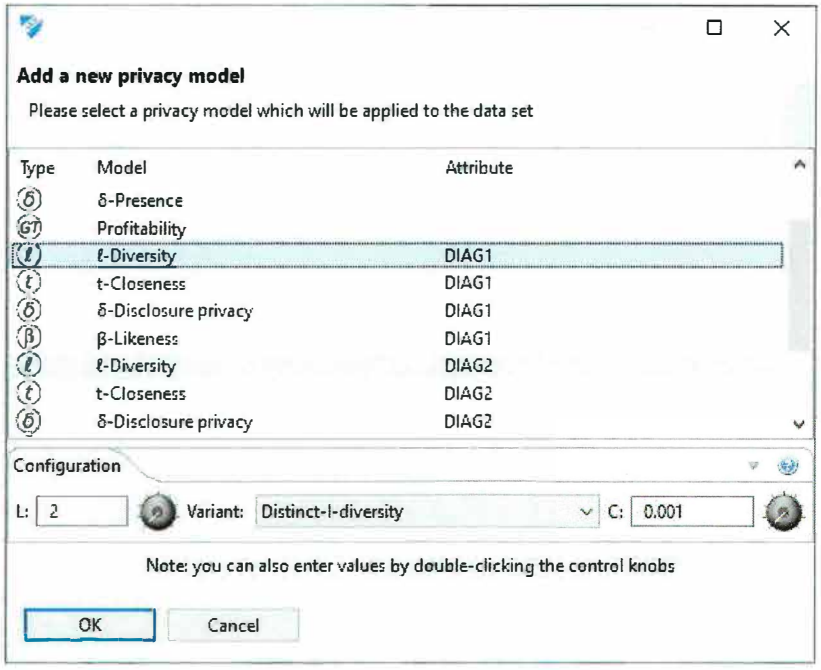
Se hace lo mismo con el DIAG2



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”.

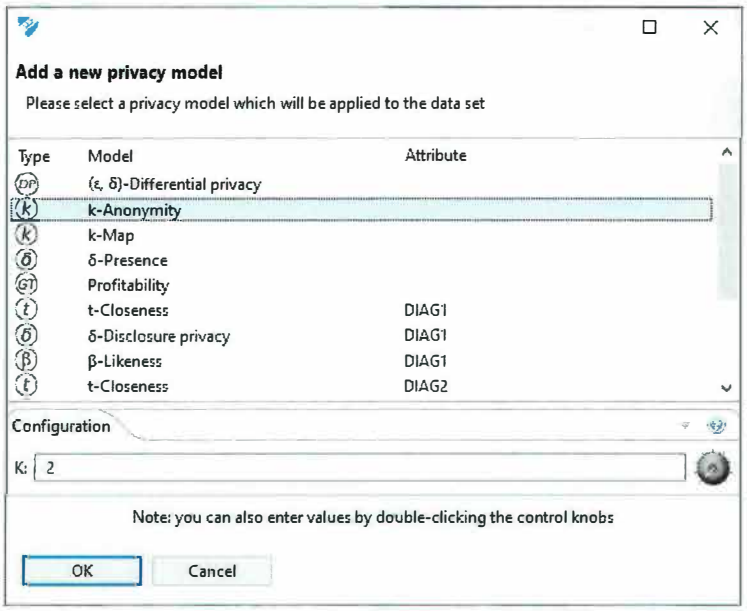
Se despliega una nueva ventana donde se selecciona el nombre del dato y el tipo de tratamiento. En este caso L-diversidad.



Hacemos clic en OK.

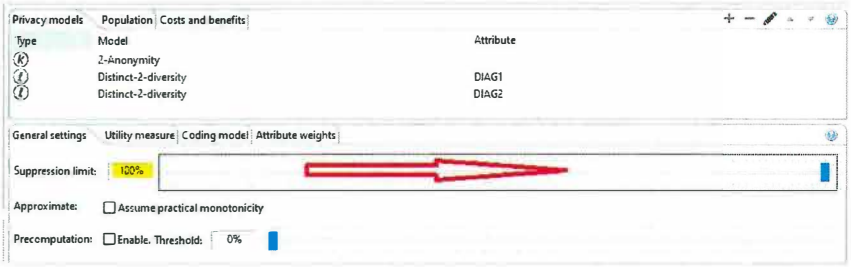
Se repite lo mismo para el DIAG2.

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



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.

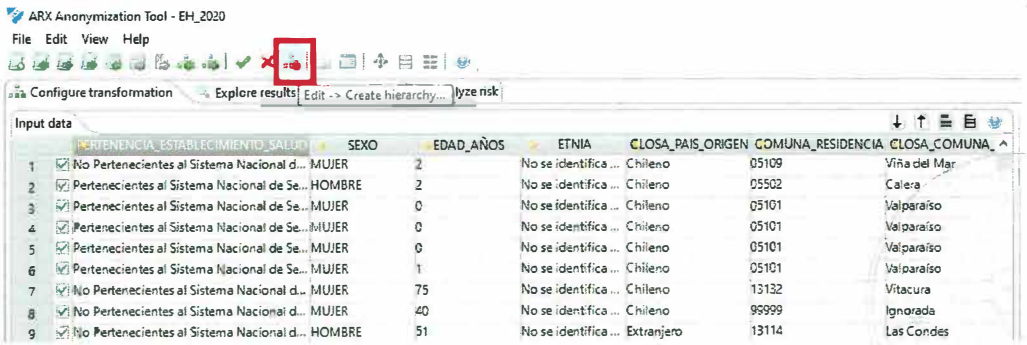


5. Creando jerarquías

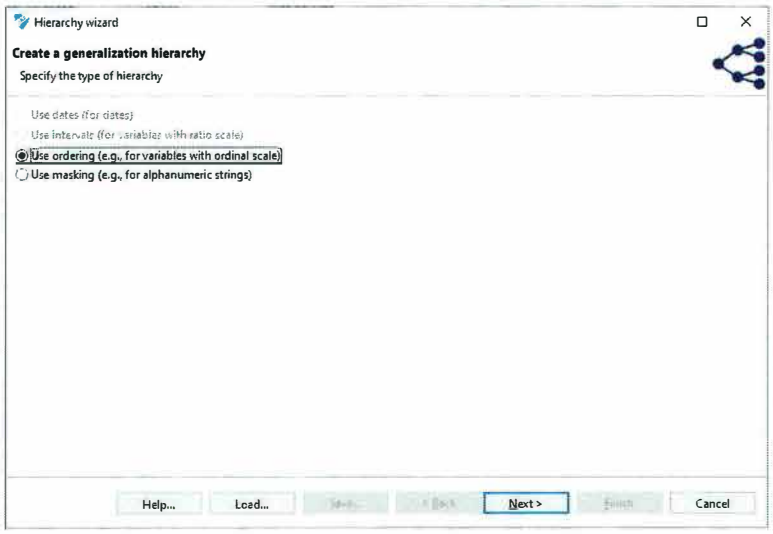
Cuando se tienen cuasi-identificadores que se pueden agregar en jerarquías o en un orden alfabético, se tiene la opción de crear jerarquías que permiten visibilizar y verificar los valores presentes. Además, se puede seleccionar la forma en que aparecen en el conjunto de resultado.

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.

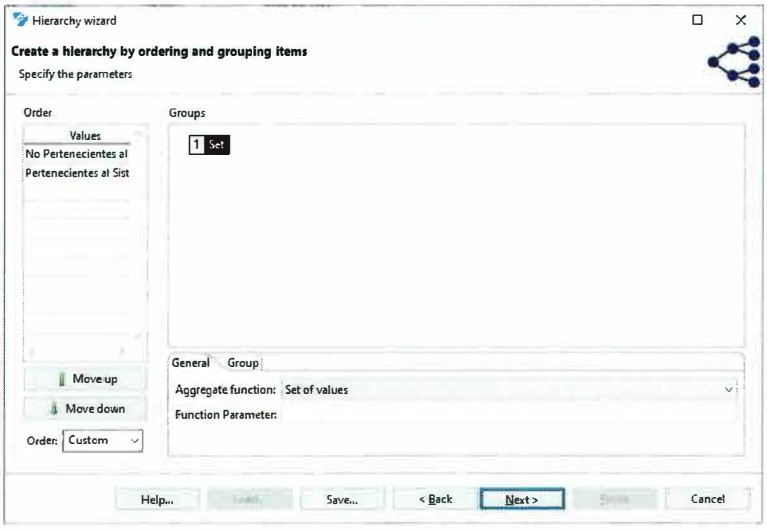


Se abrirá la ventana donde se debe seleccionar el tipo de datos que se va a utilizar y ordenar. En el ejemplo para pertenencia es “Use ordering”.



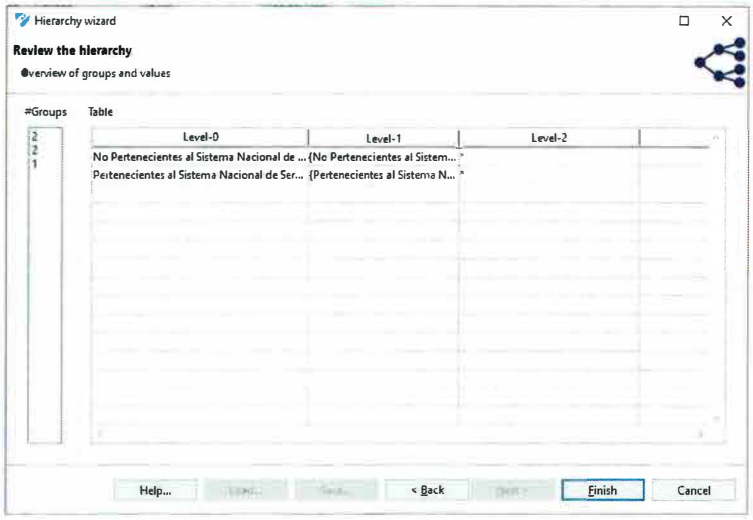
Clic en “Next >”.

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



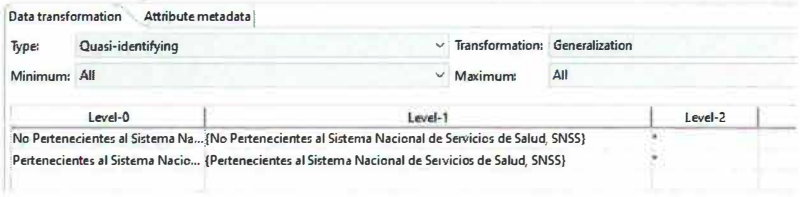
Hacemos clic en “Next >”.

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



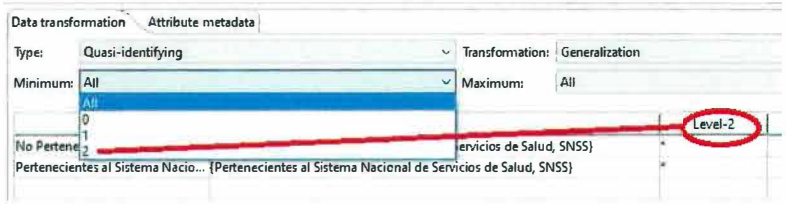
Se puede verificar que los valores se encuentren correctos y se hace clic en "Finish".

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

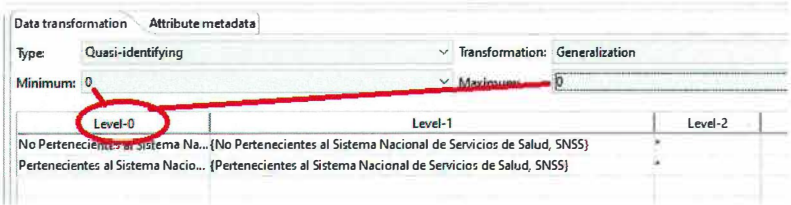


Al desplegar las alternativas de valores en el campo Minimum, se muestran los valores 0 al 2, correspondiente al valor del nivel (Level-0 al Level-2). Y All correspondiente al uso de todos los niveles. Esto se repite en el campo Maximum.

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



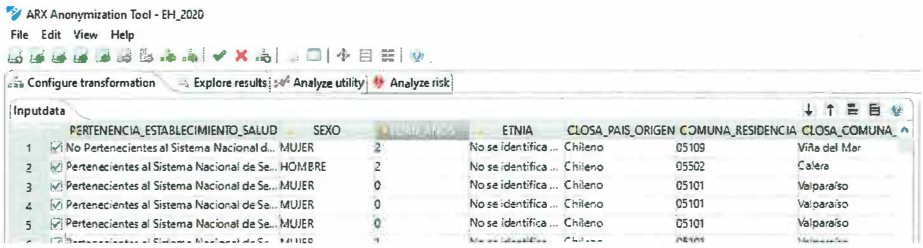
En este caso se utilizarán los valores asociados al Level-0, por lo que se selecciona el valor cero.



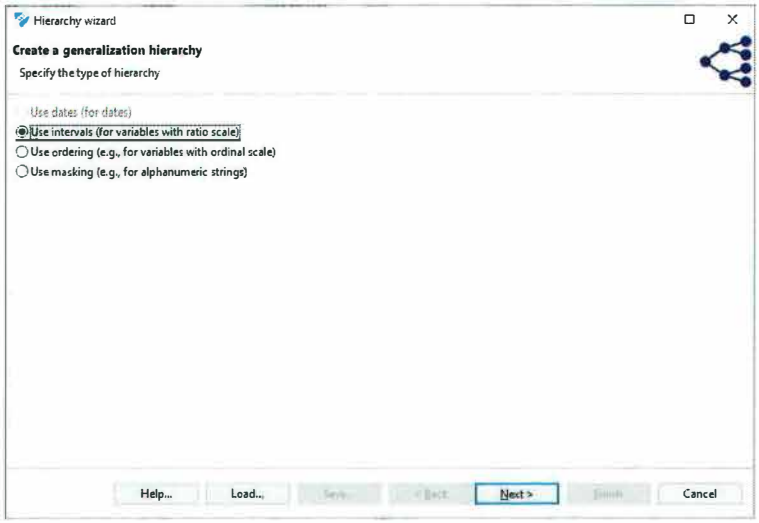
7. Jerarquía de intervalos

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.

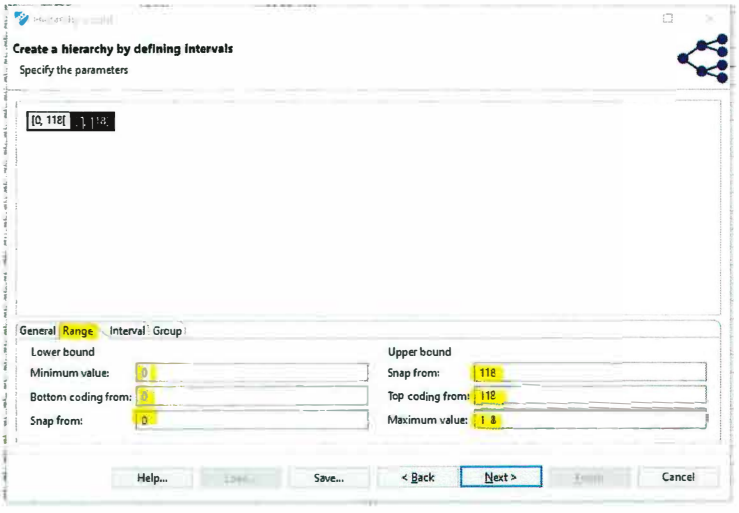


Seleccionaremos en la ventana siguiente “Use intervals”.

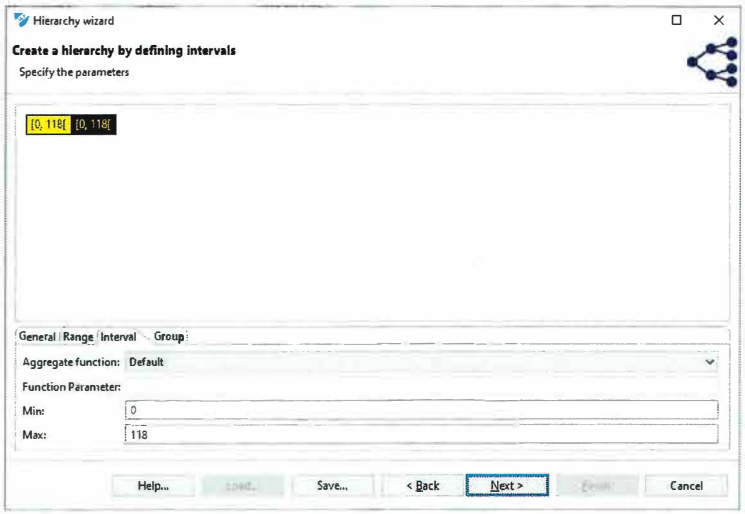


En la ventana siguiente se muestra un único set de valores desde 0 hasta 118 que es el mayor valor en la columna de datos.

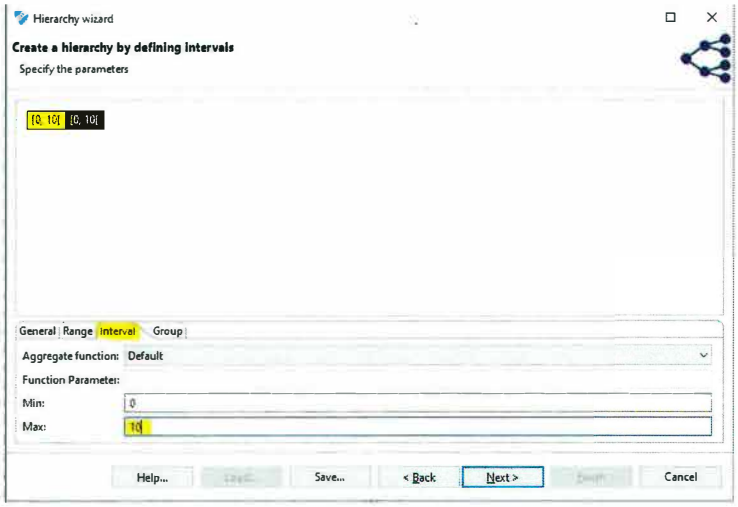
Se deben setear en la pestaña “Range” los valores lower y upper según los valores mínimos y máximos de años, o los máximos valores esperados.



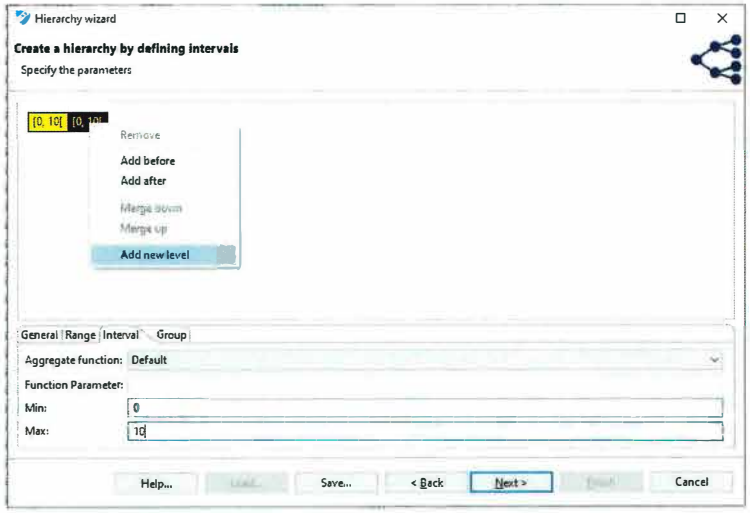
Para crear el primer intervalo de valores de datos, se hace clic en el rectángulo del set hasta que esté en color amarillo, y nos movemos a la pestaña “Interval”.



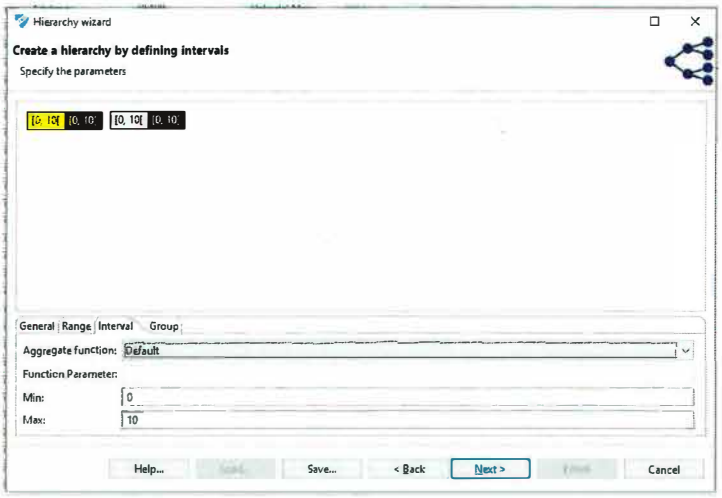
Actualmente el valor máximo del intervalo es 118, pero se quiere crear intervalos de 10 valores, por lo cual se establece el valor Max en 10.



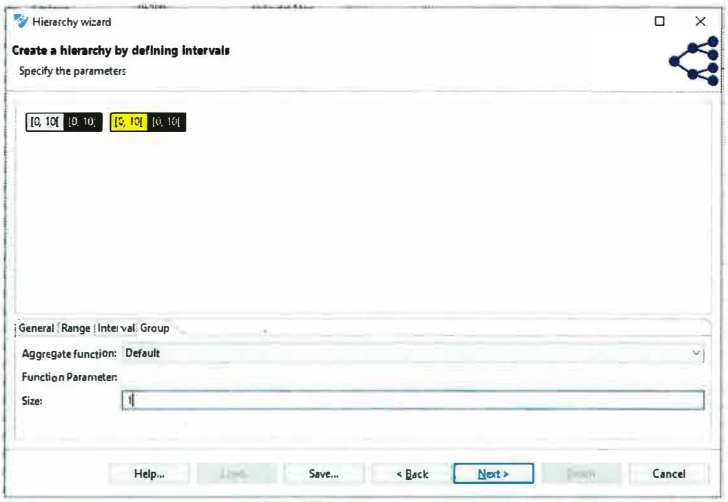
Para crear el siguiente nivel, se hace clic derecho en el rectángulo del primer intervalo, y seleccionamos “Add new level”.



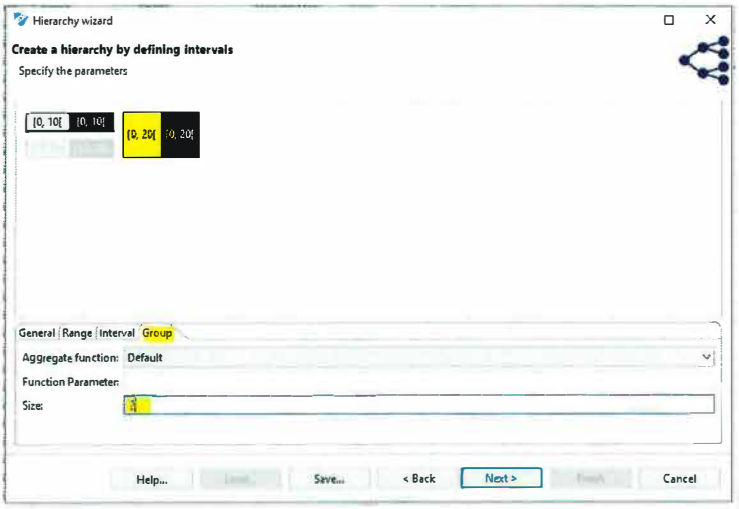
Se crea un nuevo nivel con las mismas características del anterior.



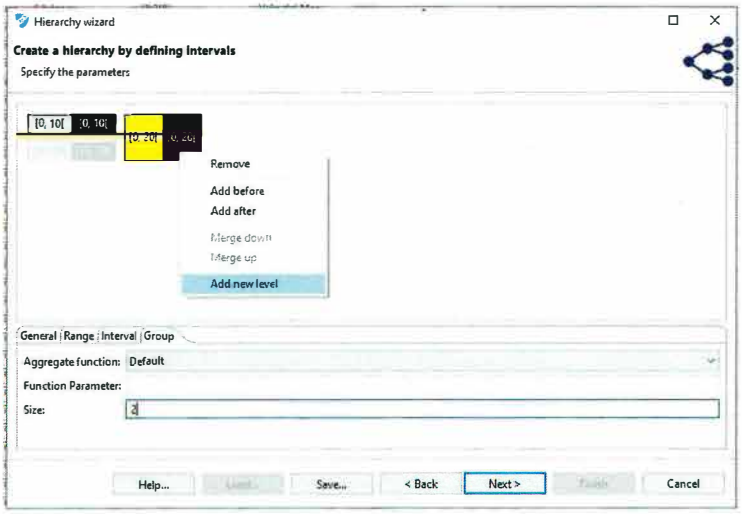
Se hace clic sobre el nuevo nivel hasta que quede en color amarillo.



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

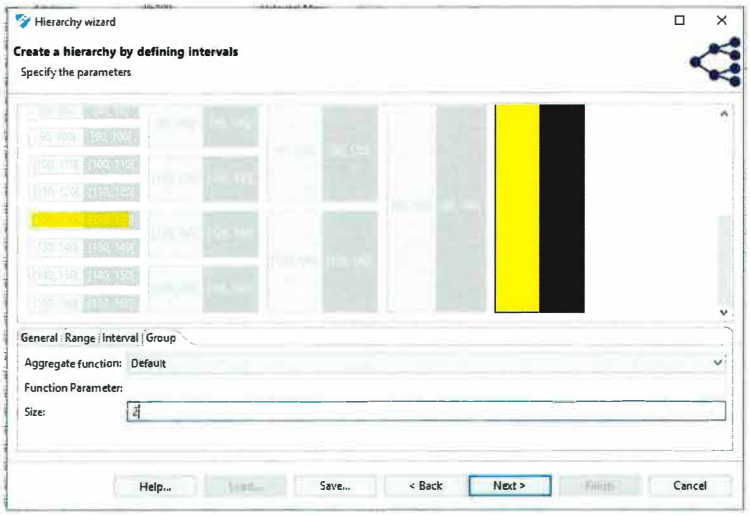


El nuevo set quedará con el doble de valores del set anterior. Para crear el siguiente nivel, se selecciona el ultimo nivel creado, se hace clic derecho sobre el set y se selecciona “Add new level”.



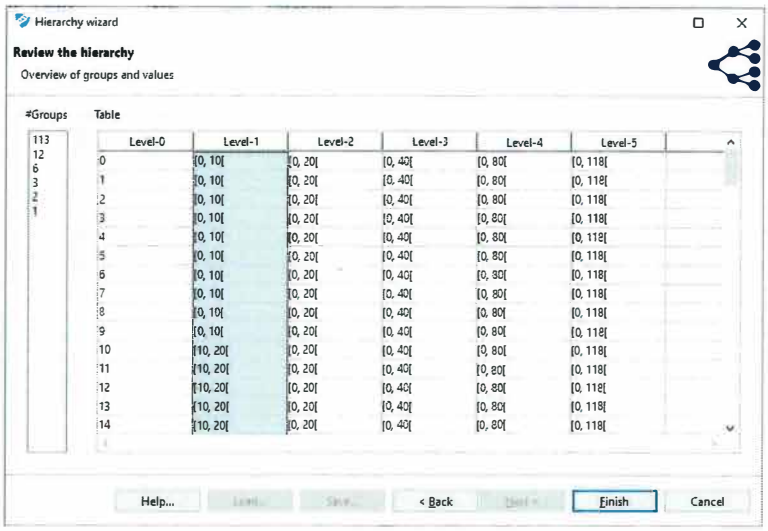
Nuevamente, se fija en 2 el valor de Size en la pestaña “Group”.

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

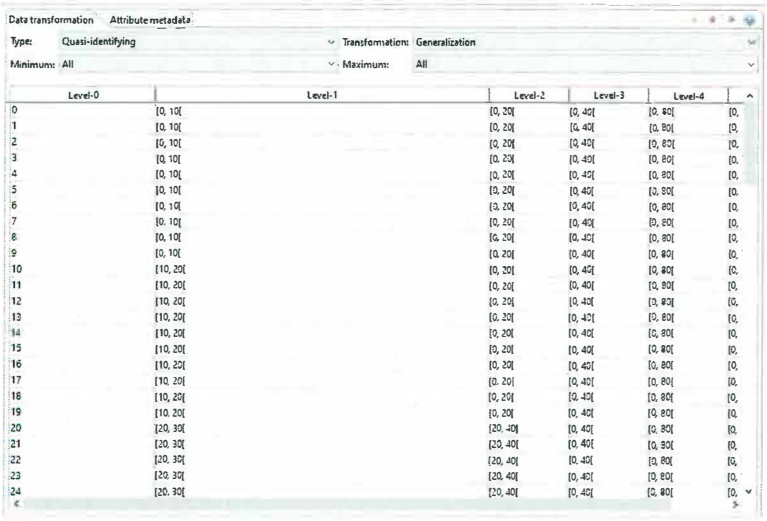


Hacemos clic en “Next >”.

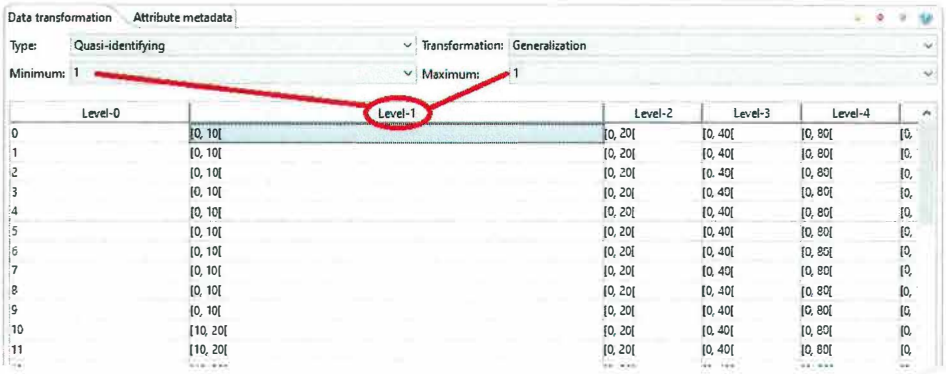
Se mostrarán distintos niveles de intervalos.



Se presiona “Finish”.

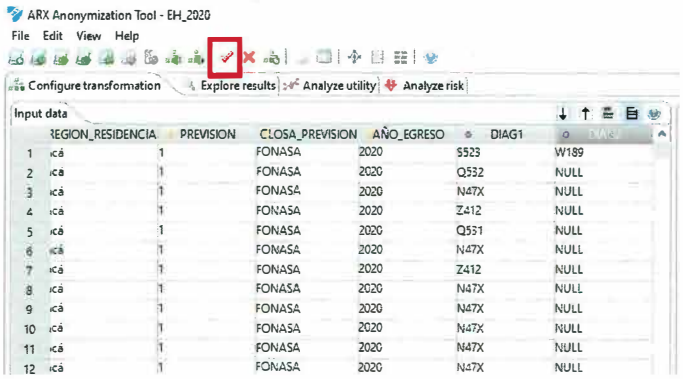


Se selecciona el valor mínimum y máximo del nivel que queremos mostrar en el proceso. En este caso el intervalo de 10 en 10 que está en el nivel 1.

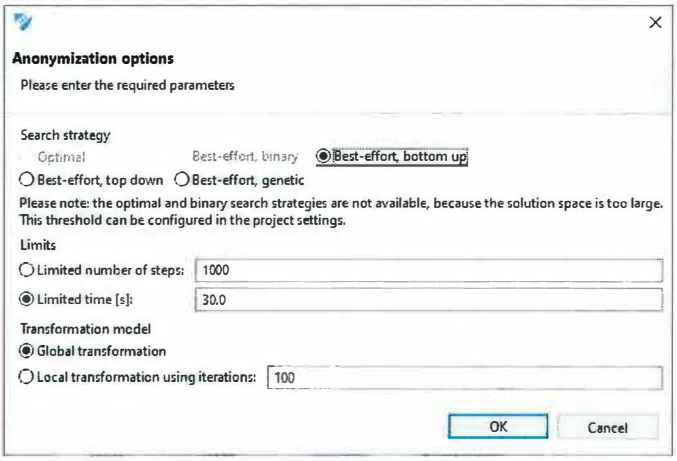


Para generar el proceso de anonimización no es necesario crear jerarquías para cada columna, por lo que se usa principalmente cuando se quiere agrupar valores o crear intervalos. Ahí si es necesario crear la jerarquía respectiva.

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.

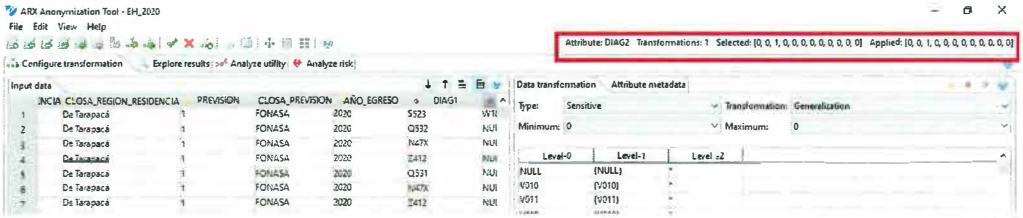


Se mostrará la siguiente ventana en la cual haremos clic en OK.



Comenzará el proceso de anonimidad.

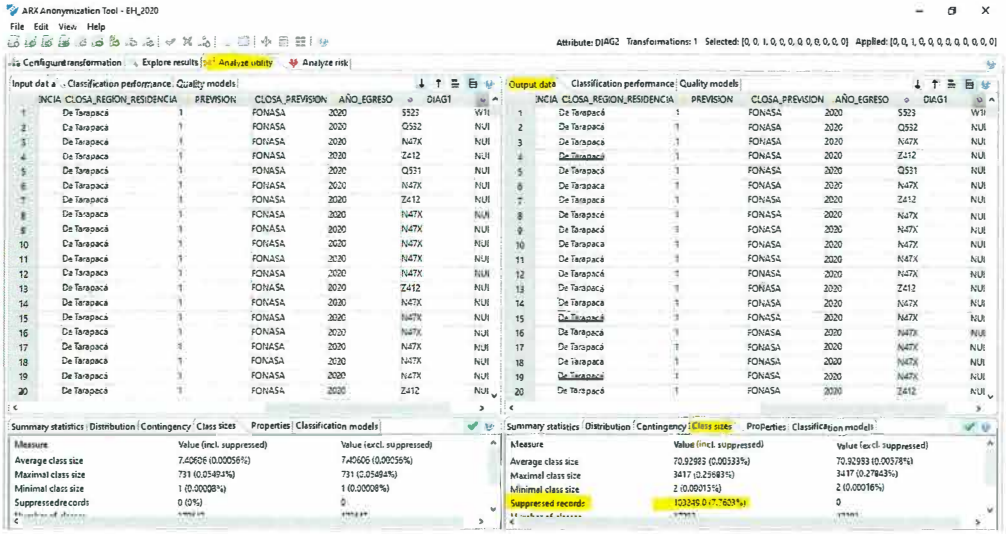
El proceso finaliza cuando aparece el resultado como en el recuadro de la imagen siguiente.



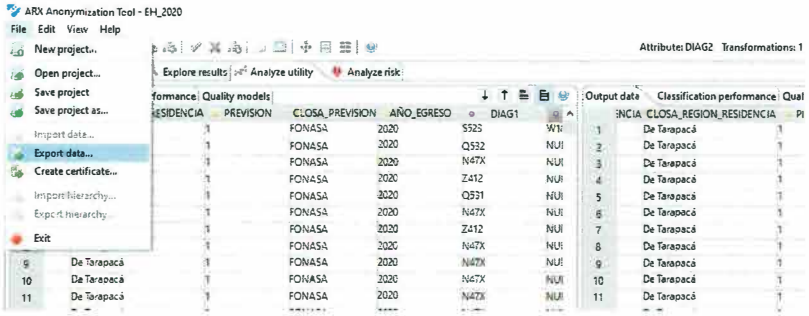
Para analizar el resultado hacemos clic en la pestaña de “Analyze utility”. Ahí podremos ver el conjunto resultante en la pestaña “Output data”. Además, en la parte inferior en la pestaña “Class sizes”,

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.

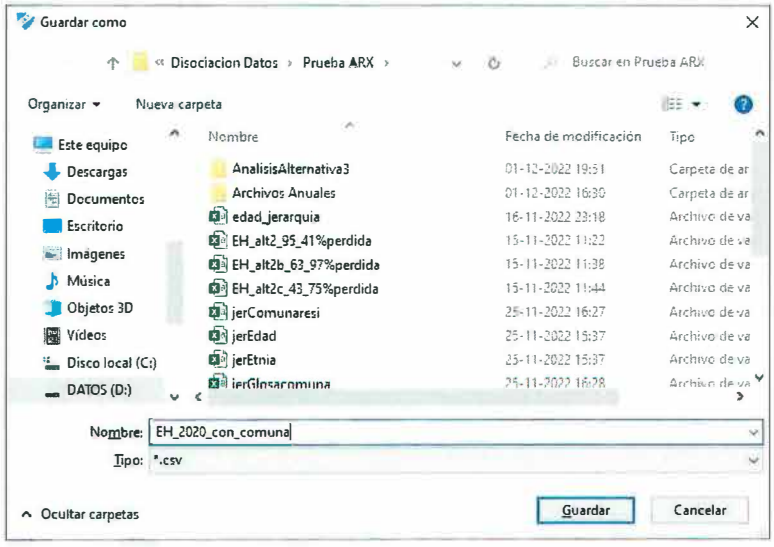


Los datos anonimizados se pueden exportar haciendo clic en “Export data” del menú “File”.



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

Ahí damos por finalizado el proceso de anonimización.



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:

```
base = base %>%
  mutate(cod_comuna_primer_nivel = codigo_comuna,
         cod_comuna_segundo_nivel = paste0(str_sub(codigo_comuna, 1, 3), "**"),
         cod_comuna_tercer_nivel = paste0(str_sub(codigo_comuna, 1, 2), "***"))
```

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:

```
base = base %>%
  select(-codigo_comuna, -K_primer_nivel, -L_primer_nivel,
         -K_segundo_nivel, -L_segundo_nivel,
         -K_tercer_nivel, -L_tercer_nivel)
```

4. Exportar la base

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