# Estructuración de la información clínica electrónica para el uso en ciencia de datos en un hospital de tercer nivel de complejidad

Isabel Olaya López

Informe de práctica para optar al título de Bioingeniera
Modalidad semestre de Industria

Asesor
Jenny Kateryne Aristizábal Nieto, Magíster (MSc) en Ingeniería Biomédica

Universidad de Antioquia
Facultad de ingeniería
Bioingeniería
Medellín, Antioquia, Colombia
2023

| Cita | | Olaya López [1] |
|---|---|---|
| **Referencia**<br><br>Estilo IEEE (2020) | [1] | I. Olaya López, “Estructuración de la información clínica electrónica para el uso en ciencia de datos en un hospital de tercer nivel de complejidad”, Práctica empresarial, Bioingeniería, Universidad de Antioquia, Medellín, Colombia, 2023. |



Coordinador de prácticas académicas Bioingeniería: Javier Hernando García Ramos.

Asesor interno: Jenny Kateryne Aristizábal Nieto.

Asesor externo: Alejandra María Restrepo Franco

Lugar de prácticas: SIGMA Ingeniería S.A

Apoyado por: Centro de desarrollo tecnológico INNVESTIGA y S.E.S Hospital Universitario de Caldas.

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

**Rector:** John Jairo Arboleda Céspedes.

**Decano/Director:** Julio César Saldarriaga Molina.

**Jefe departamento:** John Fredy Ochoa Gómez.

*A Dios y a mi familia,*
*que me han acompañado y apoyado siempre.*

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

# SIGLAS, ACRÓNIMOS Y ABREVIATURAS

| | |
|---|---|
| **EHR** | Electronic Health Record |
| **CDM** | Common data model |
| **OHDSI** | Observational Health Data Sciences and Informatics |
| **OMOP** | Observational Medical Outcomes Partnership |
| **SQL** | Structured Query Language |
| **ETL** | Extract, Transform and Load |
| **DDL** | Data Definition Language |
| **MIMIC** | Medical Information Mart for Intensive Care |
| **PCORnet** | National Patient-Centered Clinical Research Network |
| **I2b2** | Informatics for Integrating Biology and Bedside |
| **CER** | Comparative Effectiveness Research |
| **PCORI** | Patient-Centered Outcomes Research Institute |
| **HIPAA** | Health Insurance Portability and Accountability Act |
| **PHI** | Protected Health Information |
| **PII** | Personally Identifiable Information |
| **TI** | Tecnologías de la información |
| **EPS** | Entidades Promotoras de Salud |
| **IP** | Internet Protocol |

## RESUMEN

Un desafío importante en Colombia es adaptar los datos clínicos de manera adecuada para ser usados en ciencia de datos, ya que actualmente no existe un consenso en la adopción de un modelo específico para esto. Internacionalmente, la transformación de los datos clínicos a un modelo de datos común (CDM, por sus siglas en ingles), ha demostrado ser una opción válida para facilitar las investigaciones en ciencia de datos. Por esto, el presente proyecto tuvo como objetivo estructurar la información clínica almacenada en los sistemas de información del S.E.S. Hospital Universitario de Caldas para obtener una base de datos confiable que permita el desarrollo de aplicaciones en ciencia de datos. Para el cumplimiento de este objetivo, primero se realizó un análisis exploratorio de los datos del hospital, luego se seleccionó un CDM adecuado para el hospital, se diseñó e implementó el proceso de extracción, transformación y carga de datos y, por último, se elaboró una guía para el proceso de conversión de los datos y un protocolo de anonimización. Como resultado, se logró realizar la conversión de los datos del hospital a las tres primeras tablas del CDM de la asociación de resultados médicos observacionales (OMOP), que fue la estructura seleccionada debido a que presentó las mejores características entre todos los modelos de datos revisados. Se concluyó, gracias al análisis exploratorio inicial realizado, que la estructuración y armonización de los datos es vital para futuras investigaciones con los datos y que el CDM de OMOP es una opción adecuada para esto.


***Palabras clave*** — **EHR, CDM, transformación de datos, armonización de datos, OMOP.**

## ABSTRACT

An important challenge in Colombia is to adapt clinical data in a suitable way to be used in data science, as there is currently no consensus on the adoption of a specific model for this. Internationally, the transformation of clinical data into a common data model (CDM) has proven to be a valid option to facilitate data science research. Therefore, the objective of this project was to structure the clinical information stored in the information systems of the S.E.S. Hospital Universitario de Caldas in order to have a reliable database that allows the development of data science applications. To achieve this objective, first an exploratory analysis of the hospital's data was carried out, then a common data model (CDM) suitable for the hospital was selected, the data extraction, transformation and loading (ETL) process was designed and implemented, and finally, a guide for the data conversion process and an anonymization protocol were developed. As a result, the conversion of the hospital data to the first three tables of the Observational Medical Outcomes Partnership (OMOP) CDM was achieved, which was the structure selected because it presented the best characteristics among all the data models reviewed. It was concluded from the initial exploratory analysis performed that data structuring and harmonization is vital for future research with the data and that the OMOP CDM is a suitable option for this.


***Keywords*** — **EHR, CDM, data transformation, data harmonization, OMOP.**

# I. INTRODUCCIÓN

En los últimos años, se ha visto un aumento de la adopción de registros médicos electrónicos (EHR por sus siglas en inglés) a nivel nacional e internacional. Muchas instituciones médicas han comenzado a adoptar EHR, lo que ha provocado que la cantidad de datos en EHR aumente considerablemente [1]. Estos sistemas EHR almacenan datos asociados a cada encuentro médico con el paciente, incluida información demográfica, diagnósticos pasados y actuales, pruebas y resultados de laboratorio, prescripciones, imágenes radiológicas, notas clínicas y más [2].

La adopción de EHR por parte de las instituciones, se realizó principalmente para mejorar la atención médica desde un punto de vista operativo. Sin embargo, muchos científicos han encontrado un uso secundario para aplicaciones de informática clínica [3], dado que contienen datos clínicos coleccionados durante años o décadas, que de otra forma sería desgastante recopilar, en términos de tiempo y costo [4]. Dentro de las aplicaciones para estos datos, está la vigilancia automática de infecciones asociadas a la atención médica, la detección y exploración de eventos adversos de medicamentos, la asignación de códigos de diagnóstico, la minería de texto en el dominio del cáncer, el modelado temporal de eventos clínicos, la simplificación de texto de narrativas clínicas y el análisis de comorbilidad, entre otras [5].

Para utilizar este material de manera más eficiente se requiere que los datos estén de manera estructurada y ordenada para que puedan ser empleados por técnicas como el procesamiento del lenguaje natural (NLP) y algoritmos de aprendizaje profundo. Esto representa una dificultad importante, pues actualmente no hay un consenso en la adopción de un patrón específico de los EHR, donde cada EHR cuenta con su propia forma de organización de datos [1] y aproximadamente el 80% de los datos médicos se registran de forma no estructurada, lo que complica su uso en comparación con los datos estructurados y listos para usar [6].

Un desafío importante para el tema de la estructuración de los EHR en Colombia es adaptar los datos regionales de manera adecuada para el desarrollo en tecnología en ciencia de datos, como es el caso de la ciudad de Manizales, con uno de los hospitales con más impacto en la región que es el S.E.S Hospital Universitario de Caldas, el cual ha enfocado sus procesos de investigación e

innovación en el desarrollo de proyectos en áreas específicas que conduzcan a mejorar los modelos asistenciales del hospital. Actualmente, la institución viene desarrollando diferentes procesos desde la aplicación de técnicas en IA y ciencia de datos a sus procesos asistenciales, lo que ha arrojado aprendizajes valiosos enfocados en la importancia de la estructura global de los datos disponibles para la generación de modelos óptimos de ciencia de datos, lo que pone en evidencia la necesidad de estructurar la información clínica del hospital.

Por tanto, este proyecto planteó estructurar la información clínica almacenada en los sistemas de información del S.E.S Hospital Universitario de Caldas para obtener bases de datos sólidas y confiables que permitan el desarrollo de aplicaciones en ciencia de datos. Esto, con el fin de contribuir a que la gran cantidad de datos generados por los registros médicos electrónicos estén disponibles en el momento oportuno y en el formato adecuado para su análisis por parte del personal científico.

En el presente documento, se muestra el proceso seguido para lograr la conversión de la base de datos del S.E.S. Hospital Universitario de Caldas en el modelo de datos común de la asociación de resultados médicos observacionales (OMOP). Lo cual incluye primero, la realización de un análisis exploratorio de los datos hospitalarios, que demuestra la importancia del proyecto para futuras investigaciones. Posteriormente, se ilustra el proceso de selección del CDM, donde se evidencian las razones para elegir el CDM de OMOP. Luego, se muestra el diseño y la implementación del proceso de extracción, transformación y carga de datos para las tres primeras tablas del modelo. Por último, se brinda una guía para que el personal encargado de estos procesos conozca el modelo y pueda realizar la conversión de las tablas faltantes. Como complemento a lo anterior, se proporciona un protocolo de anonimización necesario para cumplir con las leyes requeridas a la hora de compartir datos clínicos.

# II. OBJETIVOS

## *A. Objetivo general*

Estructurar la información clínica electrónica almacenada en diferentes sistemas de información del S.E.S Hospital Universitario de Caldas para obtener bases de datos confiables que permitan el desarrollo de aplicaciones en ciencia de datos.

## *B. Objetivos específicos*

- Realizar un análisis exploratorio de datos (EDA) de las historias clínicas del S.E.S Hospital Universitario de Caldas para conocer la estructura y distribución de las bases de datos.
- Estructurar la información clínica relevante a partir del diseño de un modelo de datos adecuado para aplicaciones en ciencia de datos.
- Diseñar un protocolo de anonimización de datos adecuado para aplicaciones en ciencia de datos en el S.E.S Hospital Universitario de Caldas.
- Elaborar una guía de usuarios para familiarizar a los investigadores con la nueva estructura de los datos y la forma de consultarlos.

# III. MARCO TEÓRICO

## *A. Ciencia de datos en salud*

De forma general la ciencia de datos se puede definir como una ciencia interdisciplinaria, que tiene como principio la aplicación de métodos cuantitativos y cualitativos con el fin de analizar y extraer datos, información y conocimiento, para resolver problemas relevantes y predecir resultados. Esto, a través de prácticas del campo de las matemáticas, estadística, inteligencia artificial e ingeniería de la computación, que permiten responder preguntas como "qué pasó", "por qué pasó", "qué pasará" y "qué se puede hacer con los resultados" [7].

La ciencia de datos estudia la información de cuatro maneras principales que son [8]:

- Análisis descriptivo
- Análisis de diagnóstico
- Análisis predictivo
- Análisis prescriptivo

En el sector de la salud la ciencia de datos tiene un gran potencial a la hora de abordar los mayores desafíos de la atención médica moderna que existen en varios niveles diferentes, lo que eventualmente contribuirá a lograr avances en la salud global [9]. Esta, puede conseguir conducir a análisis predictivos inmediatos que se pueden usar para obtener información sobre varios procesos de enfermedades y proporcionar tratamientos centrados en el paciente. Además, puede ayudar a mejorar las habilidades de los investigadores en ciencia, estudios epidemiológicos, medicina personalizada y otros campos. Las organizaciones de atención médica de hoy pueden transformar la terapia médica y la medicina personalizada al integrar datos biomédicos y de salud. Asimismo, la ciencia de datos puede ayudar a administrar, evaluar e interpretar de manera efectiva los grandes datos, abriendo nuevos caminos en la atención médica integral [10].

### *B. Registros Electrónicos de Salud – EHR*

La historia clínica del paciente es el registro longitudinal de todos los aspectos de atención médica (clínica, científica, técnica, administrativa), en todas las etapas de la atención (promoción, prevención, diagnóstico, tratamiento, rehabilitación y paliación), y en sus aspectos biológicos, psicológicos, y sociales, que es creado, mantenido, y utilizado por el personal de salud para administrar y proveer servicios de salud al individuo. Lo cual a nivel electrónico se expresa en el término historia electrónica de salud soportada sobre Registros Electrónicos de Salud- EHR [11].

A nivel mundial se han incorporado varias notaciones y definiciones como la de historia electrónica de salud (EHR – *Electronic Health Record*), Historia clínica o medica electrónica -HCE- (en inglés EMR-*Electronic Medical Record*), o de Historia electrónica de pacientes (en inglés EPR-*Electronic Patient Record*), que se relacionan con la evolución del concepto, pero también con la definición de niveles de funcionalidad. En Colombia se utiliza el concepto de EHR para los efectos de interoperabilidad de datos de la historia clínica [11].

### *C. Bases de datos clínicas*

Una base de datos se define como un conjunto de información recolectada regularmente, organizada de tal manera que se pueda acceder, manipular y actualizar fácilmente [12]. Estas bases de datos se comenzaron a establecer gracias a que las instituciones archivaban y organizaban datos en repositorios centrales [13]. Lo que ha dado como resultado nuevas formas de presentar, comprender y utilizar los datos de atención médica.

A diferencia de los repositorios de enfermedades, estas bases de datos clínicas se caracterizan por tener datos heterogéneos a nivel de paciente, recopilados automáticamente de los EHR. Los datos de los principales servicios auxiliares, incluidas farmacias, laboratorios y estudios de radiología, se combinan con diferentes componentes de la atención clínica (por ejemplo, planes de atención de enfermería, registros, etc.). Esto da como resultado una serie de variables de alta resolución que se originan en un gran número de pacientes, lo que permite a los investigadores

estudiar tanto las interacciones clínicas como las decisiones para una amplia gama de procesos patológicos [12].

La creación de una base de datos de atención médica tiene como objetivo principal guardar la información clínica en un formato que pueda explorarse intuitivamente y procesarse rápidamente, lo que permite a los investigadores extraer conocimiento valioso de los datos [12]. Actualmente existen numerosas bases de datos clínicas, algunas de las cuales se crearon con el objetivo de impulsar la colaboración en el análisis secundario de historias clínicas electrónicas, a través de la creación de repositorios abiertos. Dos ejemplos de esto son el "*Medical Information Mart for Intensive Care*" (MIMIC) [14] y la base de datos de investigación colaborativa eICU [15].

### *D. Modelo de datos común*

Un modelo de datos común (CDM por sus siglas en inglés), es una estructura estandarizada y acordada, que se utiliza para guardar información en una base de datos con vocabularios, terminologías y esquemas de codificación comunes [16]. Para lograr esto, los CDMs especifican características como [17]:

- Tablas del modelo, con sus respectivos nombres y campos.
- Tipos de datos (por ejemplo: 'date', 'integer', 'character', 'time', entre otros)
- Restricciones (¿Son permitidos los valores nulos? ¿Cada valor debe ser único?)
- Relaciones entre filas de datos (entidades) (¿Se tiene una relación de uno a muchos, o de uno a uno? ¿Se pueden representar jerarquías que definen conjuntos de conceptos?)
- Definiciones, procedimientos y supuestos de metadatos que describen el significado y el uso previstos de cada elemento de datos, cómo se recopilarán los datos, los valores o rangos permitidos y las dependencias entre los elementos de datos.

En las investigaciones clínicas el uso de CDMs es cada vez más común, ya que permiten que los investigadores reciban datos de múltiples fuentes como lo son los EHR, documentos

administrativos, registros, entre otros. Lo que brinda el acceso a una mayor cantidad de datos del mundo real y a colaboraciones de mayor escala [16].

Actualmente, existen diversas iniciativas y redes de investigación en salud que utilizan CDMs, lo que les permite tener múltiples bases de datos para realizar estudios reproducibles utilizando una metodología de análisis estandarizada [12]. Dos ejemplos de esto son la colaboración "*Observational Health Data Sciences and Informatics*" (OHDSI) que utiliza el CDM creado por la Asociación de resultados médicos observacionales (OMOP por sus siglas en inglés) [18] y la fundación "*Informatics for Integrating Biology and Bedside"* (i2b2) [19].

### *E. Extracción, transformación y carga de datos*

El proceso de extracción, transformación y carga de datos (ETL por sus siglas en inglés) es un método para migrar datos de una o más fuentes de datos heterogéneos a otras bases de datos, almacenes o repositorios [20]. Como su nombre lo indica consta de tres fases:

- Fase de extracción, la cual tiene como finalidad extraer los datos desde los sistemas de origen, ocupándose principalmente de la heterogeneidad técnica de las diferentes fuentes para importar datos relevantes. En esta etapa se verifica que los datos extraídos cumplan con los requisitos esperados [21].
- Fase de transformación, que es el núcleo del proceso ETL. En esta, se aplican una serie de reglas o algoritmos sobre los datos extraídos para convertirlos en un modelo o esquema de datos común. Se emplean métodos de limpieza y depuración, así como directrices que pueden ser declarativas, basarse en excepciones o restricciones [21].
- Fase de carga, la última fase del proceso, donde los datos procedentes de la fase de transformación son cargados en el sistema de destino. Dependiendo de los requerimientos del proyecto, este proceso puede abarcar una amplia variedad de acciones diferentes. Dos formas básicas para desarrollar este proceso son la acumulación simple y *Rolling*. Donde con la primera se hace la carga como una única transacción y con la segunda se cargan los datos por distintos niveles [21].

# IV. METODOLOGÍA

Para el desarrollo del presente proyecto, se planteó la metodología mostrada en el esquema de la Fig. 1, la cual comprende el conjunto de actividades secuenciales que permitieron la ejecución de los cuatro objetivos específicos y por ende el cumplimiento del objetivo general.

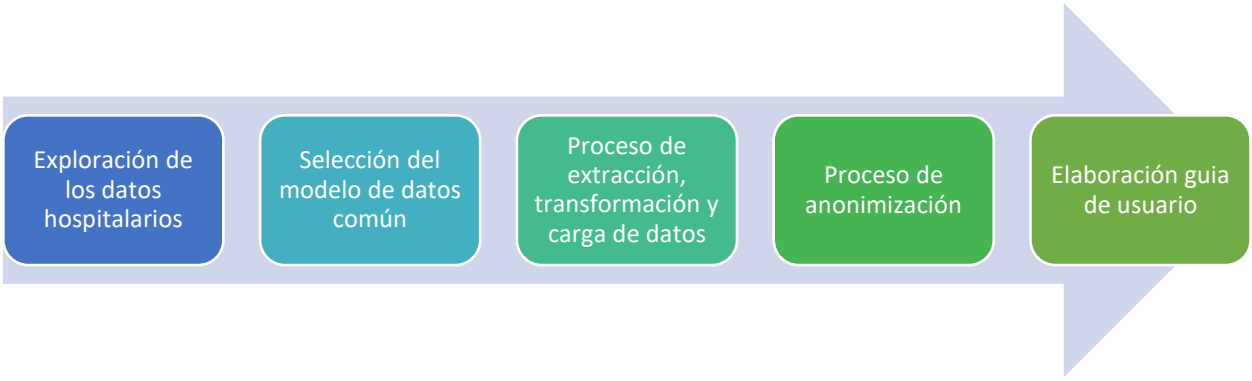



Fig. 1. Diagrama en bloques de la metodología general.

## *A. Exploración de los datos hospitalarios*

Para realizar la exploración inicial de la estructura y distribución de las bases de datos de historias clínicas del S.E.S Hospital Universitario de Caldas, fue necesario obtener la aprobación por parte del comité de ética del hospital, razón por la cual se realizó una reunión donde se expuso el proyecto y el uso de los datos. Una vez se obtuvo la aprobación del comité de ética, se ingresó a la base de datos de prueba del hospital mediante un acceso remoto con el software 'FortiClient'. El hospital tiene una base de datos relacional (SQL) y esta se maneja desde el software 'SQL Server 2014', por lo tanto, se descargó este mismo programa para acceder a los datos desde el computador propio. Para todo el proceso fueron proporcionados, por parte del hospital, usuarios y contraseñas específicas para el proyecto.

Desde 'SQL Server 2014', se hizo un reconocimiento inicial de la base de datos y luego con la herramienta 'White Rabbit' se realizó un escaneo de las tablas y campos contenidos en estas. En este escaneo, se reconocieron las variables, tipos de datos y cantidad de datos de cada tabla. En algunos casos se hicieron consultas adicionales en 'SQL Server' para conocer más de los datos.

### *B. Selección del CDM adecuado para implementar en el S.E.S Hospital Universitario de Caldas*

Se llevó a cabo una revisión sobre bases de datos clínicas libres en la comunidad científica para identificar cual era la mejor forma de estructurar la información del hospital para ser utilizada en ciencia de datos. Para esto, en las bases de datos bibliográficas (Lens, Scopus y Web of Science) se realizó una búsqueda de artículos de investigación con la siguiente ecuación:

((“evaluation” OR “analysis” OR “classification” OR “data structuring”) AND (“electronic health records” OR “EHR” OR “electronic medical records” OR “EMRs” OR “clinical data” OR “medical data information” OR “clinical information” OR “clinical databases” OR “EHRs”) AND (“common data model”))

Se realizó una identificación rápida de los artículos y se eliminaron los que estuvieran duplicados. Luego, se seleccionaron los artículos que contuvieran información relevante sobre bases de datos clínicas y modelos de datos comunes. En estos, se identificaron las bases de datos más nombradas y se extrajo información relevante como la estructura que presentaban (cantidad de tablas), la cantidad de pacientes registrados y si presentaban un CDM. Luego, se estudiaron los artículos que evaluaban los CDMs para saber cuál era el CDM con mejores características. Se revisaron algunos aspectos como las herramientas que proporcionaba cada modelo para hacer el mapeo y transformación de sus datos, la documentación que tenía disponible y las redes de apoyo con las que contaba. Luego, se seleccionaron 100 artículos aleatorios de la búsqueda inicial y se clasificaron de acuerdo al CDM que usaban o nombraban, para así tener una idea sobre la recepción de cada CDM en la comunidad científica.

Posteriormente, se realizó una reunión con el personal de tecnologías de la información (TI) del hospital para presentar la información encontrada sobre las bases de datos, los CDMs, las características y evaluaciones de estos y con base en ello se seleccionó el CDM para aplicar en el hospital.

*C. Extracción, transformación y carga de datos*

Luego de la selección del CDM, se diseñó e implementó un proceso para la extracción, transformación y carga de datos (ETL). En la Fig. 2 se muestra el resumen de este proceso.

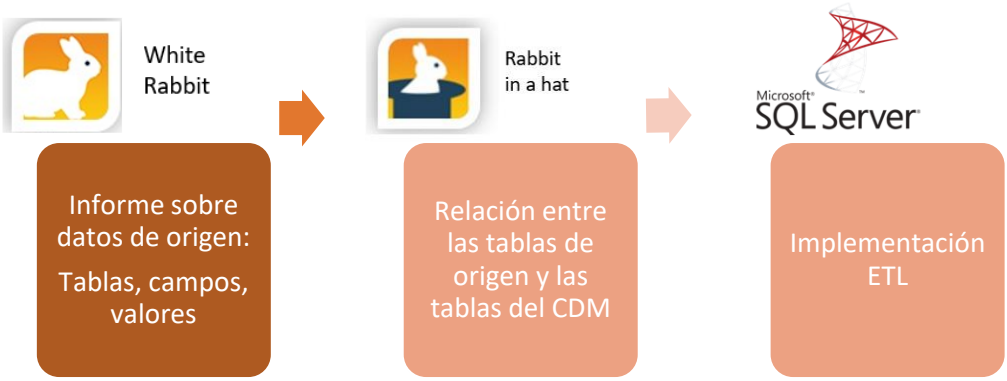



Fig. 2. Resumen proceso de extracción, transformación y carga de datos.

Para el diseño del ETL se utilizaron las herramientas 'White Rabbit' y 'Rabbit in a hat', que son softwares diseñados por colaboradores de OHDSI para facilitar la transformación de los datos al CDM de OMOP. Primero, con la ayuda de 'White Rabbit' se generó un informe sobre los datos de origen, es decir, información de las tablas, campos y valores de las bases de datos del hospital. Luego, este informe fue importado a la herramienta 'Rabbit in a hat', donde a través de su interfaz gráfica se hizo una relación entre la estructura de datos de origen a la estructura de datos del CDM. Desde 'Rabbit in a hat' se generó un informe con todas las relaciones de las estructuras entre bases de datos y los comentarios de la lógica necesaria para la conversión de los datos.

Una vez listo el informe de 'Rabbit in a hat', se descargaron los códigos encontrados en el repositorio de OHDSI: https://github.com/OHDSI/CommonDataModel/tree/v5.4.0, lo cuales contienen el lenguaje de definición de datos (DDL por sus siglas en inglés), las claves externas, las claves principales y los índices del CDM, es decir, toda la información necesaria para recrear la base de datos con el modelo de OMOP. Estos códigos se modificaron para ser utilizados en la base de datos del hospital y se ejecutaron en SQL server para crear la nueva base de datos en el servidor del hospital, con la estructura y tablas específicas del CDM.

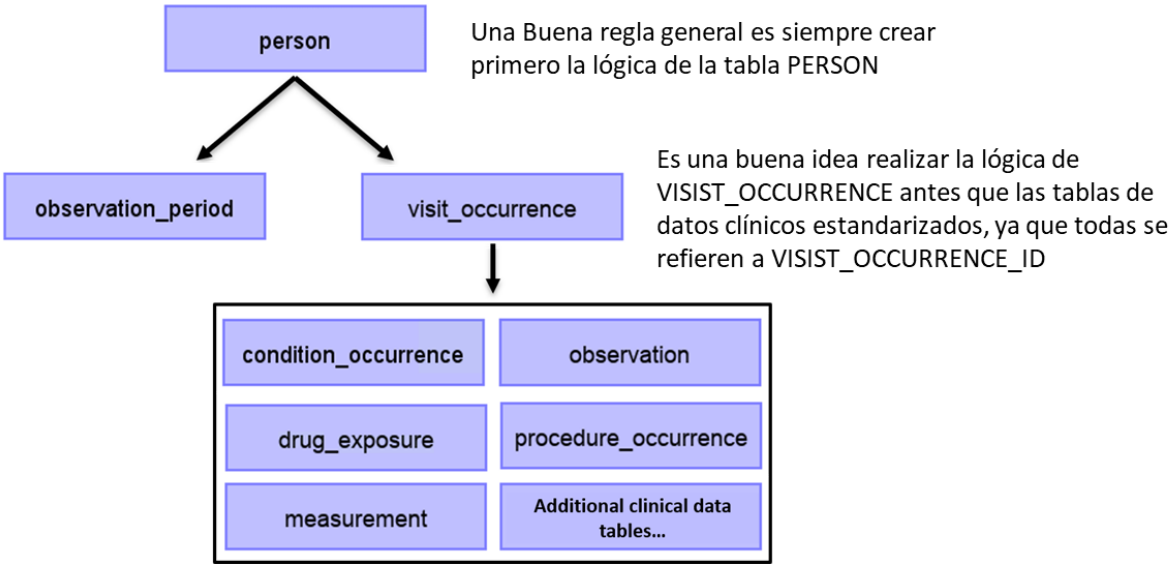



Fig. 3. Flujo general del ETL según el libro de OHDSI

Nota: Imagen adaptada de https://ohdsi.github.io/TheBookOfOhdsi/ExtractTransformLoad.html.

Según las recomendaciones del OHDSI el flujo general del proceso ETL es el mostrado en la Fig. 3, por lo tanto, se crearon nuevos scripts para insertar los datos de hospital en las tres primeras tablas del CDM (PERSON, OBSERVATION_PERIOD y VISIT_OCURRENCE). Se elaboró un script por cada tabla a llenar, en donde se identificaron las variables de la base de datos del hospital y se insertaron en las correspondientes tablas. Para esto, se tomó como ejemplo los scripts encontrados en el repositorio de OHDSI: https://github.com/OHDSI/ETL-Synthea, los cuales describen el proceso ETL de los datos de Synthea (un generador de pacientes sintéticos de código abierto que modela el historial médico de pacientes sintéticos).

Por último, se verificó si los datos se habían insertado correctamente y si se había perdido información. Para esto, se compararon la cantidad de registros de la tabla de origen con la del CDM mediante una consulta en SQL server que arrojó la cantidad de registros de cada tabla. Además, se hicieron consultas en SQL para cinco personas aleatorias que tuvieran ingresos y se confirmó si se había guardado la información de estas personas correctamente y si las fechas correspondían en ambas tablas.

*D. Proceso de anonimización*

Se realizó una revisión sobre las leyes internacionales que cumplía cada una de las iniciativas comparadas en el proceso de selección del CDM respecto a seguridad y privacidad de los datos. Además, se extrajeron los procesos que tenían los datos para cumplir con estas leyes. Posteriormente se buscó la normativa que aplicaba en Colombia para la seguridad y distribución de los datos médicos.

Sumado a lo anterior, se buscaron guías o protocolos de anonimización existentes en Colombia para tener bases teóricas sobre los procesos que debían realizarse con los datos y tomar algunas medidas necesarias a la hora de transformar los datos al CDM de OMOP.

Por último, se desarrolló un protocolo de anonimización para cuando los datos transformados vayan a compartiste. Dicho protocolo debe aplicarse una vez esté transformada toda la información, es decir, cuando todas las tablas de CDM de OMOP, estén listas.

*E. Elaboración guía de usuario*

Se elaboró un guía de usuarios donde se familiariza al investigador con el CDM de OMOP. En esta, se detalló todo el proceso de diseño e implementación del ETL para las personas encargadas de la transformación de los datos. Se especificaron las páginas web donde encontrar la información necesaria y los programas a descargar.

# V. RESULTADOS Y ANÁLISIS

## *A. Exploración de los datos hospitalarios*

Al realizar la exploración inicial de los datos desde 'SQL Server 2014' se notó que la base de datos del hospital tenía una gran cantidad de tablas, además había poca documentación de estas y no se contaba con los diagramas de entidad-relación correspondientes. Luego de realizar el escaneo con el software 'White Rabbit', se identificaron también varias tablas completamente vacías. En la TABLA I se puede observar la cantidad total de tablas en la base de datos, la cantidad de tablas que estaban completamente vacías y la cantidad de tablas que contaban con información. Así mismo, se muestra la cantidad de tablas que no pudieron ser leídas por la herramienta de análisis y que, al revisarlas, se observó que eran tablas que contenían formatos de imagen (un tipo de dato que el software no fue capaz de procesar).

TABLA I
CANTIDAD DE TABLAS DE LA BASE DE DATOS

| | |
|---|---|
| Vacías | 2400 |
| Con contenido | 1188 |
| No leídas | 3 |
| **Total** | 3591 |

Debido a la gran cantidad de tablas escaneadas no fue posible en un principio identificar las variables de cada tabla, el tipo de datos de estas y la demás información relevante. Por consiguiente, se decidió, luego de haber escogido el CDM a utilizar, escanear y extraer la información de 3 tablas solamente. Estas tablas fueron seleccionadas luego de una asesoría con el personal de tecnologías TI del hospital, el cual nos indicó donde encontrar la información de mayor interés, que en este caso era información como sexo, fechas de nacimiento y fechas de ingresos y egresos que habían tenido los pacientes.

Al realizar este segundo escaneo, se pudo extraer información importante de las 3 tablas como el número de registros de cada tabla, el número de campos por tabla y la cantidad de campos que estaban completamente vacíos en cada tabla. En la TABLA II se puede ver esta información.

TABLA II
INFORMACIÓN PRINCIPAL DE LAS TABLAS ESCANEADAS

| Tabla | No. de registros | No. de campos | No. de campos vacíos |
|---|---|---|---|
| Tabla 1 | 384,944 | 148 | 63 |
| Tabla 2 | 323,436 | 47 | 3 |
| Tabla 3 | 1,435,852 | 99 | 16 |

Nota: Los nombres de las tablas se excluyen por motivos de confidencialidad.

Se logró identificar que la tabla 1 era la que contenía mayor cantidad de campos, sin embargo, también era la que presentaba mayor cantidad de campos vacíos. Por lo tanto, se procedió a revisar que significaban los campos que estaban vacíos, pero no se encontró que fuera información relevante para el proyecto. En esta tabla había información personal de los pacientes, como identificaciones, nombres, apellidos, direcciones, fechas de nacimiento, estrato socioeconómico, entre otros. Como se puede ver en la TABLA II, existían 384,944 registros en la tabla 1, lo cual teniendo en cuenta la información contenida en esta nos indica que hay 384,944 pacientes únicos registrados en la base de datos del hospital.

Por otra parte, al verificar los campos de las tablas 2 y 3 fue posible evidencia que contenían diversas fechas y horas, las cuales al averiguar con el personal TI del hospital indicaron que eran correspondientes a los ingresos y egresos de los pacientes. Además, estas tablas informaban que tipo de ingreso habían tenido los pacientes: hospitalario o ambulatorio. Se quiso saber desde que año había registros de pacientes, por lo cual se hizo una consulta en SQL Server donde se ordenaron los registros por la fecha de ingreso. Esto mostró que había 4 registros con fecha inferior al año 2007, por lo tanto, se tomaron estos registros como erróneos, ya que eran todos de distintos años y eran muy pocos. Se concluyó entonces que la base de datos del hospital tiene registros desde el año 2007.

Fue posible evidenciar, que ninguna de las tablas tenía igual número de registros. Por ejemplo, en la tabla 3 que contiene las fechas de ingresos, se observaron muchos más registros que en la tabla 1, esto debido a que un mismo paciente puede tener varios ingresos a lo largo de los años. Por otra parte, la tabla 2 que corresponde a las fechas de egresos, mostró una menor cantidad

de registros. Según indagaciones en la mayoría de los casos cuando es un ingreso ambulatorio no se registran los egresos, lo que explicaría la diferencia de registros. Sin embargo, al hacer una consulta desde SQL para verificar esto, se probó que sí había registros de egresos en casos de visitas ambulatorias, por lo cual no se pudo explicar si había un criterio específico para registrar los egresos. Aun así, se comprobó que cuando faltaba una fecha de egreso siempre se cumplía que era un ingreso ambulatorio.

### *B. Selección del modelo de datos común adecuado para implementar en el S.E.S Hospital Universitario de Caldas*

En la TABLA III, se muestran las iniciativas más conocidas y documentadas en el campo clínico de acuerdo con la revisión bibliográfica realizada, estas incluyen bases de datos, repositorios y redes de investigación. Se da una breve descripción de cada una y se expone el número de tablas que tiene su estructura, la cantidad de pacientes registrados en estas y la clasificación de si presentan o no un CDM.

Es posible observar en la TABLA III que las dos primeras iniciativas son las que reportan una menor cantidad de registros de pacientes (omitiendo a i2b2 del cual no se encontraron registros). Esto, debido a que son bases de datos clínicas que tomaron información de una fuente específica durante un tiempo y la guardaron en una estructura. Por tal motivo, tampoco tiene un CDM ya que su objetivo es simplemente tener a disposición estos datos para investigaciones clínicas. Por otra parte, las últimas 4 iniciativas son redes colaborativas de investigación en el campo clínico, que lo que buscan es permitir la construcción de múltiples bases de datos utilizando una estructura y vocabulario común. Presentan un CDM y un mayor número de registros (Sentinel, PCORnet, OHDSI), porque tienen datos de diversas fuentes, en el caso de OHDSI datos internacional mientras que Sentinel y PCORnet datos de Estados Unidos. Por otra parte, se puede apreciar que MIMIC es la que presenta mayor cantidad de tablas en su estructura, seguido de OMOP, eICU, PCORnet, Sentinel y por último i2b2. Teniendo en cuenta que MIMIC y eICU no proporcionan documentación para la transformación de los datos y no tiene un CDM, queda OMOP como estructura con mayor cantidad de tablas. Además, revisando a detalle las tablas y campos de

las últimas 4 iniciativas, se tiene como hipótesis que el CDM de OMOP es con el que se podría tener menor pérdida de información y mayor alcance de los datos.

TABLA III
INFORMACIÓN PRINCIPAL DE LAS BASES DE DATOS CLÍNICAS INVESTIGADAS

| Iniciativa | Descripción | No. de tablas | Registros | CDM |
|---|---|---|---|---|
| **MIMIC** [14] | Es una gran base de datos de libre acceso que comprende datos no identificados relacionados con la salud de pacientes que ingresaron en las unidades de cuidados intensivos del Centro Médico Beth Israel Deaconess. | 40 | Más de 60,000 pacientes | No |
| **eICU** [22] | Base de datos de investigación con datos de unidades de cuidados intensivos en los Estados Unidos. Surgido de eICU de Philips que es un programa transformador de telesalud de atención crítica que brinda información necesaria a los cuidadores. | 31 | 139,367 pacientes | No |
| **PCORnet** [23] | Red Nacional de Investigación Clínica Centrada en el Paciente, fue establecida por el Instituto de Investigación de Resultados Centrados en el Paciente (PCORI, por sus siglas en inglés) para realizar investigaciones clínicas de eficacia comparativa (CER) centradas en el paciente. | 23 | Más de 30 millones de personas | Sí |
| **i2b2** [24] | Es una fundación que permite la colaboración eficaz para la medicina de precisión, mediante el intercambio, la integración, la estandarización y el análisis de datos heterogéneos del cuidado de la salud y la investigación. | 6 | No se encontró una cifra | Sí |
| **OHDSI-OMOP** [25] | OHDSI es un programa internacional, que ha establecido una red internacional de investigadores y bases de datos observacionales de salud con un centro de coordinación central ubicado en la Universidad de Columbia. | 39 | 810 millones de pacientes | Sí |
| **Sentinel** [26] | Sistema electrónico nacional de la FDA que ha transformado la forma en que los investigadores controlan la seguridad de los productos médicos regulados por la FDA, incluidos medicamentos, vacunas, productos biológicos y dispositivos médicos | 19 | 365 millones de identificadores únicos de pacientes | Sí |

Nota: El CDM de la iniciativa OHDSI se nombra OMOP debido a que esta iniciativa surgió de una asociación público-privada establecida en los EE. UU (llamada OMOP) que usaba este CDM.

Se encontraron varios estudios que evaluaban o comparaban CDMs con distintos criterios y métodos. A continuación, en la TABLA IV se resumen cuatro de ellos, donde se muestra el título del artículo, el nombre y país de los autores, métodos de evaluación empleados, el CDM seleccionado y la razón por la que se seleccionó el CDM.

TABLA IV
RESUMEN ARTÍCULOS QUE EVALUAN CDMs

| | **Artículo 1** | **Artículo 2** | **Artículo 3** | **Artículo 4** |
|---|---|---|---|---|
| **Título** | "Evaluating common data models for use with a longitudinal community registry" [27] | "Identifying Appropriate Reference Data Models for Comparative Effectiveness Research (CER) Studies Based on Data from Clinical Information Systems" [28] | "Data Model Considerations for Clinical Effectiveness Researchers" [29] | "Common Data Models (CDMs) to Enhance International Big Data Analytics: A Diabetes Use Case to Compare Three CDMs" [30] |
| **Autores** | Maryam Garza, Guilherme Del Fiol, Jessica Tenenbaum, Anita Waldena, Meredith Nahm Zozus. | Omolola I. Ogunyemi, Daniella Meeker, Hyeon-Eui RN. Kim, Naveen Ashish, Seena Farzaneh, Aziz Boxwala. | Michael G. Kahn, Deborah Batson, Lisa M. Schilling | Harshana Liyanage, Siaw-Teng Liaw, Jitendra Jonnagaddala, William Hinton, Simon de Lusignan |
| **País** | Estados Unidos | Estados Unidos | Estados Unidos | Reino Unido |
| **Evaluación** | Se evaluaron 4 CDMs con 11 criterios que estaban divididos en 6 categorías: cobertura de contenido, integridad, flexibilidad, facilidad de consulta, compatibilidad de estándares y facilidad y alcance de implementación. | Se compararon 4 CDMs utilizando métricas cualitativas para la recuperación de datos y la pérdida de información en una variedad de áreas temáticas de la Investigación de efectividad comparativa (CER). | Se evaluaron 8 dimensiones de la calidad del modelo de datos para seleccionar un modelo de datos que fuera propicio para la expansión con nuevos requisitos de investigación. | Se usaron quince criterios predefinidos para comparar 3 CDMs, con un caso de uso de estudio de cohorte de diabetes, evaluando el beneficio (buen control de la diabetes), el riesgo (hipoglucemia) y la rentabilidad de los medicamentos recientemente autorizados. |
| **CDM seleccionado** | OMOP | OMOP | OMOP | Ninguno |
| **Razones selección** | Obtuvo el porcentaje más alto adaptación de los datos (76 %), es decir que fue con el que menos información se perdió. Además, cumplió bien con otros requisitos y fue el que tuvo una cobertura de terminología más amplia | Cumplió con el complemento más amplio de los objetivos de CER, con este se produjo una pérdida de información mínima al mapear los datos. | Permitió agregar una amplia gama de observaciones clínicas sin cambios estructurales en el modelo (sin tablas o columnas nuevas), manejar datos faltantes sin crear celdas vacías y tener pruebas de campo extensas con muy grandes conjuntos de datos administrativos y clínicos que respaldan métodos analíticos complejos | OMOP fue el CDM que más criterios predefinidos logró (14/15) contra los otros CDMs evaluados que lograron 13/15 y 10/15. Sin embargo, ninguno de los CDMs evaluados cumplió con los requerimientos de seguridad evaluados. |

Es posible apreciar en la TABLA IV que en 3 de los 4 los artículos revisados, se eligió el CDM de OMOP como mejor modelo de datos por distintas razones, pero principalmente porque la pérdida de información con esta estructura fue mínima, lo que confirmó la hipótesis que se había hecho anteriormente. En el caso del artículo 4 se puede apreciar que a pesar de no haber sido elegido el CDM de OMOP como el mejor, sí fue el que más criterios predefinidos logó cumplir.

La revisión de los aspectos finales arrojó que la iniciativa de OHDSI junto con Sentinel eran las que contaban con un mayor número de herramientas que tenían distintos propósitos como: facilitar el proceso ETL de su CDM, revisar la calidad de los datos, visualizar consultas, entre otros. La iniciativa de i2b2 ofrecía una plataforma que incluía diversas herramientas para la exploración de datos, visualización, análisis genómicos y ETL. Para PCORnet no se encontraron herramientas de apoyo para procesos ETL o análisis de datos.

Además, se evidenció que OHDSI es el proyecto que cuenta con mayor documentación y grupos de apoyo. La iniciativa tiene un libro donde se explican diversos aspectos como el proceso de ETL, vocabularios del CDM, hasta estudios de cohortes, entre otros; además, proporciona una plataforma donde se ofrecen diversos cursos grabados e invitan a un foro donde se pueden hacer preguntas de diversos temas diariamente. Por último, posee grupos de trabajo donde se reúnen mensualmente para socializar diferentes proyectos, ya sea de ETL, herramientas nuevas, estudios con los datos del CDM, entre otros.

De la clasificación realizada a los 100 artículos, se encontró que OMOP es el CDM con mayor cantidad de citaciones, demostrando que es el CDM que ha tenido una mayor acogida en la comunidad científica y para investigaciones de ciencia de datos, especialmente para la realización de estudios predictivos y de cohortes observacionales (incluso en su modelo hay una tabla específica para formar cohortes de datos). En la Fig. 4 se muestra el porcentaje de artículos que usan o mencionan cada CDM.

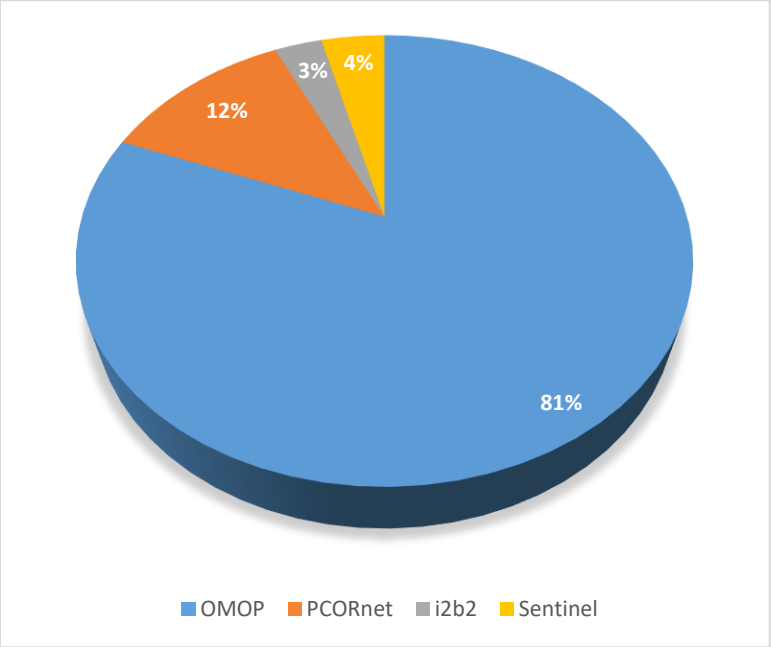



Fig. 4. Porcentaje de artículos que usan o mencionan los cuatro CDMs consultados.

Luego de la reunión con el equipo de TI y después de presentar los resultados encontrado se eligió el CDM de OMOP para implementar con los datos del hospital, debido a los amplios beneficios que este presentaba y que fueron mencionados anteriormente.

A continuación, en la Fig. 5 se muestra el esquema general de la versión actual (v5.4) del CDM de OMOP. Allí se pueden apreciar las 39 tablas que tiene, la sección a la que pertenece cada tabla y algunas conexiones entre tablas. Es importante resaltar, que este es un modelo centrado en la persona, por ello todas las tablas van a tener como centro y se van a conectar a la tabla PERSON.

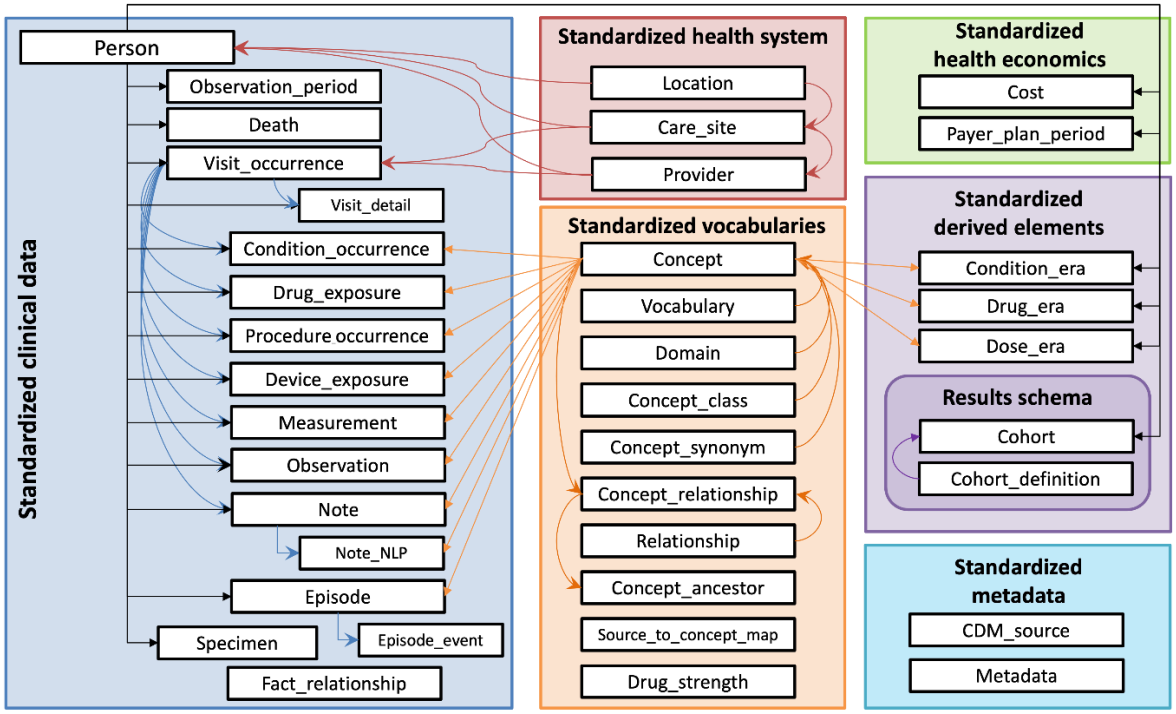



Fig. 5. Esquema de La versión actual (v5.4) de CDM de OMOP

Nota: fuente https://ohdsi.github.io/CommonDataModel/index.html

### *C. Extracción, transformación y carga de datos (ETL)*

En la Fig. 6 se puede observar el diseño del ETL realizado con ‘Rabbit in a hat’ para las 3 primeras tablas del CDM. Con esta herramienta, se hicieron las relaciones entre las tablas de origen y las del CDM. Es posible observar que la herramienta es muy amigable para realizar estos diseños, ya que al importar el reporte generado por ‘White Rabbit’ se muestra información de cada una de las tablas cuando se selecciona, como los campos y el tipo de datos de cada campo. Además, para las tablas del CDM se muestra adicionalmente una descripción de cada campo como se puede ver en la Fig. 6 donde se seleccionó la tabla PERSON.

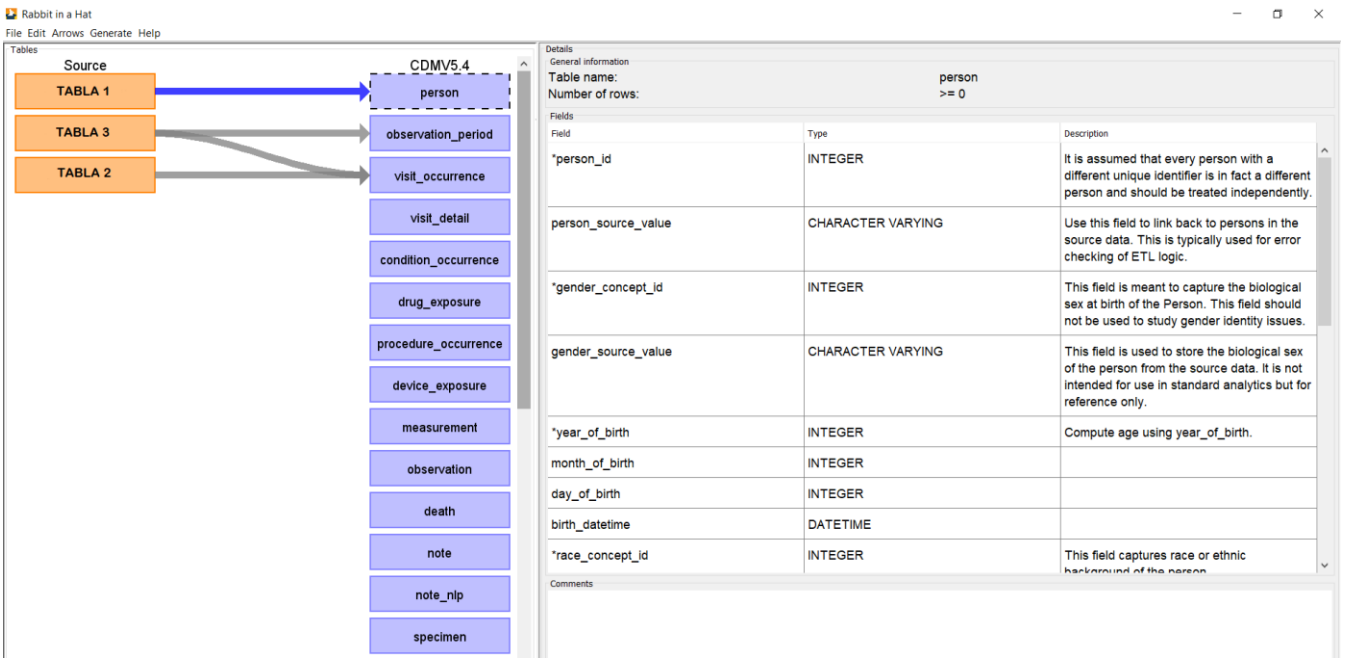



Fig. 6. Diseño del ETL con la herramienta ‘Rabbit in a hat’.

Nota: Los nombres originales de las tablas de origen se excluyen por motivos de confidencialidad.

Gracias a estas herramientas, fue posible identificar y relacionar que la tabla 1 contenía la información necesaria para insertar en la tabla PERSON del CDM, la tabla 3 contenía la información para la tabla OBSERVATION_PERIOD y para la tabla VISIT_OCURRENCE se evidenció que era preciso utilizar información contenida tanto de la tabla 2 como de la tabla 3.

Además, se realizó la conexión entre cada uno de los campos dentro de las tablas. En la Fig. 7 se muestra el proceso para la tabla 1 y la tabla PERSON. En esta primera tabla era necesario insertar datos de la persona como el género, el año de nacimiento y la raza. Es posible observar que, en este caso específico, la herramienta mostraba información o no dependiendo del campo que se seleccionaba. En este caso, que se seleccionó el campo género, se puede apreciar una lista de códigos y conceptos. Estos códigos sirvieron de guía para la transformación de los datos, ya que el CDM se basa en estos códigos para unificar la información por lo tanto fue necesario tenerlos claros para la lógica del script. Es posible ver que la tabla del CDM contiene varios campos, pero no todos son obligatorios (el * indica que no pueden ser valores nulos), es por esto que solo se hicieron las conexiones para los campos mostrados. En el caso de *race_concept_id* y *ethnicity_concept_id*,

se muestra que no pueden ser valores nulos, sin embargo, estos conceptos solo aplican para Estados Unidos, ya que en Colombia no se tienen esas clasificaciones, por lo tanto, no se realizó ninguna conexión.

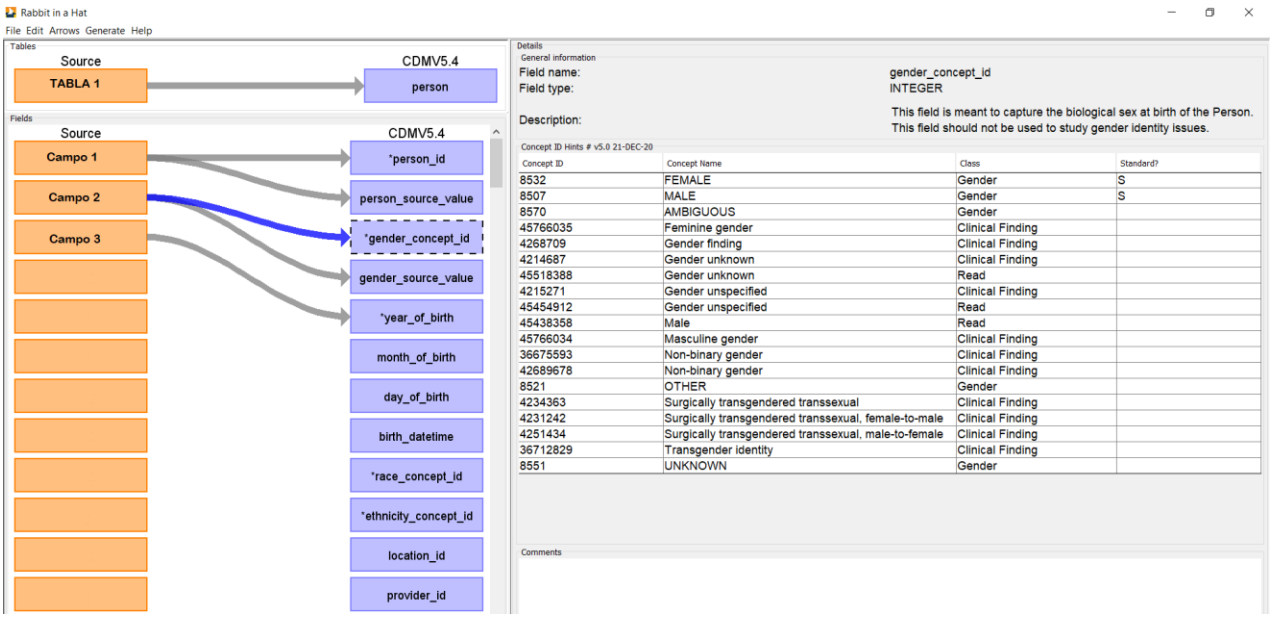



Fig. 7. Relación entre campos de la TABLA 1 y de la tabla PERSON.

Nota: Los nombres de los campos en la tabla de origen se excluyen por motivos de confidencialidad.

El informe que se generó con ‘Rabbit in a hat’ sirvió de guía para realizar los scripts y no tener que volver a la herramienta, ya que en este se exportaron todas las conexiones realizadas y la lógica que se debería seguir. En la Fig. 8 se muestra la lógica de la tabla PERSON que se exportó al informe y con la cual se realizó el script. Se puede ver por ejemplo que la lógica para el campo de género es cambiar los valores de origen (1,2,3) por los códigos correspondientes (8507,8532,8570). Para las demás tablas se realizó el mismo procedimiento y también se vieron reflejadas las conexiones y lógicas en el informe.

| Destination Field | Source Field | Logic | Comment |
|---|---|---|---|
| person_id | Campo 1 | Num. Aleatorio = person_id | |
| person_source_value | Campo 1 | Igual a campo 1 | |
| gender_concept_id | Campo 2 | 1 = Masculino = 8507<br>2 = Femenino = 8532<br>3 = 8570 | |
| gender_source_value | Campo 2 | Igual a campo 2 | |
| year_of_birth | Campo 3 | Año del Campo 3 | |

Fig. 8. Lógica propuesta para insertar los datos en la tabla PERSON.

Nota: Los nombres de los campos de origen se excluyen por motivos de confidencialidad.

En la Fig. 9 se muestra el script implementado para insertar los datos en la tabla PERSON, es posible observar que hay algunos valores que se toman como nulos, esto debido a que no eran campos obligatorios y en algunos casos no aplicaban para los datos (como el caso de raza). Para la fecha de nacimiento se decidió insertar solo el año de nacimiento de los pacientes, esto debido a que los campos de día y año no eran obligatorios, además para evitar problemas de re-identificación de los datos. Se decidió que cuando algún paciente no tuviera información sobre el género no se incluía, esto debido a que en la mayoría de los estudios como indica OHDSI es un valor obligatorio.

```
truncate table [redacted].person;

insert into [redacted].person (
            person_id,
            gender_concept_id,
            year_of_birth,
            month_of_birth,
            day_of_birth,
            birth_datetime,
            race_concept_id,
            ethnicity_concept_id,
            location_id,
            provider_id,
            care_site_id,
            person_source_value,
            gender_source_value,
            gender_source_concept_id,
            race_source_value,
            race_source_concept_id,
            ethnicity_source_value,
            ethnicity_source_concept_id
)
select
    row_number()over(order by p.[redacted]) as person_id,
    case (p.[redacted])
        when [redacted] then 8507
        when [redacted] then 8532
        when [redacted] then 8570
    end as gender_concept_id,
    datepart(year, p.[redacted]) as year_of_birth,
    NULL as month_of_birth,
    NULL as day_of_birth,
    NULL as birth_datetime,
    0 as race_concept_id,
    0 as ethnicity_concept_id,
    NULL as location_id,
    NULL as provider_id,
    NULL as care_site_id,
    p.[redacted] as person_source_value,
    p.[redacted] as gender_source_value,
    NULL as gender_source_concept_id,
    NULL as race_source_value,
    NULL as race_source_concept_id,
    NULL as ethnicity_source_value,
    NULL as ethnicity_source_concept_id
from [redacted] p
where p.[redacted] is not null;
```

Fig. 9. Script de SQL server realizado para insertar datos en la tabla PERSON.

Nota: Alguna información no se muestra por motivos de confidencialidad.

En la Fig. 10 se muestra el script realizado para insertar los datos en la tabla OBSERVATION_PERIOD. Está tabla contiene la información sobre el intervalo de tiempo que una persona ha tenido eventos clínicos en el hospital. Según las indicaciones de OHDSI esta información puede extraerse de diversas fuentes como los datos administrativos en donde se

registran periodos de inscripción o en otros casos es necesario hacer inferencias con la información de los registros EHR. Este último fue el caso del presente proyecto, ya que al ser un hospital que recibe pacientes de diversas Entidades Promotoras de Salud (EPS) no se tiene información sobre los periodos de inscripción a estas. Por lo tanto, se insertó como fecha inicial del periodo de observación (*observation_period_start_date*) el primer ingreso que había registrado de cada paciente en la base de datos del hospital y como fecha final del periodo (*observation_period_end_date*) la última fecha de ingreso. Por otra parte, para el campo *period_type_conept_id* se definió el código 32817, que según el vocabulario de OMOP indica EHR y define la procedencia de los datos en esta tabla.

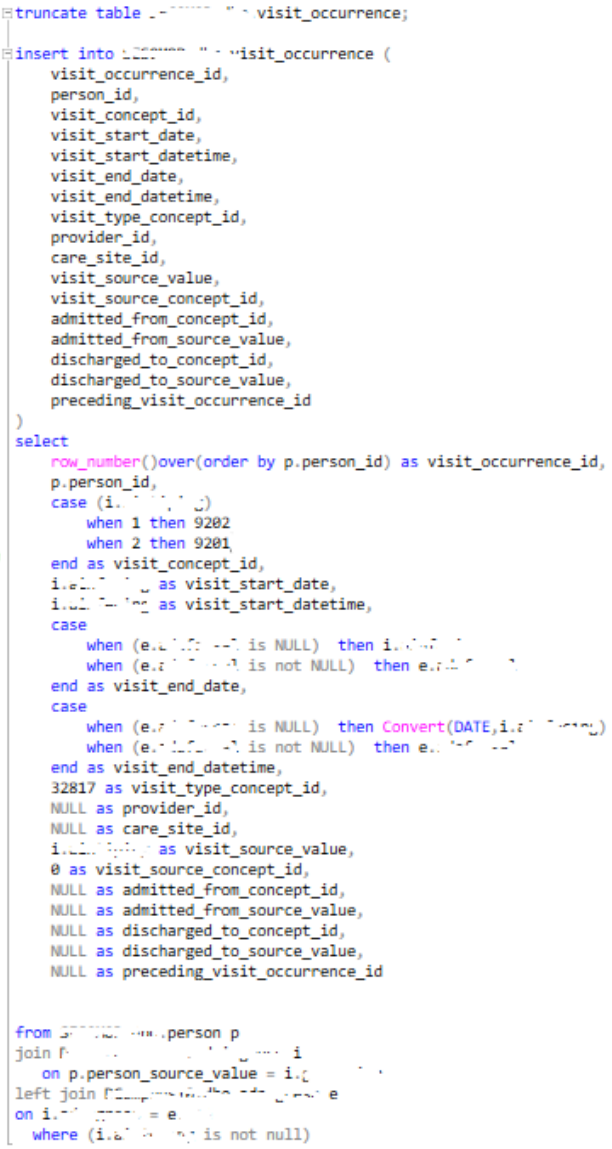
```
truncate table [redacted].observation_period;

insert into [redacted].observation_period (
observation_period_id,
person_id,
observation_period_start_date,
observation_period_end_date,
period_type_concept_id
)
select

        row_number()over(order by person_id) as observation_period_id,
        p.person_id,
        min([redacted]) as observation_period_start_date,
        max([redacted]) as observation_period_end_date,
        32817 as period_type_concept_id
from [redacted].person p
join [redacted] e
   on p.person_source_value = e.[redacted]
   where e.[redacted] is not null
group by p.person_id;
```

Fig. 10. Script de SQL server realizado para insertar datos en la tabla OBSERVATION_PERIOD

Nota: Alguna información no se muestra por motivos de confidencialidad.

El script realizado para para insertar los datos en la tercera tabla: VISIT_OCCURRENCE, se muestra en la Fig. 11 . En esta tabla se registra cada visita que tiene un paciente al hospital, la fecha de inicio de la visita, la fecha final y el tipo de visita: hospitalario o ambulatorio. Por esto, fue necesario utilizar las tablas 2 y 3 de la base de datos del hospital para insertar los datos de esta tabla. Se tomó como fecha de inicio de la visita, la fecha de ingreso de cada paciente que se encontraba en la tabla 3 y como fecha final de la visita, la fecha de egreso de cada paciente

registrada en la tabla 2. Sin embargo, como se vio en la exploración de los datos en muchos casos no había fecha de egreso, por lo tanto, se asumió para la fecha de egreso la misma del ingreso y para la hora la medianoche (00:00:0000). Esto se realizó teniendo en cuenta que en todos los casos que faltaba la información de egreso eran visitas ambulatorias. Por otra parte, se cambiaron los códigos 1 y 2 de la fuente de origen por los códigos 9202 y 9201, que indicaban si el paciente había tenido un ingreso hospitalario o ambulatorio.



```
truncate table [redacted] visit_occurrence;

insert into [redacted] visit_occurrence (
    visit_occurrence_id,
    person_id,
    visit_concept_id,
    visit_start_date,
    visit_start_datetime,
    visit_end_date,
    visit_end_datetime,
    visit_type_concept_id,
    provider_id,
    care_site_id,
    visit_source_value,
    visit_source_concept_id,
    admitted_from_concept_id,
    admitted_from_source_value,
    discharged_to_concept_id,
    discharged_to_source_value,
    preceding_visit_occurrence_id
)
select
    row_number()over(order by p.person_id) as visit_occurrence_id,
    p.person_id,
    case (i.[redacted])
        when 1 then 9202
        when 2 then 9201
    end as visit_concept_id,
    i.[redacted] as visit_start_date,
    i.[redacted] as visit_start_datetime,
    case
        when (e.[redacted] is NULL)  then i.[redacted]
        when (e.[redacted] is not NULL)  then e.[redacted]
    end as visit_end_date,
    case
        when (e.[redacted] is NULL)  then Convert(DATE,i.[redacted])
        when (e.[redacted] is not NULL)  then e.[redacted]
    end as visit_end_datetime,
    32817 as visit_type_concept_id,
    NULL as provider_id,
    NULL as care_site_id,
    i.[redacted] as visit_source_value,
    0 as visit_source_concept_id,
    NULL as admitted_from_concept_id,
    NULL as admitted_from_source_value,
    NULL as discharged_to_concept_id,
    NULL as discharged_to_source_value,
    NULL as preceding_visit_occurrence_id


from [redacted].person p
join [redacted] i
   on p.person_source_value = i.[redacted]
left join [redacted] e
on i.[redacted] = e.[redacted]
  where (i.[redacted] is not null)
```

Fig. 11. Script de SQL server realizado para insertar datos en la tabla VISIT_OCCURRENCE

Nota: Alguna información no se muestra por motivos de confidencialidad

A continuación, en la TABLA V se muestran los códigos utilizados para la transformación de los datos, con su significado correspondiente y su equivalente en la base de datos de origen.

TABLA V
CÓDIGOS DEL CDM DE OMOP UTILIZADOS

| Código CDM de OMOP | Significado | Correspondencia |
|---|---|---|
| 8507 | Género masculino | 1 |
| 8532 | Género femenino | 2 |
| 8570 | Género ambiguo | 3 |
| 32817 | EHR | No aplica |
| 9202 | Paciente ambulatorio | 1 |
| 9201 | Paciente hospitalario | 2 |

En la TABLA VI se muestra la cantidad de registros que fueron transformados y que se insertaron finalmente en el CDM de OMOP. Es posible observar que para la tabla PERSON el 100% de los registros fueron transformados, ya que como se mostró anteriormente había un total de 384,944 pacientes únicos en la base de datos de origen (tabla 1) cantidad que coincide con la de la tabla PERSON. Por otra parte, la tabla OBSERVATION_PERIOD debería tener el mismo número de registros que la taba PERSON porque se debe registrar un periodo de observación para cada paciente único que halla. Sin embargo, como se observa, en la tabla del CDM se reportan 19,365 registros menos. Para verificar si había algún error, se realizó una consulta del total de ingresos que poseía cada persona en la base de datos del hospital. La consulta arrojó que 19,365 pacientes no tenían ningún ingreso al hospital, es decir, no había un periodo de observación para ellos y por lo tanto la cantidad de registros transformados en el CDM fue correcta. En la tabla VISIT_OCCURRENCE se puede ver que el número de registros es igual al de la Tabla 3 (DB origen), lo cual es coherente ya que en esa tabla se encontraban registrados todos los ingresos de los pacientes.

TABLA VI
CANTIDAD DE REGISTROS FINALES

| Tabla | No. de registros en el CDM |
|---|---|
| PERSON | 384,944 |
| OBSERVATION_PERIOD | 365,579 |
| VISIT_OCCURRENCE | 1,435,852 |

Las consultas realizadas en las tablas transformadas al CDM de OMOP para las 5 personas escogidas aleatoriamente arrojaron la información mostrada en la TABLA VII. Al comparar estos datos con los resultados de las consultas para la base de datos del hospital, se ratificó que era la misma información y todos los datos y fechas coincidía, confirmando la correcta transformación de los datos.

TABLA VII
INFORMACIÓN OBTENIDA DE LAS CONSULTAS REALIZADAS PARA VERIFICACIÓN

| | **Paciente 1** | **Paciente 2** | **Paciente 3** | **Paciente 4** | **Paciente 5** |
|---|---|---|---|---|---|
| **Visitas** | 10: 9 ambulatorias y 1 hospitalaria | 4: 3 ambulatorias y 1 hospitalaria | 8: 5 ambulatorias y 3 hospitalarias | 2 ambulatorias | 3: 2 ambulatorias y 1 hospitalaria |
| **Fechas visitas** | 2007 a 2022 | 2008 a 2020 | 2008 a 2020 | 2007 a 2017 | 2008 a 2011 |
| **Género** | Masculino | Femenino | Masculino | Femenino | Femenino |
| **Año de nacimiento** | 1965 | 1998 | 1944 | 2007 | 1992 |

*D. Proceso de anonimización*

Después de revisar la información relacionada a la seguridad y privacidad de los datos de cada una de las iniciativas, se encontró que:

- En las dos bases de datos públicas (MIMIC [14] y eICU [15]) los datos y tablas están desidentificados de acuerdo con los estándares de la Ley de Portabilidad y Responsabilidad de Seguros Médicos (HIPAA, por sus siglas en inglés) [31]. Para

esto, los proyectos eliminaron toda la información de salud protegida (PHI, por sus siglas en inglés) de los datos. Además, asignaron a los pacientes identificadores aleatorios y eliminaron las claves que los vinculaban a los datos originales. También se descartaron todos los identificadores de hospitales y UCI para proteger la privacidad de las instituciones y proveedores contribuyentes. Para el caso de los campos de texto libre, buscaron y eliminaron la información personal utilizando un enfoque basado en reglas (por ejemplo, buscaban las palabras que seguían a “Mr.” O “Ms.” ya que normalmente eran nombres, como "Mr. Smith") [15]. Por último, miembros del personal experto hicieron una revisión manual para verificar que todos los datos habían sido desidentificados [14], [15].

- PCORnet [23] , al ser una red de investigación que incluye a varios socios, no acumula datos en un solo grupo o almacén de datos. Para ello, tienen una infraestructura que permite a cada socio permanecer con sus datos a través de su firewall protegido por HIPAA. Los socios pueden realizar consultas a través de un portal destinado para ello, el cual incluye un sistema de gobernanza sólida, controles de acceso basados en roles y auditoría. Los investigadores envían sus consultas y las respuestas, no los datos, se devuelven a los investigadores. En el caso de necesitar acceso a los datos, se puede solicitar a través de su plataforma web “Front Door” y se otorga con una política caso a caso dependiendo del proyecto [32].
- Sentinel [26] funciona de forma similar a PCORnet, donde los socios de datos conservan la administración y la posesión tanto de los datos de origen como de los datos transformados al CDM. Cuando se realizan consultas específicas, los socios de datos no comparten identificadores directos de pacientes y se adhieren al estándar HIPAA. Además, el sistema Sentinel se adhiere a la Ley Federal de Gestión de la Seguridad de la Información de 2002 (FISMA) [33].
- Los datos de i2b2 [24] se encuentran des-identificados de acuerdo a HIPAA y para tener acceso a ellos se debe firmar un acuerdo de confidencialidad donde se informe el propósito de los datos [34].
- La red de OHDSI [25] no comparte datos de forma pública en sus plataformas. Sin embargo, invita a los colaboradores que cuando hayan completado la conversión de los datos, se informe al administrador del programa OHDSI para tener el censo de

la red de datos que mantiene. Cuando se realizan investigaciones en la red, no se agrupan datos a nivel de paciente en los sitios de la red, solo se comparten los resultados. Aunque no se publiquen datos en la red, OHDSI brinda una orientación sobre el proceso general y los campos potenciales que deben monitorearse para cumplir con los protocolos de preservación de la privacidad, debido a que el CDM de OMOP al estar centrado en la persona puede retener atributos que logran considerarse información personal identificada (PII) o información de salud protegida (PHI) [35], [36].

Se puede observar que todas las iniciativas siguen la Ley de Portabilidad y Responsabilidad de Seguros Médicos (HIPAA) [31], que es la regla de privacidad que establece las políticas para proteger toda la información de salud identificable individualmente que se retiene o transmite en Estados Unidos. Según esta normativa existen 18 identificadores que se consideran información de identificación personal y los cuales deben ser tratados al usar y compartir datos de salud. A continuación se mencionan los identificadores [31]:

- Nombre
- Dirección (todas las subdivisiones geográficas más pequeñas que el estado, incluida la dirección, la ciudad, el condado y el código postal)
- Todos los elementos (excepto años) de fechas relacionadas con una persona (incluida la fecha de nacimiento, la fecha de admisión, la fecha de alta, la fecha de fallecimiento y la edad exacta si tiene más de 89 años)
- Números telefónicos
- Número de fax
- Dirección de correo electrónico
- Número de seguro social
- Numero de historia clínica
- Número de beneficiario del plan de salud
- Número de cuenta
- Número de certificado o licencia
- Identificadores de vehículos y números de serie, incluidos los números de matrícula

- Identificadores de dispositivos y números de serie
- URL web
- Dirección de Protocolo de Internet (IP, por sus siglas en inglés)
- Huella dactilar o de voz
- Imagen fotográfica: las imágenes fotográficas no se limitan a las imágenes de la cara.
- Cualquier otra característica que pueda identificar de manera única al individuo

De acuerdo a la información buscada, se encontró que en Colombia, la Ley General de protección de datos personales- Ley 1581 de 2012 [37], es la que establece los principios aplicables a las actividades de tratamiento de datos personales para garantizar el derecho fundamental de Habeas Data de las personas. En el artículo 5 de esta ley [37] se define datos sensibles como: "aquellos que afectan la intimidad del titular o cuyo uso indebido puede generar su discriminación" y dentro de los cuales se incluyen los datos relativos a la salud. Por otra parte, en el artículo 6 numeral e [37] se establece que estos datos pueden ser tratados siempre y cuando: "El Tratamiento tenga una finalidad histórica, estadística o científica", en cuyo caso " deberán adoptarse las medidas conducentes a la supresión de identidad de los Titulares."

Se hallaron algunas guías de anonimización realizadas por diferentes organizaciones y ministerios del territorio colombiano. Dentro de ellos está uno de gran importancia para el presente proyecto, elaborado por el Archivo General de la Nación en articulación con el Ministerio de Tecnologías de la Información y las Comunicaciones, la Superintendencia de Industria y Comercio, el Departamento Administrativo de la Función Pública, el Departamento Nacional de Planeación y el Departamento Administrativo Nacional de Estadística [38]. En la guía mencionada, se brindan elementos metodológicos y técnicos para que las entidades garanticen la protección de cualquier información producida, gestionada o recolectada que contenga datos personales. Todo esto, con el objetivo de masificar la disponibilidad de datos públicos digitales accesibles, usables y de calidad, pero salvaguardando los principios reconocidos por la legislación en materia de transparencia y protección de datos personales.

Teniendo en cuenta la información encontrada sobres los 18 identificadores de HIPAA [31], las recomendaciones dadas por OHDSI [36] y la guía de anonimización colombiana elaborada por Archivo general de Nación [38], se tomaron una serie de precauciones a la hora de transformar las tablas en el CDM de OMOP que consistieron en:

1. Evitar que los identificadores fueran iguales en la tabla de origen (tabla 1) y en la tabla PERSON, para lo cual se insertaron números aleatorios como identificadores en la tabla del CDM.
2. No incluir el día y el mes de nacimiento en la tabla PERSON, es decir, solo se insertó el año de nacimiento.
3. Eliminar las direcciones y códigos postales, por tal razón no se insertó ninguna información en las tablas LOCATION, CARE_SITE y PROVIDER.

Considerando que para la transformación de las demás tablas del CDM es necesario relacionar a las personas con cada uno de los eventos, alguna información que podría considerarse como PII o PHI, tuvo que conservarse. Sin embargo, cuando se termine la transformación de todas las tablas para poder compartir los datos, esta información debe anonimizarse correctamente para evitar riesgos de re-identificación. A continuación, se describe el protocolo que debe seguirse para esto:

1. Debe realizarse un análisis de riesgos de identificación para los datos del CDM. Para esto, deben clasificarse las variables por su nivel de sensibilidad, plantear los riesgos que existen en esta nueva base de datos, identificar los campos riesgosos y crear un informe de riegos. Se recomienda revisar la siguiente información:

    - Atributos de la tabla PERSON que incluyan identificadores y fechas de nacimientos (día y mes). Tener en cuenta que para personas mayores a 89 años debe excluirse el año de nacimiento también, según HIPAA.
    - Campos de fecha en todos los dominios.
    - Atributos de la tabla LOCATION.
    - Atributos de la tabla PROVIDER.

- Atributos de la tabla OBSERVATION.
- Todos los campos *_source_value.
- Atributos de las tablas NOTE y NOTE_NLP.

2. Se deben identificar y seleccionar las técnicas de anonimización a utilizar. Dentro de las que pueden estar:

   - Técnicas de aleatorización como adición de ruido y permutación.
   - Técnicas de generalización como agregación y anonimato o diversidad y proximidad.

3. Se deben implementar las técnicas de anonimización asociadas a los riesgos de identificación seleccionadas en el paso anterior. Para la aplicación de las técnicas se pueden utilizar algunos softwares como μ-Argus [39], τ-Argus [40], entre otros. Los cuales permitan implementar diferentes técnicas de anonimización. Es importante que se creen rutinas (algoritmos) que ayuden a implementar las técnicas, para que el riesgo de identificación de las unidades de observación disminuya.

4. Por último, es necesario realizar una evaluación de resultados del proceso. Para esto, se debe validar que los riesgos de identificación de las unidades de observación se hayan minimizado y que las variables de la base de datos conserven las propiedades estadísticas deseadas.

D. *Elaboración guía de usuario*

En el anexo 1 se puede encontrar la guía de usuarios donde se familiariza al investigador con el CDM de OMOP. En esta, se detalla todo el proceso ETL para las personas encargadas de la transformación de los datos, se especifican las páginas web donde encontrar la información necesaria, los programas a descargar y cómo diseñar e implementar el ETL.

# VI. CONCLUSIONES

El proyecto pretendía estructurar la información clínica electrónica almacenada en diferentes sistemas de información del SES Hospital Universitario de Caldas, para obtener bases de datos seguras y confiables que facilitaran el desarrollo de aplicaciones en Ciencia de Datos. Para lograrlo, se construyó una metodología que abarcó, desde el análisis exploratorio de los datos involucrados, hasta el diseño de una guía de usuarios para el personal del Hospital, que sirvieron de base para definir los primeros pasos de esta estructuración.

El análisis exploratorio de los datos (EDA) elaborado a las historias clínicas del S.E.S. Hospital Universitario de Caldas, confirmó la necesidad de realizar una estructuración y armonización de la información para poder ser utilizada en investigaciones de ciencia de datos. Esto, debido a que la base de datos actual no es entendible para personas que no tengan experiencia con ella y se debe depender completamente del personal TI del hospital para utilizar los datos. Adicionalmente, hay múltiples tablas y campos vacíos y no hay una documentación adecuada que sirva como guía para los investigadores. El tener una estructura definida, donde las tablas y campos sean conocidos, la existencia de una documentación adecuada y que se tenga clara la distribución y relación de las tablas es de gran importancia para las futuras investigaciones con estos datos.

Existen múltiples CDMs que son utilizados para la armonización y estructuración de datos para investigaciones clínicas, sin embargo, el CDM de OMOP fue el que presentó mejores características para ser implementado en las historias clínicas del S.E.S Hospital Universitario de Caldas. El modelo de OMOP, es el que cuenta con mayor documentación para lograr una transformación adecuada a su CDM, proporciona variados recursos como foros de ayuda donde participan personas de todo el mundo, grupos de trabajo en diversos temas y plataformas con cursos sobre su modelo. La iniciativa OHDSI también facilita herramientas para el diseño e implementación del proceso ETL. Además, investigaciones anteriores [27]–[30], calificaron el CDM de OMOP como el mejor modelo según distintos criterios de evaluación, donde se destaca por ser el que menor pérdida de información genera y el que posee la cobertura más amplia de terminologías estándar.

Las herramientas proporcionadas por OHDSI fueron de gran utilidad a la hora de realizar el diseño y la implementación del ETL. Estas, facilitaron la transformación de los datos y gracias a esto se logró realizar correctamente la conversión de información relevante como sexos, años de nacimientos, periodos de observación y fechas y tipos de ingresos de cada una de las personas registradas en la base de datos del hospital; lo cual corresponde a las tres primeras tablas de CDM de OMOP.

Las leyes consultadas tanto nacionales como internacionales confirmaron la importancia de diseñar un protocolo de anonimización de datos adecuado para aplicaciones en ciencia de datos en el S.E.S Hospital Universitario de Caldas, el cual busque la protección de los datos personales y se aplique una vez la transformación de los datos se haya completado y se desee compartir los datos. Además, fue de gran utilidad tener claridad sobre guías de anonimización existentes realizadas por establecimientos públicos del gobierno nacional colombiano, así como procesos adoptados para garantizar la seguridad de la información en bases de datos clínicas existentes, y verificar las recomendaciones dadas por OHDSI para desarrollar correctamente el protocolo de anonimización.

Al final de todo este desarrollo, la elaboración de una guía de usuarios resulta de gran utilidad para los futuros investigadores que deseen seguir con el proceso de transformación de los datos en el S.E.S Hospital Universitario de Caldas, ya que, en esta se familiariza al personal sobre la nueva estructura de los datos y se explica como diseñar e implementar el ETL para la conversión de los datos al CDM de OMOP.

# VIII. RECOMENDACIONES

Se recomienda continuar y terminar la transformación de los datos del S.E.S. Hospital Universitario de Caldas, para lograr una base de datos confiable y completa, como una herramienta en futuras investigaciones en ciencias de datos. Se requiere, además, validar está información con un estudio específico en ciencia de datos, para medir su comportamiento en las investigaciones.

Se sugiere implementar este resultado en otros centros hospitalarios nacionales y establecer relaciones con la red de OHDSI, para ampliar el alcance con los datos y utilizarlos en investigaciones de gran impacto.

# REFERENCIAS


[1] M. Jin, S. Fan, K. Zhang, and X. Bao, "A Scoping Review of Clinical Unstructured Text Information Extraction," *Proc. - 2020 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2020*, pp. 853–857, Dec. 2020.

[2] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *J. Am. Med. Inform. Assoc.*, vol. 25, no. 10, p. 1419, Oct. 2018.

[3] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis."

[4] S. Abhyankar, D. Demner-Fushman, and C. J. McDonald, "Standardizing clinical laboratory data for secondary use," *J. Biomed. Inform.*, vol. 45, no. 4, pp. 642–650, Aug. 2012.

[5] H. Dalianis, A. Henriksson, M. Kvist, S. Velupillai, and R. Weegar, "HEALTH BANK-A Workbench for Data Science Applications in Healthcare," 2015.

[6] J. Machado De Oliveira, C. André Da Costa, and R. S. Antunes, "Data structuring of electronic health records: a systematic review," *Health Technol. (Berl).*, vol. 11, pp. 1219–1235, 2021.

[7] P. P. Maslianko and Y. P. Sielskyi, "Data Science — definition and structural representation," *Syst. Res. Inf. Technol.*, vol. 2021, no. 1, pp. 61–78, Jul. 2021.

[8] Amazon Web Services, "¿Qué es la ciencia de datos?" [Online]. Available: https://aws.amazon.com/es/what-is/data-science/. [Accessed: 22-Jan-2023].

[9] A.-M.-S. Louis, R. F. Sarmiento, L. Agha-Mir-Salim, and R. F. Sarmiento, "Health Information Technology as Premise for Data Science in Global Health: A Discussion of Opportunities and Challenges," *Leveraging Data Sci. Glob. Heal.*, pp. 3–15, 2020.

[10] S. V. G. Subrahmanya *et al.*, "The role of data science in healthcare advancements: applications, benefits, and future prospects," *Ir. J. Med. Sci.*, vol. 191, no. 4, pp. 1473–1483, Aug. 2022.

[11] Minsalud, "Interoperabilidad de Datos de la Historia Clínica en Colombia - Términos y siglas ." .

[12] L. Bulgarelli, A. Núñez-Reiz, and R. O. Deliberato, "Building Electronic Health Record Databases for Research," in *Leveraging Data Science for Global Health*, Springer, Cham,

2020, pp. 55–64.

[13] J. Marshall, A. Chahin, and B. Rush, “Review of clinical databases,” in *Secondary Analysis of Electronic Health Records*, Springer International Publishing, 2016, pp. 9–16.

[14] A. E. W. Johnson *et al.*, “MIMIC-III, a freely accessible critical care database,” *Sci. Data 2016 31*, vol. 3, no. 1, pp. 1–9, May 2016.

[15] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, “The eICU Collaborative Research Database, a freely available multi-center database for critical care research,” *Sci. Data 2018 51*, vol. 5, no. 1, pp. 1–13, Sep. 2018.

[16] A. Han, A. Isaacson, and P. Muennig, “The promise of big data for precision population health management in the US,” *Public Health*, vol. 185, pp. 110–116, Aug. 2020.

[17] M. G. Kahn, D. Batson, and L. M. Schilling, “Data Model Considerations for Clinical Effectiveness Researchers,” 2012.

[18] OHDSI, “OMOP Common Data Model .” [Online]. Available: https://www.ohdsi.org/data-standardization/the-common-data-model/. [Accessed: 19-Aug-2022].

[19] TranSMART Foundation, “i2b2: Informatics for Integrating Biology & the Bedside.” [Online]. Available: https://www.i2b2.org/. [Accessed: 19-Aug-2022].

[20] L. Wyatt, B. Caufield, and D. Pol, “Principles for an ETL benchmark,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5895 LNCS, pp. 183–198, 2009.

[21] A. Albrecht and F. Naumann, “Managing ETL Processes,” *VLDB*, vol. 08, 2008.

[22] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, “The eICU Collaborative Research Database, a freely available multi-center database for critical care research,” *Sci. data*, vol. 5, Sep. 2018.

[23] J. W. Timbie *et al.*, “National Patient-Centered Clinical Research Network (PCORnet) Phase I: Final Evaluation Report,” *Natl. Patient-Centered Clin. Res. Netw. Phase I Final Eval. Rep.*, Dec. 2015.

[24] J. Klann, “2. Quick Start Guide - Bundles and CDM .” [Online]. Available: https://community.i2b2.org/wiki/display/BUN/2.+Quick+Start+Guide. [Accessed: 20-Oct-2022].

[25] G. Hripcsak *et al.*, “Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers,” *Stud. Health Technol. Inform.*, vol. 216, p.

574, 2015.

[26] FDA, “Sentinel Initiative.” [Online]. Available: https://www.sentinelinitiative.org/. [Accessed: 19-Jan-2023].

[27] M. Garza, G. Del Fiol, J. Tenenbaum, A. Walden, and M. N. Zozus, “Evaluating common data models for use with a longitudinal community registry,” *J. Biomed. Inform.*, vol. 64, pp. 333–341, Dec. 2016.

[28] O. I. Ogunyemi, D. Meeker, H.-E. Kim, N. Ashish, S. Farzaneh, and A. Boxwala, “Identifying Appropriate Reference Data Models for Comparative Effectiveness Research (CER) Studies Based on Data from Clinical Information Systems,” *Med. Care*, vol. 51, pp. S45–S52, Jan. 2013.

[29] P. J. Shih *et al.*, “Estimation of the Corneal Young’s Modulus in Vivo Based on a Fluid-Filled Spherical-Shell Model with Scheimpflug Imaging,” *J. Ophthalmol.*, vol. 2017, 2017.

[30] H. Liyanage, S. T. Liaw, J. Jonnagaddala, W. Hinton, and S. De Lusignan, “Common Data Models (CDMs) to Enhance International Big Data Analytics: A Diabetes Use Case to Compare Three CDMs,” *Stud. Health Technol. Inform.*, vol. 255, pp. 60–64, 2018.

[31] U.S. Department of Health & Human Services, “Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.” [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. [Accessed: 20-Jan-2023].

[32] “Data - The National Patient-Centered Clinical Research Network.” [Online]. Available: https://pcornet.org/data/. [Accessed: 31-Aug-2022].

[33] Sentinel, “Principles & Policies: Privacy .” [Online]. Available: https://www.sentinelinitiative.org/about/principles-policies/privacy. [Accessed: 19-Jan-2023].

[34] Informatics for Integrating Biology & the Bedside, “Data use and confidentiality agreement.” [Online]. Available: https://www.i2b2.org/NLP/DataSets/AgreementAR.php. [Accessed: 19-Jan-2023].

[35] K. Kostka, G. Klebanov, and S. Dempster, “Chapter 20 OHDSI Network Research ,” in *The Book of OHDSI*, 2021.

[36] K. Kostka, “Preserving Privacy in an OMOP CDM Implementation.” [Online]. Available:

https://ohdsi.github.io/CommonDataModel/cdmPrivacy.html. [Accessed: 20-Jan-2023].

[37] Congreso de Colombia, “Ley 1581 de 2012 .” [Online]. Available: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981. [Accessed: 19-Aug-2022].

[38] E. Serrano López, E. L. Rangel, D. R. Roa, C. Rodriguez, and C. Lozano Ortega, “Guía de Anonimización de Datos Estructurados,” pp. 1–82, 2020.

[39] “μ-ARGUS.” [Online]. Available: https://research.cbs.nl/casc/mu.htm. [Accessed: 26-Jan-2023].

[40] “τ-ARGUS.” [Online]. Available: https://research.cbs.nl/casc/tau.htm. [Accessed: 26-Jan-2023].

# ANEXOS

Anexo 1. Guía de usuario para la transformación de datos al CDM de OMOP.