UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

HACIA UN REPOSITORIO ESTANDARIZADO PARA USO SECUNDARIO: TRANSFORMACIÓN DE DATOS DEL REGISTRO CLÍNICO ELECTRÓNICO HOSPITALARIO DEL HOSPITAL DR. HHA AL MODELO COMÚN DE DATOS OMOP

Beatriz Mariane Estada Sarabia

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN INFORMÁTICA MÉDICA

Director de Tesis: Prof. Dr. Steffen Härtel

2025

UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

HACIA UN REPOSITORIO ESTANDARIZADO PARA USO SECUNDARIO: TRANSFORMACIÓN DE DATOS DEL REGISTRO CLÍNICO ELECTRÓNICO HOSPITALARIO DEL HOSPITAL DR. HHA AL MODELO COMÚN DE DATOS OMOP

Beatriz Mariane Estada Sarabia

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN INFORMÁTICA MÉDICA

Director de Tesis: Prof. Dr. Steffen Härtel

2025

ÍNDICE

Resumen

Introducción: En Chile, los Registros Clínicos Electrónicos (RCEs) presentan estructuras no estandarizadas e idiosincráticas que obstaculizan el análisis eficiente, la interoperabilidad y la reutilización de datos, limitando su aplicación en estudios multicéntricos y salud digital basada en evidencia. Para abordar esto, se emplean Modelos Comunes de Datos (CDM) que unifican contenido y estructura. El objetivo de este proyecto fue desarrollar una ETL (*Extract-transform-load*) para transformar los datos del RCE de hospitalizados en el Hospital Dr. Hernán Henríquez Aravena (HHHA) al CDM de OMOP (*Observational Medical Outcomes Partnership*), el cual fue elegido por su estandarización robusta, apoyo a estudios colaborativos y comunidad activa, para promover la interoperabilidad y utilidad de los datos en investigación en salud.

Metodología: Se construyó una ETL para transformar datos del RCE de pacientes hospitalizados (mayores de 18 años, registrados entre junio de 2020 y octubre de 2024) al modelo OMOP CDM. Tras el perfilado, mapeo sintáctico-semántico y la carga, se evaluó la calidad de la base (completitud, plausibilidad y conformidad) con el DQD (*Data Quality Dashboard*), incorporando indicadores de proporción de mapeo y cobertura de codificación para valorar su idoneidad para su uso secundario. Además se realizó la caracterización de la base de datos, con Achilles y Atlas, herramientas del ecosistema OHDSI.

Resultados: Se procesaron un total de 1.226.472 evoluciones correspondientes a 92.073 hospitalizaciones. De estas, el 85.9% incluyó diagnósticos de egreso; el 23,6% incluyó diagnósticos intra-hospitalarios; y el 37.4% incluyó prescripciones. Todos los datos de evoluciones, hospitalizaciones, resultados de laboratorio y diagnósticos de egreso fueron mapeados a conceptos estándar. El 95.2% de los diagnósticos intra-hospitalarios también fueron estandarizados, mientras que el 4.8% restante no pudo ser estandarizado debido a diagnósticos múltiples en un solo campo y abreviaturas ambiguas. Se transformaron 35 variables en seis dominios del OMOP CDM. En términos de calidad, la base alcanzó un 99% de cumplimiento en 1,894 chequeos, excluyéndose 1,117 debido a la falta de datos. Se identificaron y validaron errores menores por parte de los clínicos.

Conclusión: La base de datos transformada del HHHA representa uno de los primeros

repositorios estandarizados y armonizados en Chile, con potencial para generar evidencia científica confiable tanto a nivel local, como nacional e internacional. No obstante, la falta de disponibilidad, estandarización e inconsistencia en los registros de ciertos datos, revelan limitaciones que pueden afectar dimensiones clave de calidad como la completitud, conformidad y plausibilidad. Estos avances amplían significativamente el alcance y la relevancia de las investigaciones desarrolladas sobre la base de OMOP CDM en el país.

Abstract

Introduction: In Chile, Electronic Health Records (EHRs) present non-standardized and idiosyncratic structures that hinder efficient analysis, interoperability, and data reuse, limiting their application in multicenter studies and evidence-based digital health. To address this, Common Data Models (CDMs) are used to unify content and structure. The objective of this project was to develop an ETL (Extract-Transform-Load) process to transform EHR data of hospitalized patients at the Dr. Hernán Henríquez Aravena Hospital (HHHA) into the OMOP CDM (Observational Medical Outcomes Partnership), chosen for its robust standardization, support for collaborative studies, and active community, aiming to promote interoperability and the utility of data in health research.

Methodology: An ETL process was developed to transform EHR data of hospitalized patients (aged over 18, registered between June 2020 and October 2024) into the OMOP CDM model. After profiling, syntactic-semantic mapping, and loading, the quality of the database (completeness, plausibility, and conformity) was evaluated using the DQD (Data Quality Dashboard), incorporating mapping proportion indicators and coding coverage to assess its suitability for secondary use. Additionally, database characterization was carried out using Achilles and Atlas, tools from the OHDSI ecosystem.

Results: A total of 1,226,472 clinical progress notes corresponding to 92,073 hospitalizations were processed. Of these, 85.9% included discharge diagnoses; 23.6% included intra-hospital diagnoses; and 37.4% included prescriptions. All data on progress notes, hospitalizations, laboratory results, and discharge diagnoses were mapped to standard concepts. 95.2% of intra-hospital diagnoses were also standardized, while the remaining 4.8% could not be standardized due to multiple diagnoses in a single field and ambiguous abbreviations. 35 variables were transformed into six OMOP CDM domains. In terms of quality, the database achieved 99% compliance in 1,894 checks, with 1,117 excluded due to missing data. Minor errors were identified and validated by clinicians.

Conclusion: The transformed database from HHHA represents one of the first standardized and harmonized repositories in Chile, with the potential to generate reliable scientific evidence

at local, national, and international levels. However, the lack of availability, standardization, and inconsistencies in certain data records reveal limitations that may affect key quality dimensions such as completeness, conformity, and plausibility. These advancements significantly extend the scope and relevance of research developed using the OMOP CDM framework in the country.

1. Introducción

1.1. Sistemas de Información en Salud en Chile

En Chile, el almacenamiento y gestión de los datos clínicos se caracteriza por una marcada heterogeneidad, tanto en términos de plataformas tecnológicas como de estandarización de terminologías. A nivel de Atención Primaria de Salud (APS), uno de los sistemas más utilizados es RAYEN, una Ficha Clínica Electrónica que opera en cerca de 900 establecimientos en todo el país, principalmente en centros de atención primaria y hospitales comunitarios [1,2]. En el ámbito intrahospitalario, los establecimientos varían entre implementaciones de desarrollo propio y soluciones comerciales nacionales e internacionales. Un ejemplo destacado es el Servicio de Salud Metropolitano Occidente, que adoptó la plataforma TrakCare de InterSystems, implementada también en el Servicio de Salud Coquimbo y en la red privada RedSalud [3,4]. Este sistema permite un RCE unificado entre hospitales, urgencias, atención primaria y SAMU. Por otro lado, el Complejo Asistencial Dr. Sótero del Río, perteneciente al Servicio de Salud Metropolitano Sur Oriente, desarrolló PULSO, vigente desde el año 2014 [5].

Existen además adopciones que involucran a varios centros dentro de un mismo Servicio de Salud, como el RCE regional ALMA, implementado por el Servicio de Salud Coquimbo en hospitales como Ovalle y La Serena, con el objetivo de integrar urgencia, hospitalización, atención ambulatoria, farmacia e imagenología en una sola plataforma clínica [6,7]. El sistema SISMAULE, desarrollado por el Servicio de Salud Maule, integra hospitales y centros de atención primaria en una base de datos única que soporta registros clínicos, estadísticas, categorización de pacientes hospitalizados y listas de espera; aunque su diseño centralizado permite un registro clínico unificado a nivel regional, su escalabilidad nacional se ve limitada por la diversidad tecnológica entre los Servicios de Salud del país [8]. En adición, muchos de estos sistemas funcionan junto a subsistemas específicos como LIS (Sistema de Información de Laboratorio), RIS/PACS (Sistemas de Información Radiológica) y módulos de Farmacia, los cuales en general operan de manera independiente y sin integración completa [9].

1.2. Heterogeneidad, fragmentación y codificación local

En el contexto chileno, los sistemas de información en salud presentan una importante heterogeneidad, tanto en términos tecnológicos como de codificación de datos. Esta diversidad se evidencia en la coexistencia de múltiples plataformas que operan sin una integración efectiva entre niveles de atención o entre establecimientos de una misma red. La falta de estandarización semántica se agudiza con el uso extendido de codificaciones locales o internas para diagnósticos, procedimientos, servicios clínicos y unidades funcionales, lo que dificulta la interoperabilidad y la consolidación longitudinal de la información clínica del paciente. En la práctica, cada fuente de datos observacionales retiene sólo fragmentos del recorrido clínico del paciente, lo que impide obtener conclusiones integrales y libres de sesgos [10]. Se hace necesario entonces verificar la calidad de los registros como proceso previo, ya que no se cuenta con todos los procesos de atención normados, que son la fuente primaria en la obtención de datos [9].

Para enfrentar la situación actual de los RCEs en Chile y promulgación de la Ley N° 21.668, publicada el 28 de mayo de 2024, que introduce la interoperabilidad de las fichas clínicas en el sistema de salud chileno [11], el Ministerio de Salud está impulsando una Estrategia Nacional de Interoperabilidad que busca conectar plataformas heterogéneas mediante estándares como HL7 FHIR, SNOMED CT, CIE-10 y LOINC. Esta arquitectura incluye componentes como el Índice Maestro de Pacientes (MPI) y el Directorio de Prestadores (HPD), que permiten identificar correctamente a pacientes y profesionales. El objetivo es facilitar el intercambio seguro de datos entre niveles de atención, mejorar la continuidad del cuidado y habilitar el uso secundario de la información clínica [12].

1.3. Calidad de los datos clínicos

En la Nota Técnica realizada de la Comisión Nacional de Productividad realizada por Lobos V. [8] se evaluó la calidad de los RCEs utilizados a nivel nacional, donde se evidenció la deficiente calidad de datos almacenados. Este hecho trae múltiples consecuencias, ya que la calidad de atención de los pacientes se ve directamente afectada. Los RCEs del MINSAL se alimentan de los distintos RCEs locales, y hasta el momento no se está realizando una

estandarización de datos centralizada que permita que los sistemas conversen entre sí [9]. Al no existir una comunicación directa entre los RCEs de cada institución y el MINSAL, no existen antecedentes sobre la calidad de los datos alojados en los RCEs locales. El MINSAL, en el marco legal de sus funciones, es el encargado de tratar datos con fines estadísticos y mantener registros y bancos de datos respectos de las materias de su competencia [13]. El DEIS, dependiente del MINSAL, elaboró la Norma Técnica Nº820, la cual es una estandarización que permite diseñar, implementar y mantener actualizados RCE capaces de proporcionar datos estadísticos para la formulación control y evaluación de diferentes programas y los impactos directos que sus acciones generen sobre el estado de salud de la población y la calidad de la atención [14]. Este lineamiento no tiene como fin el intercambio electrónico de información clínica, sino más bien es una estandarización de contenido.

1.4. Uso secundario de los datos clínicos

Desde su implementación, los Registros Clínicos Electrónicos (RCE) se han convertido en extensos repositorios digitales de información sanitaria, almacenada, procesada e intercambiada de forma segura y accesible para usuarios autorizados [15]. Estos repositorios reúnen datos demográficos, prescripciones, diagnósticos, signos vitales, inmunizaciones, resultados de laboratorio y de imagen, conceptos y notas médicas, procedimientos y planes de tratamiento,etc, para facilitar la atención clínica [16]. Por lo anterior, cada vez despierta mayor interés la explotación de estos datos para realizar estudios epidemiológicos, clínicos, de utilización de recursos sanitarios, de evaluación de la calidad asistencial, planificación y gestión sanitaria.

El uso secundario de estos datos puede mejorar las experiencias en la atención médica, ampliar el conocimiento sobre enfermedades y tratamientos apropiados, fortalecer la comprensión de la efectividad y eficiencia de los sistemas de salud, respaldar los objetivos de seguridad y salud pública, fomentar la innovación en tecnologías sanitarias y ayudar a empresas y proveedores a satisfacer las necesidades de los usuarios, mediante actividades como investigación sanitaria, medición de calidad y seguridad asistencial, salud pública,

pagos, certificación y acreditación de proveedores, entre otros [17]. En este contexto, en el estudio realizado por Vuokko [18] se evaluó cómo la estructuración de los datos y la evaluación de la completitud, tanto en el punto de atención como de forma post hoc, mediante el uso de codificaciones, terminologías estandarizadas, plantillas y modelos de información de referencia, influye en la calidad y fiabilidad técnica de los registros clínicos para su uso secundario, evidenciando mejoras en la consistencia y el contenido de la información. La utilidad de los datos de salud observacionales se hace evidente, pero para poder ser utilizados, se hace fundamental evaluar la calidad en la fuente y desarrollar métodos sistemáticos para la evaluación de dimensiones importantes de la calidad de los datos, para mantener su confianza y fiabilidad [19].

Los organismos e instituciones internacionales promueven la creación de repositorios que garanticen la persistencia e integración de los datos conforme a los principios FAIR [20,21]: ***Findability***: Facilidad de búsqueda, a través de metadatos estandarizados; ***Accessibility***: Facilidad de acceso: Mediante accesos controlados y seguros; ***Interoperability***, Interoperabilidad: gracias a vocabularios y formatos semánticos comunes; y ***Reusability***, Facilidad para ser reutilizados; para compartir, reutilizar, y conformar conjuntos de datos fiables, precisos, normalizados, anonimizados, accesibles, semánticamente interoperables y puestos a disposición de la comunidad científica para la generación y replicación de nueva evidencia, la creación de redes de conocimiento biomédico, dirigidas a maximizar el impacto y los beneficios de la investigación en las necesidades de salud de la sociedad [22].

1.5. Modelos Comunes de Datos (CDM)

Los CDMs constituyen conjuntos de estándares de datos uniformes que regulan el formato y el contenido de los datos de observación, respaldan los datos de observación de diferentes fuentes, y forman una estructura de datos estandarizada a través de la extracción, transformación y carga de datos (ETL) [23]. El uso de estándares de comunicación, como los Modelos Comunes de Datos (CDM), solucionaría de manera rápida la correcta integración de los RCE de una red. La ETL genera una reestructuración de los datos en el CDM, la cual

agrega asignaciones a los vocabularios estandarizados y, por lo general, se implementa como un conjunto de secuencias de comandos automatizadas, por ejemplo, secuencias de comandos SQL. Es importante que este proceso ETL sea repetible, de modo que pueda volver a ejecutarse cada vez que se actualicen los datos de origen [10].

Los avances en la estandarización de los datos observacionales de atención médica han impulsado mejoras metodológicas, una colaboración global ágil y la generación de evidencia confiable para optimizar los resultados de los pacientes; La ETL genera una reestructuración de los datos en el CDM, la cual agrega asignaciones a los vocabularios estandarizados y, por lo general, se implementa como un conjunto de secuencias de comandos automatizadas, por ejemplo, secuencias de comandos SQL. Es importante que este proceso ETL sea repetible, de modo que pueda volver a ejecutarse cada vez que se actualicen los datos de origen [10].

Si bien existen varios CDM, en el estudio realizado por Garza et al [24], el CDM de OMOP (*Observational Medical Outcomes Partnership)* se posicionó como el que mejor cumple con los criterios para respaldar el intercambio de datos longitudinales basados en RCE. En este estudio se evaluaron seis categorías: cobertura de contenido, integridad, flexibilidad, facilidad de consulta, compatibilidad de estándares y facilidad y alcance de implementación. OMOP obtuvo el porcentaje más alto de elementos de datos (76%) y tuvo una cobertura terminológica más amplia que los otros modelos.

En resumen, ningún RCE proporciona una visión integral de los datos clínicos que un paciente acumula mientras recibe atención médica y, por lo tanto, ninguno puede ser suficiente para satisfacer todas las necesidades de análisis de resultados esperados. Los avances en la estandarización de los datos observacionales de atención médica han permitido avances metodológicos, una rápida colaboración global y la generación de evidencia confiable para mejorar los resultados de los pacientes [10]. La armonización de las estructuras de datos con CDMs, debe ir acompañada de enfoques estandarizados para la evaluación de la calidad de los datos, para así garantizar la confianza en la evidencia generada a partir del análisis de datos contenidos en los RCEs [25].

1.6. OMOP CDM y la iniciativa OHDSI

El OMOP CDM fue inicialmente desarrollado por la *Iniciativa Observacional de Resultados Médicos* (OMOP) y posteriormente adoptado, mantenido y expandido por la comunidad internacional *Observational Health Data Sciences and Informatics (OHDSI)*. Esta es una red global de investigadores, instituciones académicas y organizaciones públicas y privadas que promueven el uso de datos estandarizados para la generación de evidencia clínica confiable, utilizando un enfoque abierto, colaborativo y transparente [26].

Este modelo define una estructura relacional compuesta por tablas normalizadas que representan distintos dominios (personas, condiciones, procedimientos, medicamentos, observaciones, laboratorios, entre otros), y promueve el uso de vocabularios estándar para garantizar la interoperabilidad semántica entre fuentes de datos. Su adopción ha permitido desarrollar herramientas de análisis avanzadas, como *ATLAS*, *Achilles, CohortMethod y PatientLevelPrediction* [27]. El OMOP CDM ofrece un formato unificado y codificado que armoniza simultáneamente datos de diagnósticos, procedimientos, medicamentos, resultados de laboratorio y otros elementos clínicos procedentes de hospitales, atención primaria y registros administrativos, creando un esquema interoperable que garantiza análisis comparables y reproducibles. Esta estandarización ha impulsado estudios observacionales multicéntricos en campos como enfermedades cardiovasculares, farmacovigilancia y COVID-19, reflejando con fidelidad el recorrido clínico integral de cada paciente y promoviendo la colaboración global en investigación.

1.7. Vocabularios estándar y Athena

Una de las características fundamentales del modelo OMOP CDM es su enfoque en la interoperabilidad semántica, alcanzada mediante el uso de vocabularios estándar. Estos vocabularios permiten mapear los códigos y descripciones locales de los sistemas fuente a conceptos comunes y consistentes, asegurando que los análisis comparables entre distintas bases de datos sean válidos y reproducibles.

Los vocabularios estándar recomendados por OHDSI incluyen SNOMED CT (*Systematized Nomenclature of Medicine Clinical Terms*) para condiciones médicas, RxNorm

y ATC para medicamentos, LOINC (*Logical Observation Identifiers Names and Codes)* para laboratorios y procedimientos diagnósticos, ICD-10/11 (*Clasificación Internacional de Enfermedades, 10 u 11° revisión*) como vocabulario de origen, entre otros [28]. Para facilitar su acceso y mantenimiento, OHDSI provee el portal Athena, una plataforma pública que permite la descarga y exploración de todos los vocabularios compatibles con OMOP CDM, así como sus relaciones jerárquicas y mapeos entre conceptos de origen y estándar [29]. El uso sistemático de estos vocabularios facilita la trazabilidad de los datos, la identificación de cohortes y la estandarización de resultados clínicos en estudios poblacionales a gran escala.

1.8. Herramientas OHDSI para transformación y análisis

La implementación del modelo OMOP CDM requiere el uso de un conjunto de herramientas desarrolladas y mantenidas por la comunidad OHDSI, las cuales permiten llevar a cabo los procesos de extracción, transformación, estandarización, caracterización y análisis de datos clínicos. Entre las principales se encuentran *WhiteRabbit*, que escanea la base de datos original para generar un perfil estructural de sus tablas y campos; y *Rabbit-in-a-Hat*, que permite definir visualmente el mapeo entre las variables origen y las tablas del modelo OMOP CDM [30]. Para apoyar el mapeo semántico de los conceptos locales a vocabularios estándar, se utiliza Usagi, una herramienta que emplea técnicas de procesamiento de lenguaje natural y puntuación de similitud para sugerir conceptos estándar del vocabulario de *Athena* [31]. Una vez finalizado el proceso ETL, se puede caracterizar la base mediante herramientas como *Achilles*, que realiza análisis descriptivos automáticos sobre la población y los dominios clínicos disponibles, y *Data Quality Dashboard* (DQD), que evalúa más de 3.000 reglas para validar la integridad, plausibilidad y conformidad del modelo [32]. Además, herramientas como *ATLAS* y el paquete *FeatureExtraction* en R permiten definir cohortes, ejecutar estudios observacionales y extraer variables clínicas agregadas para estudios analíticos avanzados. El uso conjunto de estas herramientas garantiza transparencia, reproducibilidad y escalabilidad en la transformación y análisis de datos clínicos bajo el estándar OMOP.

1.9. Experiencias latinoamericanas y en Chile

En Latinoamérica, el uso del modelo OMOP CDM ha sido adoptado en proyectos piloto destacados en países como Colombia, Argentina y Brasil. En Brasil, el Hospital Israelita Albert Einstein en São Paulo lideró desde 2018 un piloto para implementar OMOP CDM sobre datos clínicos derivados de su plataforma Cerner, enfrentando desafíos de mapeo semántico, infraestructura ETL y estandarización interna, con resultados reportados hasta 2023 [33]. En Argentina, se han reportado iniciativas regionales en grupos universitarios que participaron activamente del OHDSI Latin America Working Group, adaptando conjuntos de datos observacionales a OMOP para estudios de COVID‑19 y enfermedades crónicas, aunque con publicaciones aún emergentes [34]. En Colombia, instituciones como el Hospital San Vicente Fundación han participado en proyectos colaborativos de armonización de datos clínicos usando OMOP, integrándose a redes multicéntricas para generación de evidencia y caracterización poblacional [34].

En Chile, aún no existen implementaciones OMOP a escala institucional, aunque sí se registra un proyecto académico relevante: una tesis de Magíster en la Universidad de Chile logró armonizar y estructurar los datos clínicos del Hospital Clínico Universidad de Chile bajo el estándar OMOP CDM, mostrando la viabilidad técnica del proceso en contexto local [35]. Además, plataformas como TrakCare de InterSystems, adoptadas por el Servicio de Salud Metropolitano Occidente y otros centros, representan una base técnica desde la cual podría escalarse una conversión institucional hacia OMOP CDM [36]. Estas experiencias demuestran que la adopción del modelo OMOP en Latinoamérica es una tendencia creciente con aplicaciones reales, y en Chile se identifican ya esfuerzos tempranos con potencial de escalamiento institucional mediante iniciativas académicas y la infraestructura tecnológica existente. Este contexto refuerza la relevancia y carácter pionero de tu tesis en el escenario local.

1.10. Contexto institucional: Hospital Regional de Temuco como fuente de datos clínicos

El Hospital Dr. Hernán Henríquez Aravena (HHHA) es un hospital de alta complejidad, ubicado en la ciudad de Temuco, región de la Araucanía, Chile. Administrativamente, depende del Servicio de Salud Araucanía Sur (SSASUR), y es el único establecimiento de Mayor complejidad de la Red Asistencial a la cual pertenece [37]. Cuenta con 713 camas disponibles, distribuidas en las distintas Unidades de Hospitalización. Tomando como referencia los egresos hospitalarios, de 21.395 en 2021; 22.751 en 2022; y 24.879 en 2023 [38], el HHHA acumula información longitudinal de datos clínico-epidemiológicos de más de 60 mil pacientes, constituyendo un recurso fundamental para la investigación clínico-epidemiológica de la población atendida en esta institución.

Desde el año 2019, las atenciones intra-hospitalarias del HHHA se registran en un RCE de desarrollo propio, en línea con los requerimientos definidos por el Ministerio de Salud de Chile (MINSAL), Departamento de Estadística e Información en Salud (DEIS) en la Norma Técnica Nº820 [14]. Esta Norma facilita el diseño, implementación y mantenimiento de los RCE, estandarizando principalmente el contenido, proporcionando datos estadísticos para la formulación, control y evaluación de diferentes programas y los impactos directos que sus acciones generen sobre el estado de salud de la población, por otro lado, este lineamiento no tiene como fin el intercambio, integración o escalabilidad de los datos clínicos, sino más bien es una estandarización de contenido. En ese contexto, al igual que la mayoría de los hospitales del sistema público chileno, el HHHA enfrenta desafíos relacionados con la integración de datos clínicos entre sistemas, la codificación diagnóstica heterogénea y la ausencia de una historia clínica electrónica interoperable, lo que lo convierte en un caso representativo para evaluar la factibilidad de transformación a modelos estandarizados como OMOP CDM.

En el presente estudio se realizará la transformación del RCE del HHHA al estándar OMOP CDM. Una vez realizada la transformación, evaluaremos la calidad de la base de datos transformada, determinando umbrales de decisión, para personalizar este informe de calidad de datos en una variedad de fuentes de datos observacionales en salud. Con esto se obtendrá una base de datos armonizada e interoperable, lo que facilitará el intercambio de información y su

comprensión tanto por los involucrados en el proceso clínico como los investigadores que pueden hacer uso de este Repositorio de Información en Salud creado a partir de un Hospital de Alta Complejidad como es el HHHA.

2. Objetivos

2.1. Objetivo general

Transformar datos asociados a hospitalizaciones de pacientes adultos provenientes del RCE del HHHA, al CDM de OMOP, para evaluar la calidad de los datos y caracterizar la base de datos transformada con el propósito de facilitar su uso en análisis clínicos secundarios.

2.2. Objetivos específicos

OE1. Realizar un perfilado de los datos entregados desde el RCE origen, para comprender su estructura y contenido.

OE2. Realizar un mapeo sintáctico para determinar cómo se traducirán las tablas y los campos de origen al modelo de destino.

OE3. Realizar un mapeo semántico para traducir los códigos de los datos de origen a un vocabulario estándar (LOINC, SNOMED).

OE4. Diseñar y aplicar una herramienta para la Extracción, Transformación y Carga (ETL) de datos.

OE5. Caracterizar la base de datos transformada a OMOP-CDM y evaluar la base de datos transformada en términos de calidad de datos.

3. Metodología

3.1. Diseño del estudio

El presente es un estudio retrospectivo observacional. Fue aprobado por Resolución Exenta N°4218, el 10 de abril de 2024 por el HHHA. Esto contempló la autorización por parte del Comité de Ética del HHHA y del Comité de Ética del Servicio de Salud Araucanía Sur. Se solicitó de manera complementaria el Formulario de Solicitud de Dispensa de Proceso de Consentimiento Informado (Anexo N°1). Todos los datos entregados por el HHHA, fueron previamente anonimizados, identificados con un ID único para cada paciente e ID único de la hospitalización asociada. Se extrajeron datos demográficos, de hospitalización, evoluciones clínicas, diagnósticos, resultados de laboratorio y prescripciones farmacológicas, excluyéndose los procedimientos por encontrarse en texto libre. El detalle de los campos solicitados mediante Diccionario de Datos, se encuentra en el Anexo N° 3. La selección de estas tablas obedeció a su relevancia clínica e investigativa, su alineación con los casos de uso propuestos, la disponibilidad en el centro y su estructuración mediante terminologías estándar [40].

Elegimos OMOP CDM v5.4 por ser la versión más reciente y estable recomendada por la comunidad OHDSI, que incorpora mejoras clave sobre versiones anteriores, como la tabla VISIT_DETAIL para una mayor granularidad de eventos de atención y ofrece un mapeo más robusto de vocabularios estándar (SNOMED CT, LOINC, RxNorm). Además, la v5.4 cuenta con soporte completo del Data Quality Dashboard, lo que nos permite aplicar controles de calidad automatizados y comparables de manera inmediata [6], y garantiza compatibilidad con el ecosistema de herramientas OHDSI (Achilles, ATLAS), facilitando así la caracterización y el análisis reproducible de nuestros datos.

3.2. Fuente de Datos

Los datos empleados en este estudio provienen del RCE del HHHA. Los criterios de inclusión fueron pacientes hospitalizados, mayores de 18 años, entre junio de 2020 y octubre

de 2024. Los datos se obtuvieron a partir de consultas SQL. De este RCE se extrajeron las tablas de egresos (datos demográficos, periodo de hospitalización, condición y destino al alta, diagnóstico de egreso), laboratorio (resultados cuantitativos desde el LIS), fármacos (prescripciones del módulo de farmacia), diagnósticos intrahospitalarios y evoluciones (fecha de registro y tipo de visita). Estas fuentes nativas fueron luego transformadas y armonizadas mediante un proceso ETL para conformar las respectivas tablas del OMOP-CDM.

Para facilitar la comunicación con el Departamento de Informática del HHHA, se elaboró un Diccionario de Datos, detallando el nombre, tipo y descripción de cada variable solicitada. Este diccionario (Tabla Anexa A.1) se utilizó durante todo el proceso de análisis, garantizando el acceso para todos los miembros del equipo.

3.3. Datos Clínicos

Se recopilaron 35 variables por cada paciente, las cuales se agruparon en 8 categorías: Pacientes, Hospitalizaciones, Evoluciones, Diagnósticos de egreso, Diagnósticos Intrahospitalarios, Laboratorio, Fármacos, Defunciones. Dentro de las categorías de Laboratorio, se seleccionaron algunas variables y pruebas específicas, conocidas por su relevancia clínica. La categoría de Laboratorio incluyó 33 variables: Bilirrubina Directa, Bilirrubina Total, Colesterol Total, Triglicéridos, Nitrógeno Ureico, Creatinina, Creatin-kinasa, Creatin-kinasa MB, Troponina T, Glucosa, Fosfatasas Alcalinas, GGT, GOT, GPT, Recuento de Eritrocitos, Hemoglobina, Tiempo de Protrombina, TTPA, Dímero D, Cloro, Potasio, Sodio, Fósforo, Lactato, Lactato Deshidrogenasa, Recuento de Leucocitos, PaO2, Velocidad de Filtración Glomerular (MDRD), Relación Albúmina/creatinina en orina, Proteína C Reactiva, pH y Recuento de Plaquetas. Además se añadieron en este campo los Score APACHE, APACHE II y EUROSCORE, contenido en la tabla DIAGNOSTICOS.

3.4. Infraestructura Tecnológica para la Gestión de Datos

La arquitectura se basó en una máquina virtual con sistema operativo GNU/Linux Debian versión 4.19.171-2, con 188 GB de espacio en disco y 8 GB de memoria RAM,

conectada a 50 TB de almacenamiento compartido a una velocidad de conexión de 10 Gb/s gestionada por el Departamento de Tecnologías de la Información (DTI) del HHHA. Para implementar la base de datos se utilizó pgAdmin 5.8. En cuanto al vocabulario, se utilizó la última actualización disponible en Athena v20250227. A través de un stack Docker Broadsea versión 3.5[1] con dependencias Linux, Mac o Windows con WSL, Docker 1.27.0+, Git y navegadores Chrome, Edge, etc. Dentro de Ohdsi-broadsea se montaron las aplicaciones web de OHDSI (Atlas, WebAPI), junto Postgres v13, DQD y Achilles. Se crearon las tablas destino (39 tablas) las cuales se muestran en la Figura 1 y se importaron vocabularios estándar desde Athena[2].

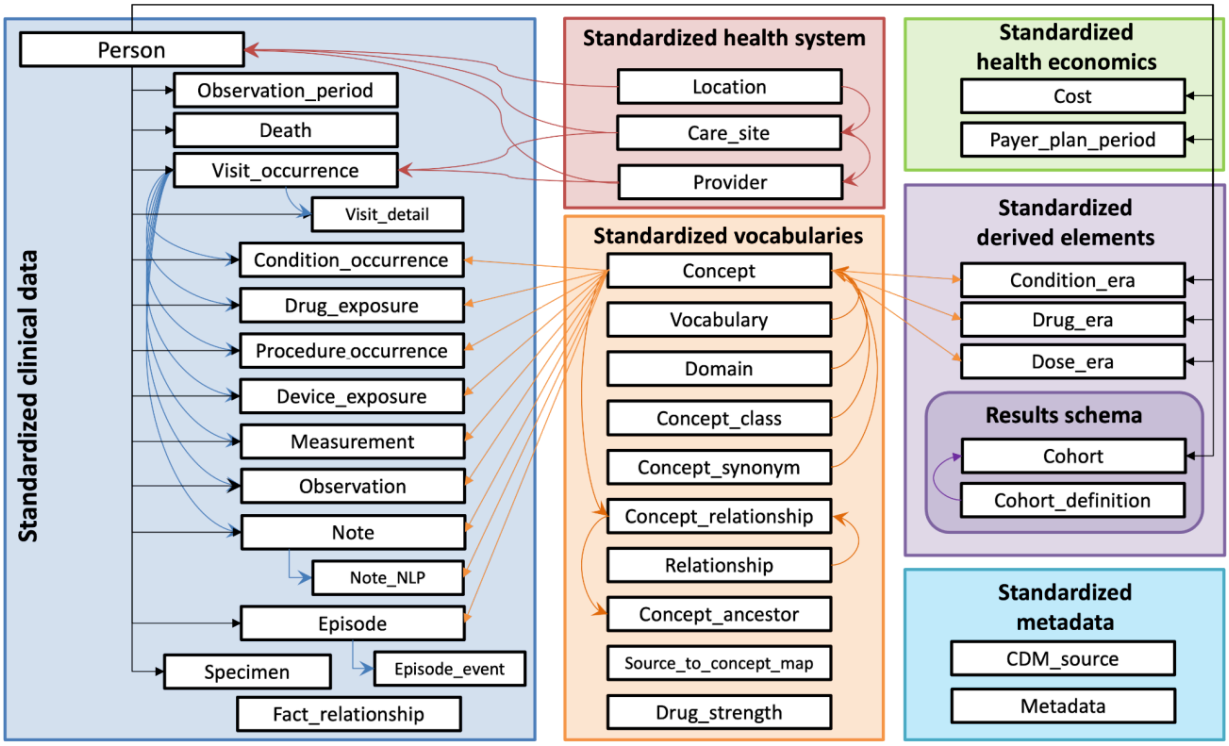


Figura 1: Diagrama del Modelo Relacional OMOP CDM v5.4 [10]. Se presentan 6 dominios: Standardized clinical data, Standardized health system, Standardized vocabularies, Standardized derived elements y Standardized metadata, junto con sus respectivas tablas.

3.5. Transformación de los Datos al Estándar OMOP CDM

El proceso de transformación de la fuente de datos del RCE al estándar OMOP CDM se

[1] https://github.com/OHDSI/Broadsea
[2] https://athena.ohdsi.org/search-terms/start

basó en un flujo bien definido de 6 pasos, que se describe a continuación (Figura 3):

1. *Extracción de datos:* Se extrajeron los datos definidos en la sección 3.2 Fuente de datos.
2. *Perfilado de datos:* Con la herramienta *WhiteRabbit*[3] Realizamos un perfilado de los datos, la cual genera una hoja de cálculo con los valores más frecuentes de cada columna, lo que permite inspeccionarlos y decidir qué columnas ignorar, especialmente aquellas con campos irrelevantes o vacíos. A través del reporte entregado por la herramienta, se realizó un análisis detallado de las tablas, campos y distribuciones de frecuencia de los datos contenidos en la fuente.
3. *Mapeo sintáctico*: Utilizando la herramienta *Rabbit-in-a-Hat*[4], se creó el diseño de la ETL, a partir de los reportes generados con *WhiteRabbit*. En este paso los elementos de datos disponibles en los tablas origen se asignan a los elementos de datos correspondientes en las tablas de destino establecidas por OMOP-CDM v5.4 (39 tablas, distribuidas en 6 categorías), visualizados a través de una interfaz gráfica, en la Figura N° 2 se muestra la interfaz de la herramienta y un ejemplo de mapeo sintáctico.

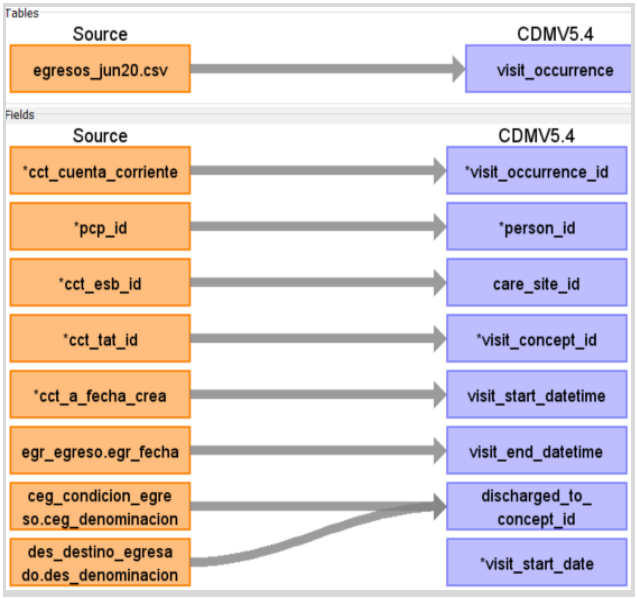


Figura 2: Interfaz gráfica de Rabbit in a Hat utilizada para mapear datos. El panel izquierdo muestra las tablas y campos del sistema origen (tabla *Egresos*). En la zona central, las conexiones visuales representan el mapeo sintáctico entre los campos origen y los del CDM. El panel derecho contiene las tablas estandarizadas del modelo OMOP v5.4 (*visit_occurrence)*. Se presentan los 7 campos mapeados.

[3] https://github.com/OHDSI/WhiteRabbit

[4] https://ohdsi.github.io/WhiteRabbit/RabbitInAHat.html

4. *Mapeo semántico:* El mapeo semántico se realizó con la herramienta Usagi[5], el cual se nutre del vocabulario Athena[1], por lo que es muy importante utilizar la última versión disponible. Además se utilizaron como apoyo browsers de terminologías médicas, principalmente, SNOMED CT. Este mapeo permite normalizar el significado de los conceptos dentro del CDM, a través de vocabularios estándar y se definen por separado según el tipo de dato [30]. La herramienta mencionada permite trabajar con reportes de frecuencia, lo que simplifica el análisis de cada terminología, además permite indicar si el concepto es igual, equivalente, más general o más específico.
5. *Implementación del ETL:* Se realizó la implementación del proceso ETL utilizando el lenguaje de programación R y PgAdmin. La definición de los campos y tablas a implementar se obtienen a partir del mapeo sintáctico, generando un esqueleto SQL que incluye todos los campos que se asignaron, lo que ahorra a los desarrolladores el tiempo de copiar los nombres de los campos a SQL [34].
6. *Caracterización y evaluación de la calidad de datos:*

   6.1 *DQD*: Se realizó una evaluación de la calidad de los datos transformados al OMOP-CDM con la herramienta DQD, que evalúa la calidad intrínseca de los datos, utilizando los criterios: Conformancia, Completitud y Plausibilidad. El detalle de la metodología utilizada en esta etapa se abordó en la siguiente sección del artículo, donde se realizaron análisis más detallados para evaluar la calidad de los datos mapeados, incluyendo la validación de los mapeos sintáctico y semántico.

   6.2. *Achilles/Atlas*: Se realizó la caracterización de la base de datos OMOP CDM. Para ello, se emplearon herramientas del ecosistema OHDSI que permiten describir en detalle las características poblacionales, clínicas y de utilización de servicios de salud. Inicialmente, se utilizó Achilles[6] para generar un perfil descriptivo general de la base, incluyendo distribuciones por edad, sexo, condiciones más frecuentes y evolución temporal de eventos clínicos. Posteriormente, se utilizó la plataforma web ATLAS[7], donde se visualizan los reportes de diagnósticos, medicamentos y procedimientos en distintos contextos clínicos, generados a partir de los análisis de Achilles.

[5] https://ohdsi.github.io/Usagi
[6] https://github.com/OHDSI/Achilles
[7] https://github.com/OHDSI/Atlas

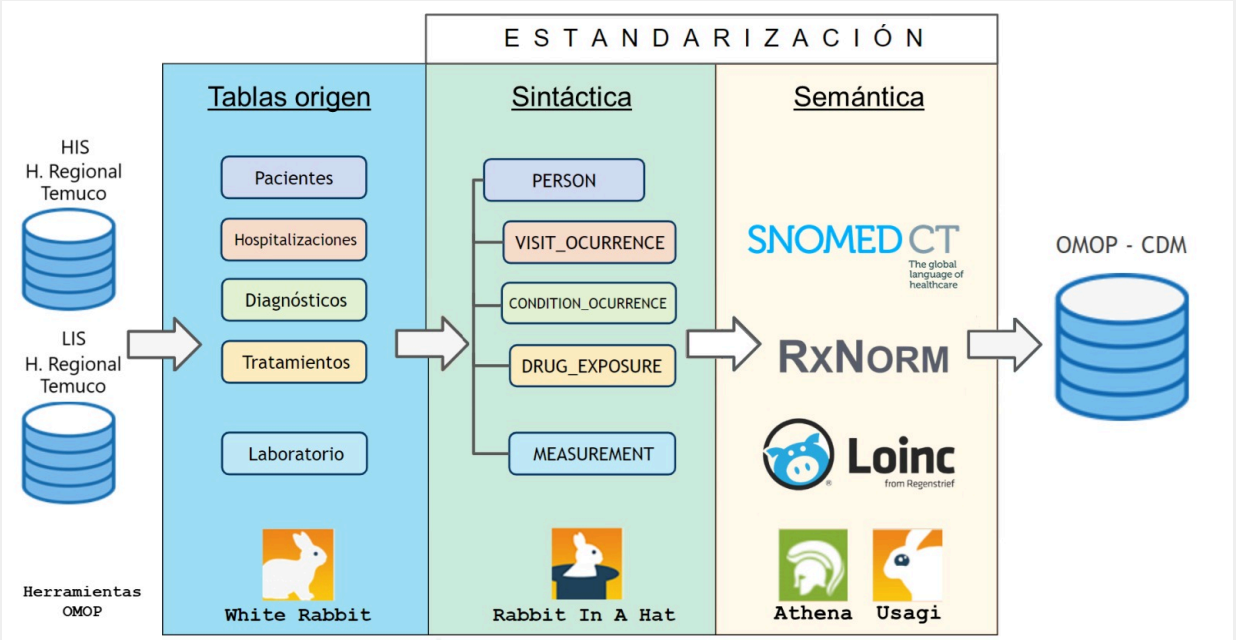


Figura 3: Diagrama del proceso de transformación de Tablas de Origen a OMOP CDM. A partir de las tablas origen, se realiza un perfilado a las tablas origen con la herramienta White Rabbit. El proceso continúa con la Estandarización Sintáctica, con la herramienta Rabbit In A Hat, para luego realizar la Estandarización Semántica con la herramienta Usagi - Athena. Se finaliza el proceso con la carga de los datos transformados a la base de datos OMOP CDM.

4. Resultados

4.1 Extracción de los Datos contenidos en el RCE del HHHA

Se lograron extraer datos clínico-administrativos de pacientes hospitalizados entre junio de 2020 y octubre de 2024. Estos datos fueron proporcionadas por la DTI del HHHA obtenidos desde las bases de datos del HHHA y sometidos a un proceso de transformación y carga utilizando PgAdmin 4 v8[8], plataforma diseñada para la gestión de bases de datos postgresSQL. Se utilizó un esquema de carga temporal conocida como Staging, donde los datos se transforman siguiendo la estructura del Diccionario de Datos (Anexo N°3). Esto permite realizar una revisión y corrección de errores antes de cargar el OMOP-CDM final.

[8] www.pgadmin.org

4.2 Perfilamiento de los Datos

Se generó el perfilamiento de la fuente de datos usando *WhiteRabbit*. Esta herramienta entregó un informe en formato Excel llamado ScanReport, el cual se muestra en el Anexo N°2. Se identificaron un total de 27 campos distribuidos en 5 tablas, mostrando consistencia con la estructura planificada del Diccionario de Datos. A continuación se muestran los resultados obtenidos a partir de los reportes, lo cual se muestra además en la figura N°zzz:

1. *Egresos*: El RCE incluye datos de 92.073 hospitalizaciones de 68.358 pacientes. Esta tabla contiene el 100% de los datos asociados a sexo, fecha de nacimiento, previsión y fecha de ingreso; 85.9% (79.072) datos de fecha de egreso, diagnóstico principal y condición al alta; 94.6% (87.120) del destino al alta. Se incluyen los campos condición y destino al alta por encontrarse discrepancias en el registro, por lo que ambos campos se complementan.

2. *Evoluciones*: De las 92,073 hospitalizaciones, 53.9% tienen evoluciones asociadas. Se recopilaron datos de 1,268,938 evoluciones. Esta tabla contiene el 100% de los datos de: Servicio de Hospitalización, ID de la Hospitalización, ID de paciente, ID de evolución y el tipo de atención; 98.6% fecha de evolución.

3. *Diagnósticos*: En esta tabla se registran 21,748 (24%) hospitalizaciones con diagnósticos adicionales asociados. Del total de 80,270 diagnósticos asociados, se tiene el 100% de los datos de: ID paciente, ID de la hospitalización e ID del diagnóstico y fecha de registro del diagnóstico; 29.1% de los códigos CIE-10. Al revisar los datos, existen varios registros equivalentes al mismo diagnóstico, por ejemplo: Diabetes Mellitus, Diabetes Mellitus tipo 2, Diabetes Mellitus 2, Diabetes Mellitus Tipo II, entre otros, por lo que una vez se realice la estandarización semántica, se tendrá la frecuencia real de cada diagnóstico.

4. *Fármacos:* La tabla de fármacos contiene la información de 34,402 hospitalizaciones (37.4%). De un total de 492,116 indicaciones, se tiene el 100% de los datos de: ID de paciente, ID de hospitalización, nombre, Dosis/Unidad, vía de administración y fecha de indicación del fármaco.

5. *Laboratorio:* Las tablas de laboratorio contienen en total 6,262,942 registros de exámenes. De estas tablas se tiene el 100% de los datos de: paciente, nombre del examen, fecha de registro, fecha de validación, ID de la solicitud del examen y resultado del examen.

6. *Defunciones:* La tabla defunciones fue extraída del reporte de egresos hospitalarios, cuya condición de egreso fue fallecido, se complemento con el campo estado al alta. En total se reportaron 2335 defunciones de las cuales el 1.9% (45 en total) no registran diagnóstico de egreso, por lo que no se pudo determinar la causa de fallecimiento.

4.3 Mapeo Sintáctico

El mapeo sintáctico realizado con *Rabbit-In-a-Hat* permitió realizar el diseño de la ETL. Durante este proceso, se seleccionaron las tablas del OMOP-CDM que en una etapa posterior se poblaron con los datos obtenidos desde el RCE fuente. Basándonos en las tablas origen del RCE del HHHA, se pudieron mapear 35 campos. Los campos mapeados se pueden visibilizar en el Anexo N°3. La mayoría de las tablas tienen en común el identificador anonimizado del paciente y el identificador único de la hospitalización asociada. La excepción se presentó en las tablas asociadas a resultados de laboratorio, ya que el registro estaba asociado a un identificador distinto. Con la herramienta R[9], se logró hacer un merge de las tablas para que tuvieran un identificador común. Las tablas del OMOP CDM v5.4 que se lograron poblar fueron: PERSON, OBSERVATION_PERIOD, VISIT_OCCURRENCE, VISIT_DETAIL, MEASUREMENT, CONDITION_OCCURRENCE y DRUG_EXPOSURE y DEATH. Se lograron transformar exitosamente un total de 8,286,845 registros asociados a 68.358 pacientes hospitalizados desde el RCE del HHHA al estándar OMOP CDM.

En la Tabla 1 se muestra la cantidad de conceptos origen y la cantidad de conceptos mapeados al estándar OMOP. Se logró migrar en promedio un 96.3% de los registros. La principal fuente de incompletitud proviene de los egresos hospitalarios, ya que existen 13,002 hospitalizaciones sin fecha de egreso ni diagnóstico asociado, por lo que se ven afectadas las tablas VISIT_OCCURRENCE (85.8%) y CONDITION_OCCURRENCE (85.8%).

[9] www.r-project.org

Fuente de datos RCE		OMOP CDM v5.4		
Nombre tabla	Registros (N=8.318.728)	Nombre tabla	Registros (N=8.286.845)	Proporción migrada
Pacientes	68.358	person	68.358	100%
Hospitalizaciones	92.073	visit_occurrence	79.072	85.8%
Diagnóstico egreso	92.073	condition_occurrence	81.497	85.8%
Diagnóstico intra	80.408	condition_occurrence	72.590	95.2%
Fármacos	492.565	drug_exposure	492.565	100%
Pruebas de laboratorio	6.262.942	measurement	6.262.449	99.9%
apache_score	1.507	measurement	1.507	100%
Evoluciones	1.226.472	visit_detail	1.226.472	100%
Defunciones	2.335	death	2.335	100%

Tabla 1. Comparación de registros migrados exitosamente desde la fuente de datos del RCE a las tablas OMOP CDM.

4.4. Mapeo Semántico

Los datos contenidos en el RCE son exclusivamente de pacientes hospitalizados, por lo tanto, en la tabla VISIT_OCCURRENCE, en el campo VISIT_CONCEPT_ID se utilizó el concepto 8717 (Inpatient Hospital) el cual pertenece al dominio visit y se utilizó el concepto 32817 (EHR) en el campo VISIT_TYPE_CONCEPT_ID, el cual hace alusión a la procedencia del dato.

La tabla CONDITION_OCCURRENCE tuvo dos fuentes de datos, por un lado los diagnósticos de egreso, para los cuales se utilizó el concepto 32823 (EHR discharge record) perteneciente al dominio type_concept, que indica la procedencia del registro, y el concepto 32896 (Discharge diagnosis) perteneciente al dominio Condition status, concepto que representa el momento durante la visita en que se registró el diagnóstico. La segunda tabla contiene los diagnósticos intra-hospitalarios, por lo que se utilizó el concepto 32817 (EHR) en el campo CONDITION_TYPE_CONCEPT_ID y los concepto 32899 (Preliminary diagnosis); 32893 (Confirmed diagnosis) para el campo CONDITION_STATUS_CONCEPT_ID.

La tabla VISIT_DETAIL, contiene información de las visitas realizadas durante el periodo de hospitalización. Además, contiene datos sobre el Servicio o Unidad donde fue realizada la visita y el tipo de evaluación realizada. OMOP incluidos en sus conceptos, de dominio *visit*, los conceptos estándar para los Servicios, ya que este campo debe poblarse obligatoriamente con conceptos del dominio visit, por lo que, se siguió la recomendación de la comunidad, la cual indica que: UCI y UTI se mapean al concepto 32037 (Intensive care); Los servicios de Urgencia, tanto obstétrica como adulto, se mapean al concepto 262 (Emergency Room and Inpatient Visit) y todos los demás servicios se mapean al concepto 9201 (Inpatient visit). Si bien este mapeo no responde exactamente al concepto, se incluyen los VISIT_DETAIL_SOURCE_VALUE, donde se registra el concepto fuente de manera complementaria. Esto permitió capturar y registrar de manera precisa la información sobre las hospitalizaciones y visitas asociadas a cada paciente.

En la tabla MEASUREMENT, se observaron algunos resultados que contenían operadores como (Mayor o Menor), junto al resultado numérico. Para mapear estos conceptos, se realizó una separación de los registros para aislar el operador, y luego se mapearon a los conceptos OPERATOR_CONCEPT_ID: 4139823 (Larger); 4126988 (Lower) y 4084765 (Above reference range).

De la tabla origen DIAGNOSTICOS, la cual contenía 80,408 diagnósticos, se lograron mapear 95.2% (72,590) de los conceptos, esto debido principalmente al registro en formato texto desestructurado, no fue posible asociar los conceptos origen a uno estándar.

El proceso de mapeo semántico se logró exitosamente, alcanzando el 100% para los 5,993 conceptos origen (Tabla 2). Los datos del RCE fueron mapeados mediante Usagi, Athena y buscadores SNOMED CT, sumado a la colaboración de dos expertos, hacia los estándares LOINC, UCUM, SNOMED-CT, RxNorm y OMOP.

Fuente de datos RCE			OMOP CDM v5.4		
Categoría	Nombre tabla	Conceptos origen	Nombre tabla	Conceptos Destino	Proporción migrada
Género	egresos	3	person	3	100%
Destino al alta	egresos	10	visit_occurrence	10	100%
Diagnóstico de egreso	egresos	2470	condition_occurrence	2470	100%
Diagnóstico Intrahospitalario	diagnosticos	2214	condition_occurrence	2212	100%
Fármacos	farmacos	1159	drug_exposure	1159	100%
Vía de administración	farmacos	17	drug_exposure	17	100%
Pruebas de laboratorio	laboratorio	32	measurement	32	100%
Unidades	laboratorio	13	measurement	13	100%
Operador - resultado	laboratorio	3	measurement	3	100%
Servicios	evoluciones	35	visit_detail	35	100%
Tipo de evolución	evoluciones	15	visit_detail	15	100%
Total		5995		5993	100%

Tabla 2. Mapeo semántico. Muestra los conceptos mapeados a la terminología estándar OMOP.

4.5. Implementación de la ETL

Para la implementación de la ETL se utilizó el stack docker OHDSI-Broadsea, que agrupa las principales herramientas OMOP que se utilizaron: DQD, Achilles, ATLAS junto con el contendor Postgres que contiene el OMOP CDM. Tras una exitosa implementación del proceso ETL, se logró una integración eficiente de los datos origen al modelo OMOP CDM. Para llevar a cabo esta implementación se utilizó además la Documentación ubicada en los repositorios de la comunidad OHDSI a través de GitHub y el foro OHDSI[10], el cual por un lado contiene mucha información sobre el proceso de implementación de una ETL y además existen miembros que continuamente responden dudas y ayudan a solucionar problemas asociados a las herramientas OHDSI. Como resultado, se poblaron las tablas de los dominios Standardized clinical data y Standardized vocabularies.

4.6. Caracterización y Evaluación de la calidad de los datos OMOP

La caracterización de la base de datos se realizó con la herramienta Achilles en combinación WebAPI y Atlas, para algunas consultas específicas se realizaron consultas SQL.

[10] https://forums.ohdsi.org

En la figura 4 se visualiza la interfaz de la herramienta ATLAS, con la cual es posible crear cohortes específicas para realizar estudios avanzados.

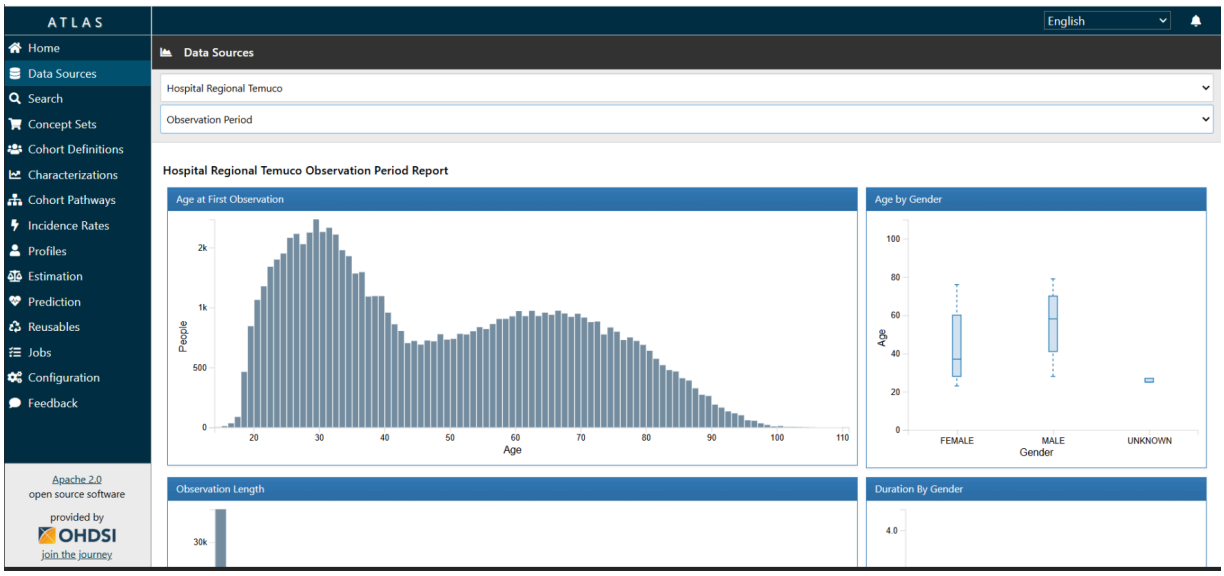


Figura 4. Visualización de la Interfaz de la Herramienta ATLAS, se observa la distribución de las hospitalizaciones por edad y por género.

4.6.1. Caracterización de la Base

- **Hospitalizaciones**

Durante el período comprendido entre junio de 2020 y octubre de 2024 se registraron un total de 92,073 hospitalizaciones en la base de datos, correspondientes a 68,358 pacientes únicos. Con un promedio de 1.35 hospitalizaciones por paciente. Del total de pacientes, 64.7% (44,253) corresponden a sexo femenino y un 35.3% (24,103) de sexo masculino. En la figura N°5 se observa la distribución de edad de la primera hospitalización la cual muestra un patrón bimodal: un peak entre los 10-20 y otro entre los 40-70 años. La frecuencia disminuye después de los 70 años y se observan casos aislados en edades avanzadas, reflejando posibles ingresos tardíos o consolidación retrospectiva de datos.

Los resultados obtenidos desde la base de datos armonizada al modelo OMOP CDM con la fecha de egreso registrada, indican un total de 16,948 egresos hospitalarios en el año 2021, 17,339 en 2022 y 19,070 en 2023. Frente a la falta de fechas de egreso registradas, se

realizó una corrección a partir del registro de evoluciones, agregando la fecha de la última evolución como fecha de egreso. Los resultados finales son: 19,876 en el 2021, 19,808 en el 2022 y 22,116 en el 2023. Al comparar estas cifras con las reportadas oficialmente en la Cuenta Pública del HHHA, excluyendo los egresos de pacientes menores de 18 años: 17,886 egresos en 2021, 18,171 en 2022 y 19,991 en 2023; se observa un registro en el CDM superior de egresos en cada año. En 2021, la diferencia absoluta fue de 1,990 egresos, lo que representa un 10% menos respecto al total obtenido. En 2022, la discrepancia fue de 1,637 egresos, equivalente al 8.2%, mientras que en 2023 se registraron 2,125 egresos más que los informados oficialmente, lo que corresponde a una diferencia relativa del 9.6%.

En la figura 5 se puede observar que la mayor densidad de datos corresponde a exámenes de laboratorio, aunque experimenta una fuerte caída en el mes de enero de 2021. Además se logra observar un creciente aumento de datos asociados a exposición a fármacos a partir de octubre de 2021 en adelante, provocando también un aumento de la categoría DRUG_ERA.

En la figura 6 se visualiza la densidad de datos en cuanto a la cantidad de registros y la cantidad de conceptos agrupados en categorías. Se observa que las categorías de DRUG_EXPOSURE y MEASURMENT presentan una mediana más alta y una mayor dispersión en comparación a CONDITION_OCCURRENCE y CONDITION_ERA, teniendo asociados menos de 5 conceptos por persona en esas categorías.

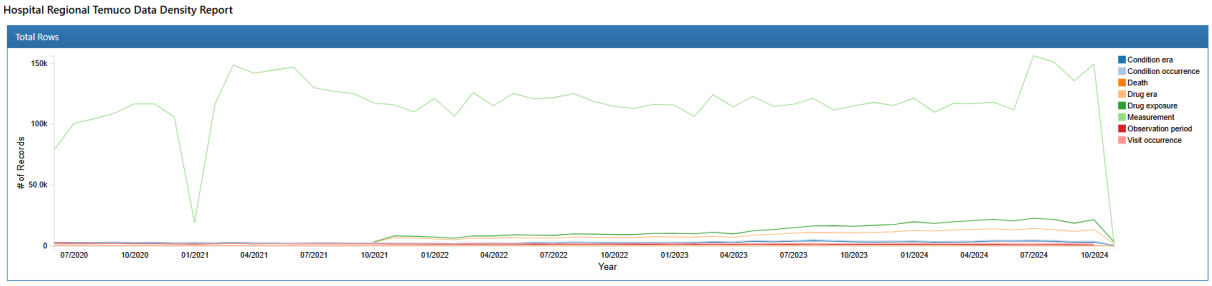


Figura 5. Densidad de datos. Muestra la densidad de la población de datos, separados por categorías como condiciones, medicamentos, procedimientos, mediciones y observaciones según el modelo OMOP CDM.

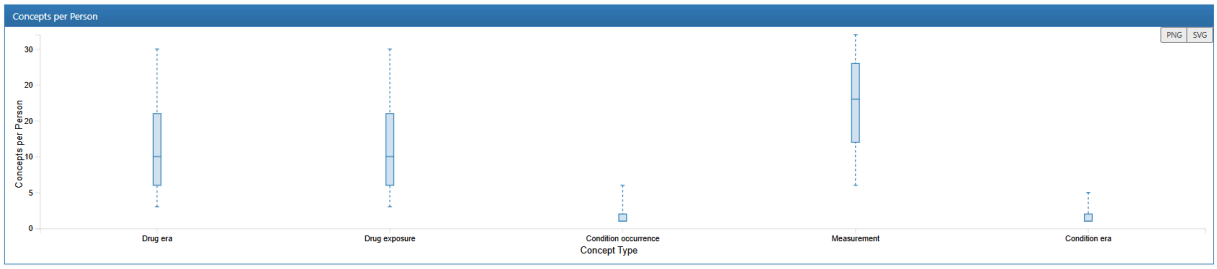


Figura 6. Conceptos por persona. Muestra la distribución del número de conceptos registrados por persona en categorías como condiciones, medicamentos, procedimientos, mediciones y observaciones según el modelo OMOP CDM.

- **Visitas o evoluciones**

El total de visitas intrahospitalarias realizadas es de 1,226,472. Las visitas origen se registran en relación al Servicio de Hospitalización asociado. La terminología estándar no incorpora a los Servicios de Hospitalización con la granularidad esperada, ya que el CDM OMOP requiere que el concepto utilizado pertenezca al dominio *visit* y dentro de ese dominio no existen los Servicios de Hospitalización. Visitas en hospitalización (*Inpatient Visit*), con 710,859 registros (57,96%). Le siguen las visitas en cuidados intensivos más UTI (Intensive Care), que suman 481,341 registros (39.25%); (Emergency Room and Inpatient Visit) con 34,077 visitas (2.78%) y 195 registros (0.02%) sin concepto asociado por ausencia de registro.

- **Diagnósticos**

Del total de 68,358 pacientes registrados en la base de datos, 60,856 personas presentaron al menos una condición clínica documentada, lo que equivale al 89.0%. El 11.0% restante corresponde a pacientes sin diagnósticos clínicos codificados, lo cual se debe principalmente a la ausencia del registro de condiciones en la tabla egresos hospitalarios. A continuación se muestra la figura 7 que contiene los diagnósticos hospitalarios más frecuentes agrupados en dos grandes categorías: diagnósticos obstétricos y diagnósticos generales o médicas. Se muestra además la tabla N°4 que muestra la distribución de Diagnósticos por grupo etario y género. Del total de 161,905 diagnósticos, existen 7,818 (4.8%) registros sin concepto estandarizado ("No matching concept") debido a que no fue posible interpretar la

descripción del diagnóstico y tampoco presentaban una codificación que permitiera identificarlo.

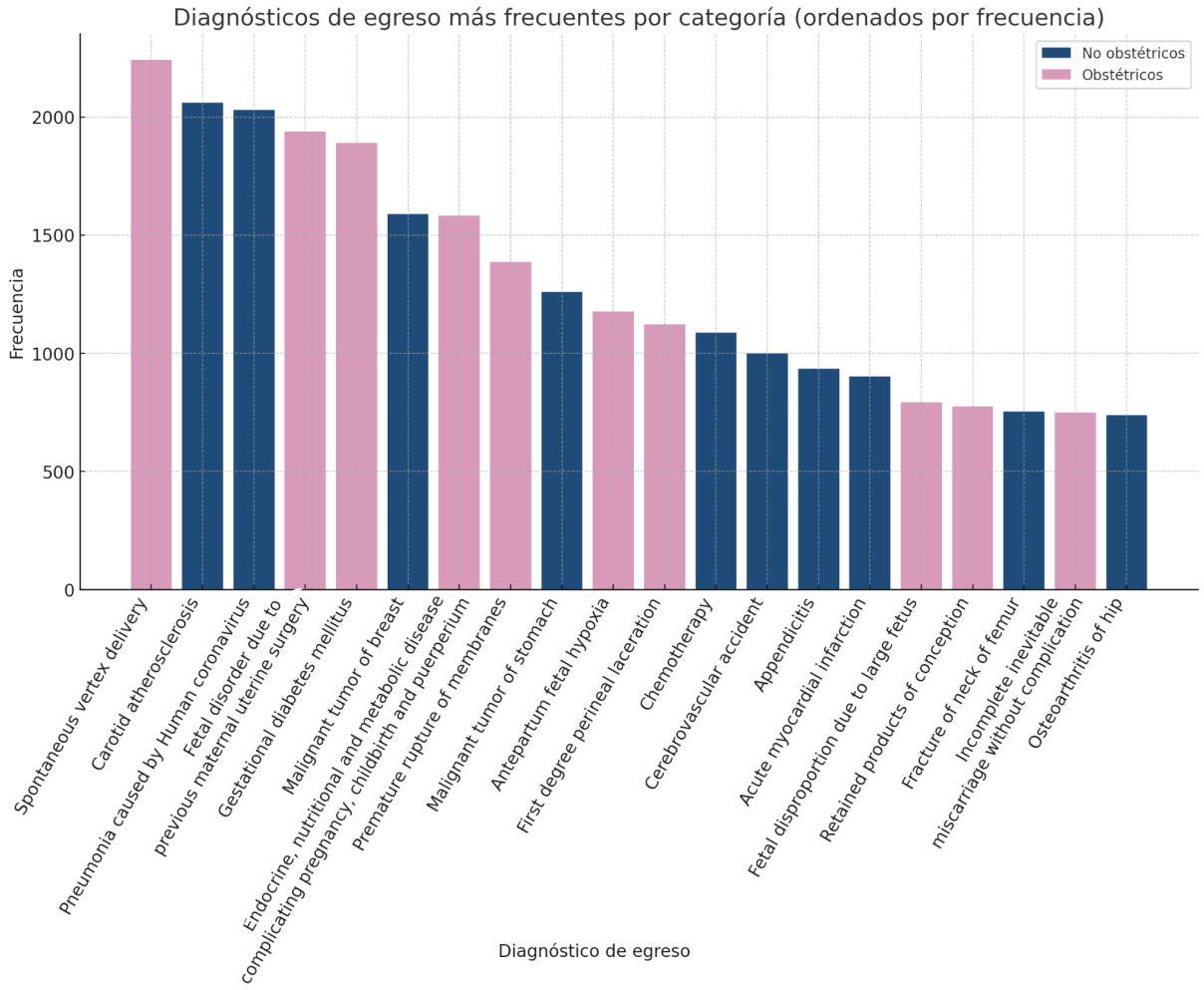


Figura 7. Diagnósticos más frecuentes separados por categorías.

- **Fármacos**

Durante el periodo analizado, se registraron un total de 492,565 exposiciones a medicamentos, correspondientes a 27,138 (39.7%) pacientes únicos. A partir de lo anterior se analizó la polifarmacia, obteniéndose que 22,645 pacientes (83.44%) de los pacientes presentan exposición a más de 5 fármacos y 4,493 pacientes (16.56%) con exposiciones menores o iguales a 5, reflejando un grupo menor con terapias más simples.

- **Resultados de laboratorio**

Se seleccionaron por su relevancia 32 análisis de laboratorio y se adicionaron 3 provenientes de la tabla diagnósticos que corresponden a scores (Euroscore, Apache y Apache II). Se tiene el registro de 6,264,449 de mediciones registradas, donde 68,358 pacientes (100%) tienen por lo menos un examen asociado. La figura 8 muestra las 20 combinaciones más frecuentes de exámenes de laboratorio y su estado respecto al rango de referencia. Destacan resultados "dentro de rango" en pruebas como plaquetas, leucocitos, creatinina y enzimas hepáticas. También se observa una alta frecuencia de valores elevados en pruebas como tiempo de protrombina (PT) y proteína C reactiva, así como valores bajos en hemoglobina y glomerular filtrado estimado (MDRD), lo que refleja alteraciones comunes en pacientes hospitalizados.

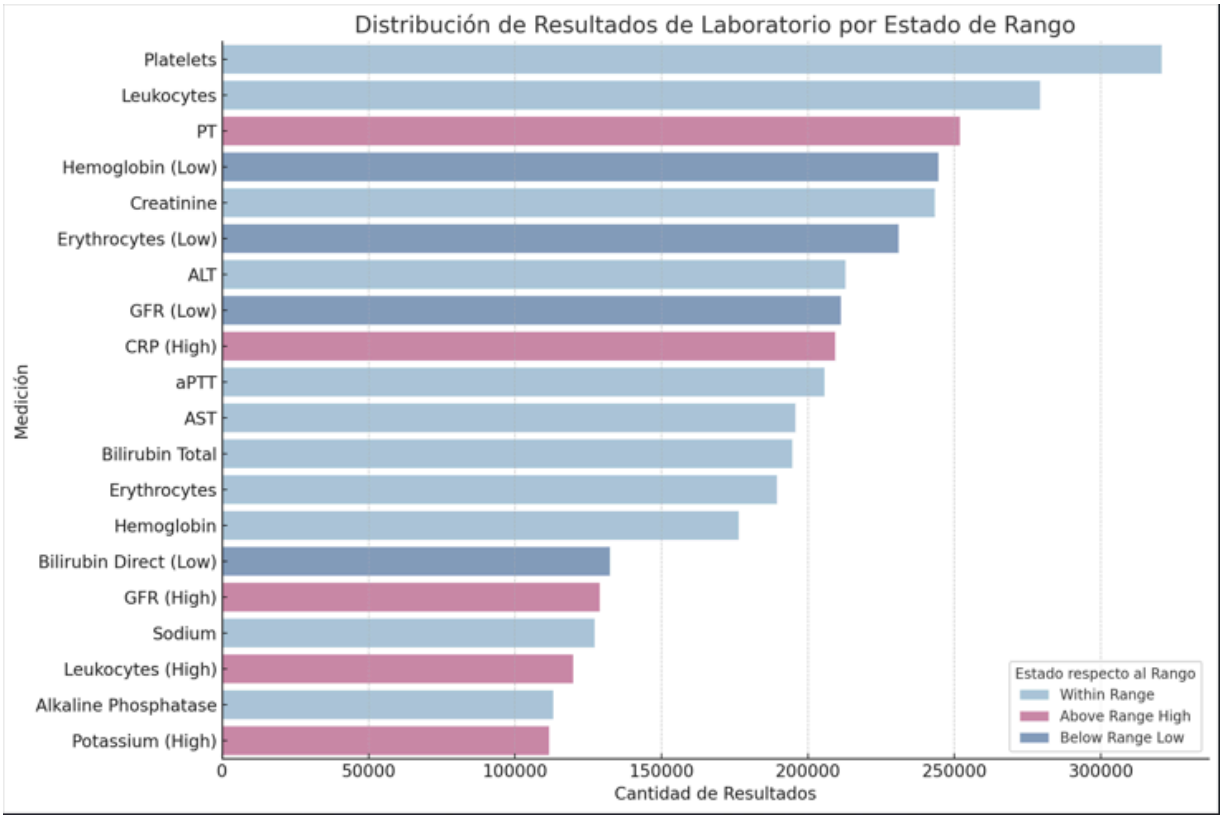


Figura 8. Resultados de laboratorio distribuidos por estado: Dentro del Rango, Bajo el rango y sobre el rango de referencia.

- **Defunciones**

Existe un total de 45 defunciones mapeadas como “No matching concept”, lo que indica muertes sin codificación estándar, esto se debió principalmente a la ausencia del registro en los egresos hospitalarios entregados. Durante el período 2020–2024, se registraron en la base OMOP un total de 2,335 defunciones, 1,529 defunciones entre 2021 y 2023: 399 en 2021, 536 en 2022 y 594 en 2023. Al comparar estas cifras con los valores oficiales reportados por el HHHA, se observan discrepancias en los tres años: En 2021, se reportaron 593 muertes oficialmente, pero solo 399 figuran en la base (67.2% de cobertura); en 2022 se reportaron 516, comparado con lo obtenido en la base de datos transformada, que fueron 536, se evidencia un aumento del 3.8% y en 2023, se reportan oficialmente 565 muertes, en la base de datos transformada se reportaron 594, 5.1% superior al valor reportado. Estas diferencias podrían deberse a falta de completitud en los egresos hospitalarios entregados, registros sin codificación de defunción, o errores de mapeo durante la transformación al modelo OMOP. Es crucial considerar esta limitación al interpretar los resultados de mortalidad y realizar análisis longitudinales o comparativos con fuentes institucionales.

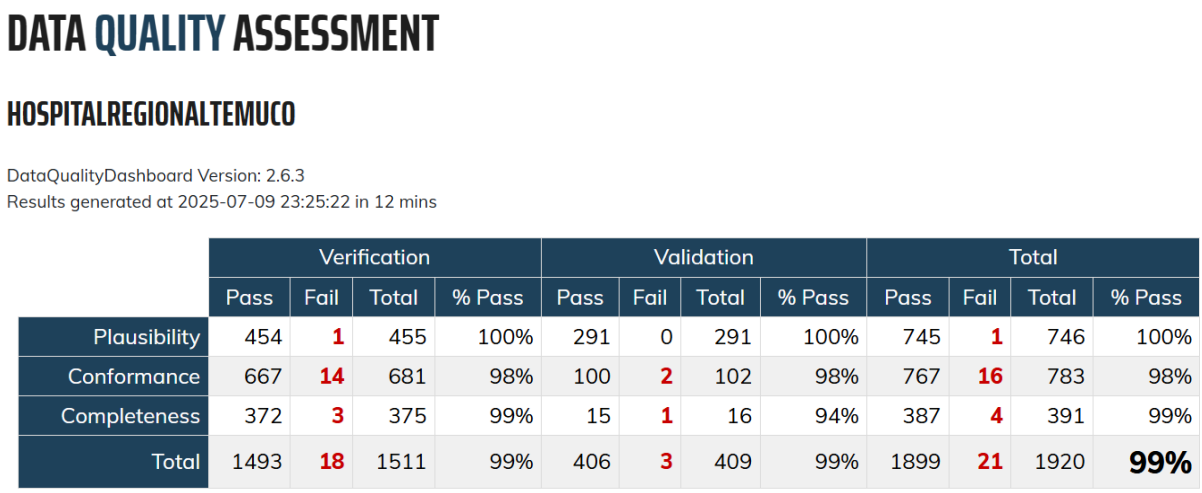
- **Rehospitalizaciones**

Los resultados obtenidos muestran la evolución anual de la tasa de rehospitalización a 30 días en la base de datos analizada, considerando el total de hospitalizaciones índice (primer ingreso hospitalario en el período): En 2021, se registraron 20,193 hospitalizaciones, con 1,383 rehospitalizaciones en 30 días, lo que representa una tasa de 6.85%; En 2022, hubo 19,932 hospitalizaciones y 1,528 rehospitalizaciones, con una tasa de 7,67%; En 2023, se observaron 22,008 hospitalizaciones, de las cuales 2,004 resultaron en rehospitalizaciones dentro de 30 días, con una tasa de 9.11%. El motivo más común de re-hospitalización fue la Quimioterapia, con una frecuencia elevada y sostenida a lo largo de los tres años. Le siguen las neoplasias malignas, destacando el tumor del estómago, el colon, el recto, el linfoma no Hodgkin y el tumor del testículo, todos con cifras significativas de rehospitalización. Entre las condiciones metabólicas y cardiovasculares se encuentran la hipertensión arterial y la diabetes mellitus tipo 2, así como la enfermedad aterosclerótica del corazón. En el ámbito obstétrico, se

reportaron con frecuencia la diabetes gestacional, la atención materna por cicatriz uterina, y el trabajo de parto prematuro sin parto, evidenciando una importante carga de re-hospitalización en mujeres embarazadas.

4.6.2. Calidad de los Datos

La Figura 9 presenta la evaluación obtenida de la calidad de datos en las dimensiones Conformancia y Completitud. En la dimensión de Plausibilidad, se realizaron un total de 455 chequeos, de los cuales 2 resultaron en errores, lo que representa un cumplimiento del 100%. En la dimensión de Conformancia, se llevaron a cabo 681 chequeos, de los cuales 18 presentaron fallas, con un cumplimiento del 97%. En la dimensión de Completitud, se efectuaron 375 chequeos, de los cuales 3 arrojaron errores, con un cumplimiento del 99%. En general, la base de datos transformada a OMOP ha obtenido un alto porcentaje de aprobación, con un cumplimiento total del 99% en un total de 1,894 chequeos. La visualización ha sido generada utilizando la herramienta DQD, proporcionando una visión precisa del estado de la calidad de los datos. La tabla 3 muestra los análisis que fallaron. Los errores fueron principalmente por fallos en la conformancia (16 fallos), principalmente asociados a la ausencia de conceptos, con respecto a la raza (RACE) y etnia (ETHNICITY), ya que no existe registro de esos campos en las tablas origen, o conceptos mapeados a un dominio distinto del requerido por OMOP. Se identificaron 4 fallos asociados a completitud: ausencia de VISIT_END_DATE en la tabla VISIT_OCCURRENCE (6.62%), ausencia de la tabla CONDITION_ERA (100%), situación que fue subsanada previo a la ejecución de Achilles; ausencia de concepto estándar en el campo VISIT_DETAIL_CONCEPT_ID (0.02%); ausencia de CONDITION_SOURCE_VALUE de la tabla CONDITION_OCCURRENCE, relacionado con la ausencia de registro codificado en las tablas origen (11.9%); Por último, un fallo asociado a la plausibilidad, en el campo QUANTITY de la tabla DRUG_EXPOSURE.



DATA QUALITY ASSESSMENT

HOSPITALREGIONALTEMUCO

DataQualityDashboard Version: 2.6.3
Results generated at 2025-07-09 23:25:22 in 12 mins

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	454	1	455	100%	291	0	291	100%	745	1	746	100%
Conformance	667	14	681	98%	100	2	102	98%	767	16	783	98%
Completeness	372	3	375	99%	15	1	16	94%	387	4	391	99%
Total	1493	18	1511	99%	406	3	409	99%	1899	21	1920	**99%**

Figura 9. Análisis detallado de la calidad de los datos en las dimensiones Plausibilidad, Conformancia y Completitud. Visualización obtenida desde la herramienta OHDSI DataQualityDashboard (DQD) v1.4.1

STATUS	TABLE	FIELD	CHECK	CATEGORY	SUBCATEGORY	LEVEL	NOTES	% RECORDS
FAIL	CONDITION_ERA	None	measureConditionEraCompleteness	Completeness	None	TABLE	None	100.00%
FAIL	PERSON	RACE_CONCEPT_ID	isForeignKey	Conformance	Relational	FIELD	None	100.00%
FAIL	PERSON	ETHNICITY_CONCEPT_ID	isForeignKey	Conformance	Relational	FIELD	None	100.00%
FAIL	PERSON	RACE_SOURCE_CONCEPT_ID	isForeignKey	Conformance	Relational	FIELD	None	100.00%
FAIL	PERSON	ETHNICITY_SOURCE_CONCEPT_ID	isForeignKey	Conformance	Relational	FIELD	None	100.00%
FAIL	VISIT_OCCURRENCE	VISIT_CONCEPT_ID	isForeignKey	Conformance	Relational	FIELD	None	100.00%
FAIL	CONDITION_OCCURRENCE	CONDITION_SOURCE_VALUE	sourceValueCompleteness	Completeness	None	FIELD	None	11.39%
FAIL	CONDITION_OCCURRENCE	CONDITION_CONCEPT_ID	isForeignKey	Conformance	Relational	FIELD	None	29.45%
FAIL	VISIT_OCCURRENCE	VISIT_END_DATE	measureValueCompleteness	Completeness	None	FIELD	None	6.62%
FAIL	CONDITION_OCCURRENCE	CONDITION_CONCEPT_ID	fkDomain	Conformance	Value	FIELD	None	5.74%
FAIL	DRUG_EXPOSURE	QUANTITY	plausibleValueHigh	Plausibility	Atemporal	FIELD	None	2.04%
FAIL	DRUG_EXPOSURE	DRUG_CONCEPT_ID	isForeignKey	Conformance	Relational	FIELD	None	1.89%
FAIL	VISIT_OCCURRENCE	DISCHARGE_TO_CONCEPT_ID	isForeignKey	Conformance	Relational	FIELD	None	0.59%
FAIL	DRUG_EXPOSURE	DRUG_CONCEPT_ID	fkDomain	Conformance	Value	FIELD	None	0.47%
FAIL	VISIT_DETAIL	VISIT_DETAIL_CONCEPT_ID	isForeignKey	Conformance	Relational	FIELD	None	0.02%
FAIL	VISIT_DETAIL	VISIT_DETAIL_CONCEPT_ID	standardConceptRecordCompleteness	Completeness	None	FIELD	None	0.02%
FAIL	CONDITION_OCCURRENCE	CONDITION_CONCEPT_ID	isStandardValidConcept	Conformance	Value	FIELD	None	0.01%
FAIL	DRUG_EXPOSURE	ROUTE_CONCEPT_ID	isStandardValidConcept	Conformance	Value	FIELD	None	2.80%
FAIL	DEATH	CAUSE_CONCEPT_ID	isForeignKey	Conformance	Relational	FIELD	None	27.79%
FAIL	VISIT_DETAIL	PROVIDER_ID	isForeignKey	Conformance	Relational	FIELD	None	4.34%

Tabla 3. Conceptos por persona. Muestra la distribución del número de conceptos registrados por persona en

5. Discusión

5.1. Transformación de la base de datos a OMOP CDM

La estandarización de estructura y contenido de una base de datos es clave para crear repositorios clínicos y evaluar la calidad de los datos con un vocabulario común[11]. Para la estandarización, se incluyeron en total: 92,073 hospitalizaciones, 1,226,472 evoluciones,

[11] Kahn MG et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). Sep 11;4(1):1244, 2016.

81,497 diagnósticos principales, 80,408 diagnósticos secundarios, 492.565 fármacos indicados y 6,262,449 resultados de laboratorio. En cuanto a la estandarización sintáctica, se mapearon 35 campos, mientras que para la estandarización semántica, se transformaron 5,993 conceptos. Existen 7,787 diagnósticos que no se lograron mapear, principalmente porque no se pudo interpretar el texto descriptivo, estos diagnósticos fueron cargados al CDM, ya que uno de los objetivos es que se vayan subsanando con el tiempo los fallos entregados por el DQD. No fue posible transformar los datos relativos a los procedimientos realizados, debido a que se encuentran registrados como texto libre.

La etapa de Perfilamiento de los datos es crucial, ya que se puede definir el punto de partida para definir qué campos mapear y si es posible solicitar que se entreguen archivos adicionales para completar los campos ausentes. Es importante mencionar que debido a la gran cantidad de datos del RCE del HHHA, el proceso se hizo más complejo, tal como menciona Alonso Carvajal en su estio [35] en donde menciona que a medida que el tamaño y la complejidad de los datos aumentan, puede ser necesario contar con una infraestructura más robusta. La elección de la infraestructura adecuada debe basarse en un análisis de los requisitos del proyecto y en la evaluación de los recursos disponibles.

5.2. Conceptos Estándar - Vocabulario Athena

Es recomendable estar familiarizado con la nomenclatura estandarizada y su estructura jerárquica, ya que el mapeo se va complejizando a medida que aumenta la cantidad de conceptos fuente. Si bien USAGI aporta una interfaz visual para la realización del mismo, el hecho de que la validación se realice manualmente, hace que el proceso sea más lento. Además, si bien Athena es una buena guía para buscar y definir los conceptos exactos a mapear, es muy importante revisar la documentación en cuanto a qué dominio de codificación estándar requiere cada campo, los primeros errores que presentó el DQD, ya que el entregado fue la segunda iteración, correspondían a mapeos realizados a dominios incorrectos.

5.3. Calidad de los Datos

En cuanto a la calidad de los datos, actualmente no existe un enfoque único para esta evaluación, ya que depende cada caso de uso, realizándose de forma ad hoc según cada proyecto[12]. En este estudio, seguimos el marco de Khan[9], diseñado para evaluar de manera integral grandes conjuntos de datos en redes de intercambio de datos clínicos. Este marco analiza la calidad intrínseca de los datos mediante el **DQD**, evaluando: **Completitud** (presencia de datos a lo largo del tiempo); **Conformancia** (cumplimiento de estándares); y **Plausibilidad** (coherencia lógica). Si bien los resultados del *Data Quality Dashboard* reflejan un cumplimiento global del 99%, lo que evidencia una adecuada estructuración y conformidad técnica con el modelo OMOP CDM, persisten importantes limitaciones asociadas a la completitud de los datos.

5.4. Errores de calidad y oportunidades de mejora

La ausencia de información en campos clínicos críticos, como por ejemplo, fechas de egreso, fechas de defunción, causas de defunción, o códigos estandarizados en exposiciones y condiciones, compromete la interpretación y confiabilidad de los análisis realizados. Esta falta de completitud no necesariamente se detecta mediante herramientas automatizadas de control estructural, pero sí afecta directamente la validez del conocimiento derivado de los reportes, gráficos y tablas generadas. En consecuencia, se destaca la necesidad de abordar la calidad del dato desde su origen, incorporando mecanismos de captura más robustos, validaciones al momento del ingreso y auditorías de completitud, especialmente si se pretende un uso secundario riguroso en investigación, vigilancia epidemiológica o inteligencia sanitaria.

Una de las ventajas de generar repositorios basados en modelos comunes es la escalabilidad de los datos, ya que una vez que se tengan los registros faltantes se pueden incorporar en el CDM OMOP realizando iteraciones con el fin de obtener una mejor calidad de datos progresivamente en el tiempo.

[12] Lewis AE, et al. Electronic health record data quality assessment and tools: a systematic review. J Am Med Inform Assoc. 2023 Sep 25;30(10):1730-1740.

5.5. Lecciones aprendidas hasta la fecha

Factor tiempo: El proceso de autorización del uso de los datos clínicos inició en junio de 2023, finalizando en abril de 2024. Los Comités de Ética del HHHA y SSASUR tienen la responsabilidad de proteger los derechos de los pacientes y deben velar por la seguridad de los datos clínicos, por lo que es probable que la demora sea principalmente por el tipo de proyecto que se está ejecutando, ya que involucra el acceso a un gran volumen de datos clínicos confidenciales.

Factor completitud: En el primer análisis, sólo se incluyeron indicaciones farmacológicas para el 57% y diagnósticos asociados para el 42% de las 35.771 hospitalizaciones. Tras una reunión explicativa y una nueva solicitud de datos, se modificó el periodo de análisis (junio 2020 - octubre 2024), excluyendo el periodo de implementación. Con las tablas corregidas, se lograron mayores porcentajes de completitud: diagnósticos principales en el 86% y fármacos asociados en el 37.4% de las 92.073 hospitalizaciones.

Factor trazabilidad: Ante la demora en la respuesta y la lenta comunicación con el Subdepartamento de Desarrollo TICs del HHHA, se identificó que la falta de comprensión sobre la utilidad de transformar el RCE al estándar OMOP-CDM y la alta carga laboral del equipo dificultaron el proceso. Se realizó una presentación explicando el problema, los beneficios de OMOP-OHDSI y las discrepancias. Como resultado, el equipo TICs realizó un nuevo envío de los datos directamente desde las tablas origen.

6. Conclusión

Se transformaron a vocabulario estándar el 100% de los datos, excepto por los diagnósticos intra-hospitalarios, cuyo mapeo fue del 95,2%. Esto se debió principalmente a la ausencia de un estándar local, que sirviera de referencia para el correcto proceso de transformación a estándares internacionales. En cuanto a la estandarización sintáctica, se estandarizó la estructura de 35 campos, asociados a seis dominios de OMOP CDM.

Los desafíos identificados en este estudio — como la falta de disponibilidad y estandarización, junto a las inconsistencias en los registros de fallecimientos, diagnósticos y fechas ausentes — revelan limitaciones que afectan directamente a dimensiones clave de calidad, como la completitud, conformidad y plausibilidad. Estas limitaciones no necesariamente se encuentran en el diseño y estructura de los RCEs, sino que también involucra al personal clínico, como médicos, enfermeras y otros profesionales de la salud, quienes son los responsables del registro de los datos. La precisión, coherencia y exhaustividad de estos registros son cruciales para garantizar la calidad de la atención.

A pesar de estas dificultades, este trabajo establece una base metodológica sólida para abordar dichas limitaciones en el contexto del HHHA, mejorando progresivamente la validez del repositorio creado, permitiendo por ejemplo, la carga progresiva de datos, para mejorar la calidad de la base actual. Este enfoque no solo permitirá elevar la calidad y confiabilidad de los datos transformados, sino que también potenciará su aplicación en estudios observacionales de alto impacto, como la farmacovigilancia, el modelamiento predictivo y la evaluación de calidad de atención.

Es importante mencionar la necesidad de implementar estándares y protocolos comunes en el marco de la Gobernanza de Datos Clínicos en Chile, para asegurar que los datos de salud se gestionen de manera ética, eficiente y segura. La implementación de un marco robusto de gobernanza es fundamental no solo para cumplir con las normativas legales, sino también para optimizar y facilitar el acceso y uso de los datos en la toma de decisiones clínicas, la investigación y las políticas públicas de salud.

El modelo de datos comunes OMOP CDM constituye una herramienta robusta para alcanzar la interoperabilidad, la integración y estandarización de datos clínicos, facilitando su reutilización en investigaciones colaborativas a nivel global. En conjunto, estos avances amplían significativamente el alcance y la relevancia de las investigaciones desarrolladas sobre la base de OMOP CDM en el país, brindando la posibilidad de escalar a otras instituciones y tipos de atención, y constituye un insumo valioso para impulsar políticas públicas de estandarización en Chile y Latinoamérica.

7. Perspectivas futuras

A medida que el ecosistema OMOP CDM continúa consolidándose, la red global de OHDSI ya agrupa más de 534 fuentes de datos en 49 países y ha estandarizado 928 millones de registros de 41 naciones, demostrando la escalabilidad y eficacia de los estudios federados sin intercambio de datos individuales [40]. En Europa, iniciativas como EHDEN y DARWIN EU utilizan herramientas de caracterización y el DQD para generar evidencia multinacional estandarizada, mientras que en Latinoamérica comienzan a surgir proyectos pioneros, por ejemplo, la primera cohorte DATASUS[13] de Brasil en formato OMOP y las discusiones del Latin America Workgroup de OHDSI sobre mapeo de vocabularios y variables socioeconómicas, que sientan las bases para redes regionales de datos interoperables. En Chile, la progresiva adopción del OMOP CDM en hospitales de alta complejidad abre la puerta a la creación de repositorios nacionales uniformes, capaces de soportar estudios multicéntricos sobre enfermedades prevalentes y de validar algoritmos predictivos en contextos locales.

8. Referencias

1. Ministerio de Salud de Chile (2020). Calidad de datos y sistemas de información en salud pública: Nota Técnica. www.minsal.cl
2. RAYEN Salud. (2022). Ficha Clínica Electrónica APS. www.rayensalud.cl
3. López, R. (2023). SSVQ implementa TrakCare de InterSystems. LinkedIn. www.linkedin.com/posts/rodrigolopez_ssvq-implementa-trakcare-de-intersystems-activity-7052675752449695745--WJl
4. UDD Ventures (2023). RedSalud e InterSystems sellan alianza para digitalizar la atención de salud privada en Chile. Extraído el 30 de mayo de 2025 desde: https://uddventures.udd.cl
5. SSMSO. (2022) Hospitales de la Red de Salud Sur Oriente implementan soluciones tecnológicas ante las necesidades de sus usuarios. Servicio de Salud Metropolitano Sur

[13] Oliveira, J.C.B. et al (2023). Data Standardization in Brazil: An OMOP Common Data Model Approach in a DATASUS Cohort. Value in Health, Volume 26, Issue 12, S539

Oriente. Extraído el 20 de junio de 2025 desde: https://redsalud.ssmso.cl/hospitales-de-la-red-de-salud-sur-oriente-implementan-soluciones-tecnologicas-ante-las-necesidades-de-sus-usuarios

6. Hospital Provincial de Ovalle (2025). ALMA: el nuevo sistema de ficha clínica digital que se instalará en el Hospital de Ovalle. Extraído el 25 de mayo de 2025 desde: https://hospitaldeovalle.cl

7. Hospital de La Serena (2025). El Hospital de La Serena se suma a la implementación regional del sistema de informatización hospitalaria. https://hospitaldelaserena.cl

8. Lobos Ossandón, V. (2020). Calidad de datos y sistemas de información en salud pública: Nota técnica. Comisión Nacional de Productividad. Extraído el 10 de abril de 2025 desde: https://cnep.cl/wp-content/uploads/2021/01/nota_tecnica_ti_2021-final.pdf

9. CNP & Circular-HR Fundación Chile (2022). Informe Final: Eficiencia en la Gestión de Atención Primaria en Salud. Extraído el 25 de junio de 2025 desde: https://cnep.cl/estudios-finalizados/eficiencia-en-gestion-de-atencion-primaria-de-la-salud

10. OHDSI (2022). OMOP CDM Background. Recuperado el 02 de enero de 2024, desde: https://ohdsi.github.io/CommonDataModel/background.html.

11. Ley N° 21.668. (2024). Ley de interoperabilidad de las fichas clínicas. Biblioteca del Congreso Nacional de Chile. Recuperado el 10 de abril de 2025 desde: www.bcn.cl/leychile/navegar?idNorma=1203827

12. Ministerio de Salud de Chile (2023). Estrategia de interoperabilidad en salud. Disponible en: https://interoperabilidad.minsal.cl

13. Decreto N°136 (2024). Reglamento Orgánico del Ministerio de Salud. Recuperado el 02 de enero de 2024. Biblioteca del Congreso Nacional de Chile. Extraído el 15 de mayo de 2025 desde: www.bcn.cl/leychile/navegar?idNorma=237230.

14. Ministerio de Salud de Chile (2023), Departamento de Estadística e Información en Salud. Norma Técnica N° 820: Estándares de Información en Salud. Publicado el 12 de enero de 2023.

15. International Organization for Standardization (ISO). Health informatics — Electronic health record — Definition, scope and context. ISO/TR 20514:2005(en). Recuperado el 16 de julio de 2023 desde: www.iso.org/obp/ui/#iso:std:iso:tr:20514:ed-1:v1:en.

16. Sarwar, T. et al (2022). The Secondary Use of Electronic Health. ACM Computing Surveys (CSUR), Volume 55, Issue 2.

17. Safran C. et al (2007). Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association, Volume 14, Issue 1, Pages 1-9, 2007.

18. Riikka V. et al (2017). “Impacts of structuring the electronic health record: Results of a systematic literature review from the perspective of secondary use of patient data”. International Journal of Medical Informatics, Volume 97, Pages 293-303.

19. Declerck J, et al (2024). Building a Foundation for High-Quality Health Data: Multihospital Case Study in Belgium. JMIR Med Inform. 2024 Dec 20;12:e60244. doi: 10.2196/60244. PMID: 39727158; PMCID: PMC11683741.

20. EHDEN. (2020). Implementing FAIR in OHDSI. The Hyve. Recuperado de www.thehyve.nl/articles/implementing-fair-in-ohdsi

21. Annika Jacobsen, et al (2020). FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2020; 2 (1-2): 10–29. doi: https://doi.org/10.1162/dint_r_00024

22. NIH (2020). Common Data Model Harmonization (CDMH) and Open Standards for Evidence Generation. US Food & Drug Administration and National Institutes of Health. Recuperado el 20 de abril de 2025 desde: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://aspe.hhs.gov/sites/default/files/private/pdf/259016/CDMH-Final-Report-14August2020.pdf

23. Zhang X, Wang L, Miao S, et al. Analysis of treatment pathways for three chronic diseases using OMOP CDM. J Med Syst. 42(12). pmid:30421323. 2018.

24. Garza, M. et al (2016). Evaluating common data models for use with a longitudinal community registry. Journal of Biomedical Informatics, Volume 64, Pages 333-341.

25. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek P. Increasing trust in real-world evidence through evaluation of observational data quality, Journal of the American Medical Informatics Association, Volume 28, Issue 10, October 2021, Pages 2251–2257. 2021

26. Observational Health Data Sciences and Informatics (OHDSI). (2023). The OHDSI Collaboration. Recuperado de www.ohdsi.org

27. Hripcsak, G., et al. (2015). Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Studies in Health Technology and Informatics, 216, 574–578. https://doi.org/10.3233/978-1-61499-564-7-574

28. Pacaci, A. et al (2018). A Semantic Transformation Methodology for the Secondary Use of Observational Healthcare Data in Postmarketing Safety Studies. Front Pharmacol. 2018 Apr 30;9:435.

29. Ward R et al (2024). The OMOP common data model in Australian primary care data: Building a quality research ready harmonised dataset. PLOS ONE 19(4): e0301557

30. OHDSI. (2023). WhiteRabbit and Rabbit-in-a-Hat: ETL Design Tools. Observational Health Data Sciences and Informatics. Recuperado de www.ohdsi.org/data-standardization/the-common-data-model/

31. Schuemie, M. J., & Rijnbeek, P. R. (2020). Usagi: Semi-automated mapping of source codes to standard concepts. OHDSI. Recuperado de https://ohdsi.github.io/Usagi/

32. Blacketer, C. et al. (2021). Increasing trust in real-world evidence through evaluation of observational data quality. Journal of the American Medical Informatics Association, 28(10), 2251–2261. https://doi.org/10.1093/jamia/ocab130

33. Abrahão R, et al. Challenges and opportunities in adopting OMOP-CDM in Brazilian hospital settings: a case report from Hospital Israelita Albert Einstein. OHDSI, 2023. Recuperado el 01 de julio del 2025 desde: www.ohdsi.org/wp-content/uploads/2023/10/12-Abrahao-BriefReport.pdf

34. OHDSI Latin America Working Group. Proyectos de armonización de datos clínicos a OMOP en Colombia y Argentina. Comunidad OHDSI LatAm, 2022–2023. (información interna de grupo de trabajo)

35. Carvajal A. Armonización y análisis de datos clínicos basado en OMOP CDM en Hospital Clínico Universidad de Chile. Tesis Magíster, Universidad de Chile, 2024. Disponible en: https://cimt.uchile.cl/wp-content/uploads/2024/03/TesisACarvajal.pdf

36. InterSystems. TrakCare Implementation in Chile's Western Metropolitan Health Service. Disponible en: www.intersystems.com/resources/intersystems-omop/

37. Hospital Dr. Hernán Henríquez Aravena. "Quienes Somos" recuperado el 10 de julio de 2023, desde: www.hhha.cl/?page_id=200.

38. Hospital Dr. Hernán Henríquez Aravena. "Cuenta Pública año 2022" Recuperado el 15 de julio de 2023 desde: www.hhha.cl/?page_id=972.

39. MII (2023). The Medical Informatics Initiative's core data set. Medical Informatics Initiative. Recuperado el 17 de febrero de 2025.

40. You SC, Lee S, Choi B, Park RW. Establishment of an International Evidence Sharing Network Through Common Data Model for Cardiovascular Research. Korean Circ J. 2022 Dec;52(12):853-864. doi: 10.4070/kcj.2022.0294. PMID: 36478647; PMCID: PMC9742390.

Anexos

Anexo N°1: Formulario de Solicitud de Dispensa de Proceso de Consentimiento Informado

FORMULARIO DE SOLICITUD DE DISPENSA DE PROCESO DE CONSENTIMIENTO INFORMADO

I. **Identificación de la Propuesta de Investigación:**

NOMBRE PROYECTO INVETIGACION:	Transformación y evaluación del Registro Médico Electrónico (HIS) del Hospital Dr. Hernán Henríquez Aravena (HHHA) de Temuco, utilizando el Modelo de Datos Común (CDM) de la Asociación para Resultados Médicos Observacionales (OMOP).
Número Protocolo (Código si corresponde):	
Fecha de la Versión:	14-11-2023
Patrocinador:	
Nombre Investigador:	Beatriz Mariane Estrada Sarabia
Lugar de Ejecución del Estudio:	Hospital Dr. Hernán Henríquez Aravena, Temuco

II. **Fundamentación de la Dispensa:**

Se solicita la dispensa del Documento de Consentimiento Informado considerando lo siguiente:

1.- **Fundamentación legal y ética:**

Si bien la ley chilena no contempla excepciones al consentimiento informado en su ley 20.584 y 20.120. El año 2015 la Comisión Asesora Ministerial de Investigación en Salud (CMEIS), elaboró algunas recomendaciones orientadas a que los comités que revisan estudios con fichas clínicas, en casos excepcionales puedan autorizar su uso sin el consentimiento del titular. Cuando solicitar el consentimiento informado para conocer la incidencia o prevalencia se requiera incluir a la mayoría o incluso a la totalidad de los pacientes afectados, pues del contrario la negativa de alguno de ellos generaría la invalidez de la investigación.

Es importante tener en consideración que la ley autoriza en ciertos casos y a determinadas entidades, a utilizar datos sensibles relativos a la salud sin previo consentimiento informado, cuando se tiene por finalidad proteger la salud de la población (por ejemplo, para una investigación epidemiológica) o existe un interés público (por ejemplo, mantener registros estadísticos).

A nivel internacional la declaración de Helsinki del año 2013 afirma: "Podrá haber situaciones excepcionales en las que será imposible o impracticable obtener el consentimiento para dicha investigación. Esta situación la investigación solo puede ser realizada de ser considerada y aprobada por un comité de ética de investigación".

Anexo N°2: Reporte White Rabbit

Egresos

G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
CCT_ESB_	Frequency	PCP_ID	Frequency	PCP_SEX_I	Frequency	PCP_PI	Frequency	PCP_FECH	Frequency	EGR_EGRE	Frequency	DIA_DIAGI	Frequenc	DIA_DIAGNOSTICO.DIA_DESCRI	Frequency	CEG_CONI	Frequency	DES_DEST	Frequency
51	92073	453679	29	2	59481	66	50092	13-07-196	35		13002		13002		13002	VIVO	76599	DOMICILO	75478
		848599	28	1	32590	65	19055	01-11-195	31	22-12-202	20	O800	2241	PARTO UNICO ESPONTANEO, PI	2241		13001	DERIV. OT	9108
		75387	26	List trunca		68	11315	06-08-199	30	24-05-202	17	I251	2061	ENFERMEDAD ATEROSCLEROTI	2061	FALLECIDO	2473		4953
		33572	25			67	10239	09-01-195	30	13-05-202	17	U071	2030	CASO CONFIRMADO CORONAVI	2029			DERIV. A C	762
		987579	25			64	420	29-08-198	29	18-10-202	16	O342	1938	ATENCION MATERNA POR CICA	1938			ALTA VOL	556
		675777	23			57	367	11-06-197	28	06-02-202	16	O244	1890	DIABETES MELLITUS QUE SE OR	1890			FUGA DEL	383
		124530	22			55	144	26-07-199	28	19-07-202	16	C509	1589	TUMOR MALIGNO DE LA MAMA	1589			DERIV. A II	373
		724189	21			58	60	24-11-196	27	23-02-202	16	O992	1582	ENFERMEDADES ENDOCRINAS,	1582			FALLECIDO	335
		26377	20			59	58	20-08-199	27	10-04-202	16	O363	1177	ATENCION MATERNA POR SIGN	1177			HOSPITALI	125
		1371636	19			63	57	15-07-195	26	12-09-202	16	O700	1122	DESGARRO PERINEAL DE PRIME	1122				
		55685	19			62	52	07-10-199	25	17-06-202	16	Z512	1086	OTRA QUIMIOTERAPIA	1086				
		418838	18			71	49	29-04-198	25	11-11-202	15	I678	984	OTRAS ENFERMEDADES CEREBF	984				
		795825	18			56	34	09-05-197	24	15-12-202	15	K358	933	OTRAS APENDICITIS AGUDAS Y I	933				
		105737	17			34	23	05-06-198	24	09-12-202	15	I219	902	INFARTO AGUDO DEL MIOCARD	902				
		232589	17				22	06-10-197	24	12-08-202	15	O335	793	ATENCION MATERNA POR DESP	793				
		376137	17			31	15	15-08-195	23	13-10-202	15	O021	775	ABORTO RETENIDO	775				
		237579	16			54	13	03-12-199	23	27-09-202	14	S720	753	FRACTURA DEL CUELLO DE FEM	753				
		506060	16			47	12	28-12-199	22	06-02-202	14	O064	748	ABORTO NO ESPECIFICADO, INC	748				
		144661	16			11	6	18-08-195	22	22-10-202	14	O429	745	RUPTURA PREMATURA DE LAS I	745				
		385804	16			70	5	23-08-194	22	01-03-202	14	M169	737	COXARTROSIS, NO ESPECIFICAD	737				
		616898	16			1	5	27-08-195	22	24-12-202	14	O069	689	ABORTO NO ESPECIFICADO, CO	689				
		427377	15			60	5	23-09-199	21	03-05-202	14	O600	642	TRABAJO DE PARTO PREMATUR	642				
		515962	15			List tru		16-09-198	21	07-07-202	14	O420	641	RUPTURA PREMATURA DE LAS I	641				
		1164626	15					01-02-194	21	29-10-202	14	C169	630	TUMOR MALIGNO DEL ESTOMA	628				
		279252	15					15-11-195	20	17-03-202	14	O009	578	EMBARAZO ECTOPICO, NO ESPE	578				
		305828	15					27-07-196	20	31-08-202	14	O809	577	PARTO UNICO ESPONTANEO, SI	577				
		277156	15					12-11-196	20	10-02-202	14	K810	575	COLECISTITIS AGUDA	575				
		318187	15					16-07-198	20	03-10-202	13	C189	509	TUMOR MALIGNO DEL COLON,	509				
		563245	15					18-09-194	20	24-08-202	13	S069	500	TRAUMATISMO INTRACRANEAL	500				
		456885	14					01-08-198	20	09-06-202	13	O13X	493	HIPERTENSION GESTACIONAL (I	493				
		81672	14					04-10-196	20	28-10-202	13	C20X	475	TUMOR MALIGNO DEL RECTO	475				
		397857	14					10-01-199	19	19-03-202	13	D291	474	TUMOR BENIGNO DE LA PROST	474				
		481300	14					02-08-198	19	16-09-202	13	O339	465	ATENCION MATERNA POR DESP	465				
		884743	14					27-09-199	19	13-09-202	13	E115	459	DIABETES MELLITUS NO INSULI	459				
		375420	14					08-03-198	19	28-05-202	13	O321	457	ATENCION MATERNA POR PRES	457				
		700187	14					30-11-195	19	29-03-202	13	O343	446	ATENCION MATERNA POR INCC	446				
		1588144	14					03-09-198	19	19-01-202	13	D259	439	LEIOMIOMA DEL UTERO, SIN OT	439				
		759978	13					02-03-195	19	15-07-202	13	O365	426	ATENCION MATERNA POR DEFI	426				
		902835	13					21-09-195	19	13-12-202	13	C859	420	LINFOMA NO HODGKIN, NO ESF	420				

Anexo N°3: Diccionario de datos

Nombre tabla OMOP V5	Nombre campo OMOP	Variable origen	Tipo de dato	Tabla origen RCE
person				
	person_id	PCP_ID	INTEGER (PK)	egresos
	gender_concept_id	SEXO_ID	INTEGER (FK)	egresos
	birth_datetime	FECHA_NACIMIENTO	DATETIME	egresos
	care_site_id	ESTABLECIMIENTO_ID	INTEGER	egresos
visit_occurrence				
	visit_ocurrence_id	CUENTA_CORRIENTE_ID	INTEGER (OK o FK)	egresos
	visit_start_datetime	INGRESO_FECHA	DATETIME	egresos
	visit_end_datetime	EGRESO_FECHA	DATETIME	egresos
	visit_detail_source_value	TAT_ID	VARCHAR	evoluciones
	discharged_to_source_value	DESTINO_EGRESO	VARCHAR	
	visit_detail_source_value	TAT_ID	VARCHAR	evoluciones
visit_detail				
	visit_detail_id	EVOLUCION_ID	INTEGER (PK)	evoluciones
	visit_detail_start_datetime	EVOLUCION_FECHA_CREA	DATETIME	evoluciones
	provider_source_value	ANA_TAN_ID	VARCHAR	evoluciones
	visit_detail_source_value	SER_DESCRIPCION	VARCHAR	evoluciones
condition_occurrence				
	condition_occurrence_id	DIAGNOSTICO_ID	INTEGER (PK)	diagnosticos
	condition_start_datetime	DIAGNOSTICO_FECHA_CREA	DATETIME	egresos/diagnosticos
	condition_source_value	DIAGNOSTICO_DESCRIPCION	VARCHAR	egresos/diagnosticos
	condition_type_concept_id	DIAGNOSTICO_TIPO	INTEGER (FK)	egresos/diagnosticos
	condition_status_concept_id	DIAGNOSTICO_SITUACION	INTEGER (FK)	egresos/diagnosticos
drug_exposure				
	drug_source_value	FARMACO_DESCRIPCION	VARCHAR	farmacos
	drug_exposure_start_datetime	FARMACO_FECHA_CREA	DATETIME	farmacos
	dose_unit_source_value	UNIDAD_APLICACION_FARMACO	VARCHAR	farmacos
	route_source_value	VIA_SUMINISTRO	VARCHAR	farmacos
	drug_exposure_end_datetime	FARMACO_FECHA_CREA	DATETIME	farmacos
	quantity	INDICACION_FARMACO_DOSIS	VARCHAR	farmacos
	sig	FARMACO_FRECUENCIA	VARCHAR	farmacos
	sig	FARMACO_FREC_UNIDAD	VARCHAR	farmacos
measurement				
	person_id	PCP_ID	INTEGER (FK)	laboratorio
	measurement_datetime	FECHA_VALIDACION	DATETIME	laboratorio
	value_as_nomber	RESULTADO	FLOAT	laboratorio
	measurement_source_value	DESCRIPCION_LIS	VARCHAR	laboratorio
	measurement_type_source_value	LIS	VARCHAR	laboratorio
death				
	death_datetime	FECHA EGRESO/EVOLUCION_FECHA_CREA	DATETIME	egresos
	death_type_soruce_value	EGRESO	VARCHAR	egresos
	cause_source_value	DIAGNOSTICO_DESCRIPCION	VARCHAR	egresos