Calidad de datos en el proceso ETL de entrega de medicamentos y asignación de citas médicas para las EPS

Universidad del Cauca

Trabajo de grado

Ing. Daisy Yisel Meneses Lopez

Director: PhD. Martha Eliana Mendoza Becerra

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Programa de Maestría en Computación
Grupo de Investigación y desarrollo en Tecnologías de la Información GTI
Áreas de Investigación: Inteligencia de negocios
Popayán, Diciembre de 2023

Calidad de datos en el proceso ETL de entrega de medicamentos y asignación de citas médicas para las EPS

Ing. Daisy Yisel Meneses Lopez

Trabajo de grado presentado a la Facultad de Ingeniería
Electrónica y Telecomunicaciones de la Universidad
del Cauca para la obtención del Título de:
Magíster en Computación

Director: PhD. Martha Eliana Mendoza Becerra

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Maestría en Computación
Grupo de Investigación y desarrollo en Tecnologías de la Información GTI
Áreas de Investigación: Inteligencia de negocios
Popayán, Diciembre de 2023

Tabla de Contenido

Lista de Figuras

LISTA DE TABLAS

Lista de Anexos

Anexo 1. Articulo "*Kahn's Data Quality Categories for Prescription delivery and Medical Appointment Assignment Reports*" publicado en la revista de la Universidad Pedagógica y Tecnológica de Colombia (UPTC), Vol. 32 No. 65 (2023): July-September 2023 (Continuous Publication).
Link publicación: https://revistas.uptc.edu.co/index.php/ingenieria/article/view/16314/1352. (Ver documento anexo).

Resumen

En el sector de la salud, los reportes de entrega de medicamentos y asignación de citas médicas generados por las Instituciones Prestadoras de Servicios de Salud para las Entidades Promotoras de Servicios de Salud presentan problemas de coherencia estructural, inconsistencias de formato, datos inexistentes, incompletos o no normalizados. Estos problemas afectan la calidad de los datos y dificultan la confiabilidad de la información.

Con el objetivo de abordar este problema, se propone adaptar las categorías de calidad de datos de Kahn a estos reportes, considerando su aceptación en el sector salud. Estas categorías no solo contemplan la estructura y dominio del dato, sino también la completitud y plausibilidad (credibilidad) del mismo.

Para llevar a cabo esta investigación, se siguieron dos enfoques metodológicos. En primer lugar, se realizó una revisión sistemática para identificar estudios relevantes sobre el tema. Posteriormente, se utilizó la metodología de Patrón de Investigación Iterativa de Pratt, donde se observaron los reportes de entrega de medicamentos y asignación de citas médicas, para comprender en detalle el problema y sus implicaciones y posteriormente se procedió a adaptar las categorías de calidad de datos propuestos por Kahn a estos reportes evaluándolas mediante la técnica de grupo focal. Los resultados según la percepción de los expertos demostraron que para el reporte de entrega de medicamentos la adaptación realizada obtuvo un 66.7% en “Completamente de Acuerdo” y 33.3% en “De Acuerdo”; y para asignación de citas médicas un 73.3% en “Completamente de Acuerdo” y un 26.7% en “De Acuerdo”, según la escala de Likert.

Por último, se incorporó la adaptación en un proceso ETL y se evaluó mediante el indicador de confianza. Los resultados obtenidos mostraron un alto nivel de confiabilidad en general, reflejando una consistencia significativa en la mayoría de las categorías y subcategorías evaluadas.

Agradecimientos

Culminar este viaje de aprendizaje y crecimiento ha sido un logro que no podría haber alcanzado sola. Durante este tiempo, he tenido el privilegio de conocer a personas excepcionales que han sido fundamentales en cada etapa de mi formación como Magister en Computación. Quiero expresar mi más profundo agradecimiento a cada uno de ustedes.

En primer lugar, agradezco a Dios por otorgarme la fortaleza, la sabiduría, la dedicación y el amor por lo que hago. A mi familia, mi pilar fundamental, por su apoyo incondicional. Sus consejos, amor y ánimo fueron el impulso que necesitaba en momentos desafiantes. A mi madre, Adíela López, quien ha sido mi modelo de vida y cuyas enseñanzas han forjado mi ser actual. A mi hijo, Thiago Garzón, mi fuente de inspiración, cuyo futuro se verá enriquecido por este logro. A mis hermanos, Diana y David por todo su apoyo incondicional, espero les sirva de ejemplo de que todo se puede lograr.

Agradezco de corazón a mi directora, Martha Eliana Mendoza Becerra, por guiar con sabiduría, paciencia y experiencia este proyecto de investigación.

Quiero dedicar un reconocimiento especial a la Universidad de Granada, España, y en particular al Dr. Salvador García López, por brindarme la oportunidad de llevar a cabo mi estancia de investigación. Sus consejos y críticas han sido fundamentales en mi crecimiento profesional y personal.

Finalmente, me gustaría expresar mi sincero agradecimiento a la Universidad de Cauca, una institución a la que me enorgullece pertenecer, por haberme brindado la invaluable oportunidad de crecimiento académico y profesional durante mi trayectoria como estudiante

Este logro es un tributo a todos ustedes, pero especialmente a mi familia, quienes son mi motor y mi mayor inspiración. Gracias por ser parte de este importante capítulo en mi vida. Con cariño y agradecimiento sincero,

Daisy Yisel Meneses López

Diciembre 2023

CAPITULO 1

1 INTRODUCCIÓN

En el ámbito de la salud, la calidad de los datos desempeña un papel crucial en la toma de decisiones informadas y respaldadas por información precisa y confiable. La correcta gestión de datos en el proceso ETL (Extracción, Transformación, Carga) es esencial para garantizar que los reportes normativos relacionados con la entrega de medicamentos y la asignación de citas médicas cumplan con los estándares más exigentes.

En este capítulo, se detalla el planteamiento del problema que motiva la presente investigación. Asimismo, se presenta la hipótesis principal y la hipótesis alternativa que servirán para el análisis. Además, se define tanto el objetivo general como los objetivos específicos definidos para abordar el problema. Finalmente, con el propósito de facilitar la comprensión y navegación del lector a lo largo de este documento, se proporciona una visión general de su estructura.

1.1 Planteamiento del Problema

En el ámbito de la salud, la generación masiva de datos es una realidad cotidiana, no obstante, la precisión y confiabilidad de estos datos se ha convertido en un desafío para su análisis, ya que la información de baja calidad puede llevar a conclusiones y decisiones erróneas. En su lugar, el uso de datos precisos y confiables es fundamental para la toma de decisiones informadas y la generación de valor en el sector de la salud [1].

Las Instituciones Prestadoras de Servicios de Salud (IPS), los proveedores de medicamentos y los prestadores independientes, también conocidos como prestadores de servicios de salud, proporcionan datos mensuales a las Entidades Promotoras de Servicios de Salud (EPS), los cuales en su mayoría se presentan en archivos de Excel. Sin embargo, en estos informes, y más concretamente en los relacionados con la entrega de medicamentos de acuerdo con la resolución 1604 de 2013 del Ministerio de Salud y Protección Social (MSPS) [2] y la asignación de citas médicas según la resolución 1552 de 2013 del MSPS [3], se encuentran problemas frecuentes como: (i) atributos que carecen de una estructura coherente e inconsistencia en el formato; (ii) código único de medicamentos

incompletos, inexistentes o registrados con los códigos de Clasificación anatómica Terapéutica (ATC), lo que dificulta la identificación de la presentación comercial del medicamento entregado al afiliado; (iii) Clasificación internacional de enfermedades (CIE10) inexistentes o incompletos; (iv) presencia de información sobre afiliados que no se encuentra en la base de datos o que pertenecen a otras EPS; (v) errores y discrepancias en los campos que indican cantidades prescritas, entregadas, días de tratamiento, fechas de radicación y fechas de entrega, como cantidades entregadas superiores a las formuladas o fechas de entrega anteriores a la fecha de formulación [4].

Una bodega de datos (DW, por sus siglas en inglés) es un sistema de gestión de datos que centraliza y consolida grandes cantidades de información de distintas fuentes, con el fin de permitir a las empresas organizar, comprender y utilizar los datos para la toma de decisiones [5]. Sin embargo, los datos provenientes de múltiples fuentes pueden ser inconsistentes y de mala calidad [6].

El proceso de Extracción, Transformación y Carga (ETL) que forma parte de las bodegas de datos, se encarga de eliminar y corregir errores causados por datos faltantes [7] por medio de tres etapas: (i) extracción de datos a partir de fuentes existentes, (ii) transformación de los datos en un formato común, y (iii) carga de los datos en una DW [8].

En la revisión del estado del arte, se identificaron varios estudios relacionados con la calidad de datos. Uno de ellos enfocado en la estandarización de la estructura de los datos para detectar y corregir errores en las variables definidas [9]. Otro analiza las discrepancias entre los datos de origen y destino utilizando tres categorías: integridad, coherencia y validez sintáctica [10]. Por último, un enfoque de limpieza de datos que abarca la evaluación de datos, metadatos, valores atípicos y duplicados, seguido de la detección de anomalías y un seguimiento de las inconsistencias para estandarizar los datos [11]. Por otro lado, Kahn [12] propone tres categorías de calidad de datos: conformidad, completitud y plausibilidad, que evalúan aspectos como restricciones sintácticas o estructurales, la presencia o ausencia de datos en diferentes momentos y la credibilidad o veracidad de los valores. Los estudios [13],[14][15] han abordado la calidad de datos en el sector de la salud, centrándose en la estructura y el dominio de los datos, así como en la estandarización y la detección de anomalías para corregir errores en las variables. Sin embargo [13] no considera la categoría

de plausibilidad planteadas por el mismo autor, que es fundamental para garantizar la veracidad de los datos. Cabe destacar que estas categorías han sido ampliamente utilizadas y aceptadas en el sector de la salud en otros países [13],[14][15].

Una alternativa para mejorar la calidad de los datos en el proceso ETL es primero adaptar las categorías completitud, conformidad y plausibilidad a las características específicas de los reportes de entrega de medicamentos y asignación de citas médicas, buscando detectar errores y una mayor calidad en los datos que permita la realización de análisis por parte de las EPS; segundo incorporar las categorías de calidad de datos adaptadas en un proceso de ETL; por último, evaluar la calidad de los datos al aplicar el proceso ETL a través de los indicadores confianza y soporte propuestos por N. Duque [9].

Por todo lo anterior y teniendo en cuenta que en la literatura actual no se evidencian estudios aplicados en el sistema de salud Colombiano sobre calidad de datos en el proceso ETL, surge la siguiente pregunta de investigación ¿Qué tipo de adaptación se debe realizar a las categorías del Framework propuesto por Kahn para verificar y validar los datos de los reportes de medicamentos y citas médicas reportados por las Instituciones prestadoras de servicios de salud (IPS) a las Entidades promotoras de servicios de salud (EPS)?.

De acuerdo con la pregunta anterior, se plantea la hipótesis principal de tipo causal para esta investigación, así como la hipótesis alternativa.

1.2 Hipótesis Principal

Mediante la adaptación de las categorías de calidad de datos planteadas por Kahn, incorporadas en el proceso ETL de las Entidades Promotoras de Servicios de Salud, se mejora la calidad de los datos de entrega de medicamentos y asignación de cita médicas reportados por las Instituciones Prestadoras de Servicios de Salud.

1.3 Hipótesis Alternativa

Mediante la adaptación de las categorías de calidad de datos planteadas por Kahn, incorporadas en el proceso ETL de las Entidades Promotoras de Servicios de Salud, NO se

mejora la calidad de los datos de entrega de medicamentos y asignación de cita médicas reportados por las Instituciones Prestadoras de Servicios de Salud.

1.4 Objetivo General

Proponer un proceso de extracción, transformación y carga para las Entidades Promotoras de Servicios de Salud Colombianas, que incorpore las categorías de calidad de Kahn adaptadas a los datos de los reportes de entrega de medicamentos y asignación de citas médicas generados por las Instituciones Prestadoras de Servicios de Salud.

1.5 Objetivos Específicos

1. Adaptar las categorías de calidad de datos propuesto por Kahn, para verificar y validar los reportes de entrega de medicamentos y asignación de citas médicas generados por las Instituciones Prestadoras de Servicios de Salud, mediante la definición de criterios específicos para cada característica de las categorías.

2. Incorporar las categorías de calidad de datos adaptadas en un proceso de extracción, transformación y carga, para Entidades Promotoras de Servicios de Salud, buscando que puedan acceder a datos más confiables para futuros análisis y reportes a los entes territoriales.

3. Evaluar la calidad de los datos al aplicar el proceso ETL propuesto usando los archivos consolidados de entrega de medicamentos y asignación de citas médicas facilitados por Asmet Salud EPS SAS, a través de los indicadores confianza y soporte propuestos por N. Duque [9].

1.6 Estructura del documento

Este documento está estructurado de la siguiente manera:

En el Capítulo 1 se plantea el problema, se establece la hipótesis principal y alternativa, y se define el objetivo general y los específicos para llevar a cabo esta investigación. Además, se presenta la estructura del documento para guiar al lector a través de los componentes de este.

El Capítulo 2 proporciona la base conceptual necesaria para comprender el proceso ETL y la importancia de la calidad de los datos en entornos de la salud. Se exploran conceptos como la bodega de datos, el proceso ETL, la calidad de datos y la evaluación de la calidad de datos mediante indicadores. Además, se revisa el estado actual del campo objeto de esta investigación.

En el Capítulo 3 se describe en detalle la metodología utilizada en esta investigación, que se divide en dos ciclos: la adaptación de las categorías de calidad, y la incorporación y evaluación de la calidad de datos en el proceso ETL. Cada ciclo se desglosa en cuatro etapas, que van desde la observación hasta la prueba de la solución.

El Capítulo 4 se presenta en detalle la adaptación realizada a las categorías de calidad de datos de Kahn (Conformidad, Completitud y Plausibilidad) a los reportes de entrega de medicamentos y la asignación de citas médicas.

El Capítulo 5 aborda la incorporación de las categorías de calidad de datos de Kahn adaptadas en el proceso ETL, detallando la arquitectura y los paquetes implementados por cada categoría, utilizando la herramienta Pentaho.

En el Capítulo 6, se explica la validación realizada a la adaptación propuesta por parte de un grupo de expertos y se evalúa la calidad de los datos en el proceso ETL desarrollado para los reportes de entrega de medicamentos y la asignación de citas médicas.

En los Capítulo 7 y 8, se presentan las conclusiones derivadas de esta investigación y se plantean posibles líneas de trabajo futuro.

El documento también contempla la lista de referencias bibliográficas en el Capítulo 8 y anexos relevantes en el Capítulo 9.

CAPITULO 2

2 MARCO TEÓRICO Y ESTADO DEL ARTE

En esta sección se presentan los conceptos necesarios para el desarrollo de esta tesis de maestría y una revisión sistemática en el tema de investigación.

2.1 Marco Teórico

Detalla los conceptos esenciales que sientan las bases para comprender y abordar esta investigación, incluyendo la definición de una bodega de datos, el proceso de extracción, transformación y carga, así como la calidad de datos y su evaluación mediante indicadores.

2.1.1 Bodega de datos

Una bodega de datos (DW, por sus siglas en inglés), se define como un depósito centralizado diseñado para llevar a cabo el almacenamiento, recuperación, análisis, consulta y visualización de datos de manera eficiente [7]; cuyas características son [16]:

- **Integrada**: busca eliminar las inconsistencias en los sistemas operacionales para que los datos se puedan integrar en una estructura homogenizada.

- **Temática**: todos los datos se organizan por temas, así los usuarios finales pueden acceder y entender fácilmente la información almacenada, la cual está debidamente categorizada.

- **Histórica**: los datos almacenados en una DW guardan sus registros temporales, con lo cual se puede acceder a información histórica permitiendo realizar comparaciones.

- **No volátil**: la información ingresada en las DW es permanente, puede ser consultada, pero no eliminada, lo que respalda la fiabilidad de los datos almacenados.

2.1.2 Proceso ETL

El proceso de extracción, transformación y carga (ETL, por sus siglas en inglés) es el conjunto de acciones que permiten extraer datos para integrarlos en un solo lugar [7] y está compuesto por tres fases que son:

- **Extracción**. Es la encargada de leer los datos desde múltiples fuentes (ERP, Excel, Open Data, IOT, entre otros).

- **Transformación**. Los datos extraídos, se someten a una serie de transformaciones como: limpieza de datos, conversión de formatos, enriquecimiento con datos adicionales, agregación, filtrado y otros procesos para que los datos sean coherentes, estandarizados y listos para el análisis.

- **Carga**. Los datos transformados se cargan en una fuente de destino, que puede ser una base de datos, una DW, entre otras; los cuales estarán disponibles para herramientas como la minería de datos o el procesamiento analítico en línea (OLAP, por sus siglas en inglés).

2.1.3 Calidad de datos

La calidad de datos (DQ, por sus siglas en inglés) se refiere a la medida en que los datos han pasado por un proceso como el definido en la fase de transformación del proceso ETL, con el propósito de garantizar que los datos sean coherentes, estandarizados y estén preparados de manera óptima para su posterior análisis [17].

Una forma de evaluar la DQ es la planteada por Kahn [12], la cual plantea tres categorías compuestas de subcategorías, así:

- **Conformidad.** Evalúa si los valores cumplen con las restricciones sintácticas o estructurales, y se divide en tres subcategorías que son:

  - **Conformidad de valor.** Identifica si los elementos de datos registrados están de acuerdo con una arquitectura de datos preespecificada y basada en restricciones como: de formato internas, valores o rangos permitidos, o representaciones basadas en estándares externos. Ej. el sexo solo debe tener los valores "M" o "F".

  - **Conformidad relacional.** Se enfoca en determinar si los valores de los datos cumplen con las restricciones relacionales establecidas, ya sea por estándares internos de la organización o por externos. Ej. el número de historia clínica del paciente se vincula a otras tablas según sea necesario.

- **Conformidad computacional.** Se centra en la exactitud del valor de salida de los cálculos frente a las especificaciones técnicas funcionales. Ej. índice de masa corporal (IMC) calculados son idénticos con una fuente externa.

- **Completitud.** Evalúa la ausencia de datos en un solo o en múltiples momentos a lo largo del tiempo. Ej. los códigos ICD-9CM coincide con la implementación de ICD-10CM.

- **Plausibilidad.** Describe la credibilidad o veracidad de los valores de los datos, y se divide en tres subcategorías que son:

  - **Plausibilidad de unicidad.** Verifica si los objetos (entidades, observaciones, hechos) no se duplican o no pueden distinguirse dentro de una base de datos o cuando se comparan con una referencia externa Ej. los pacientes de una sola institución no tienen varios números de historia clínica.

  - **La plausibilidad atemporal.** Establece si los valores, distribuciones o densidades de datos observados concuerdan con el conocimiento local o "común" o de comparaciones con fuentes externas Ej. la temperatura oral y axilar son similares.

  - **La plausibilidad temporal.** Comprueba si los valores observados o derivados tienen propiedades temporales similares en uno o más comparadores externos. Ej. la fecha de admisión es anterior a la fecha de alta.

Todas las categorías de DQ planteadas por Kahn son evaluadas mediante dos estrategias que son:

- **La verificación.** Los valores de los datos coinciden con las expectativas con respecto a las limitaciones de los metadatos, los supuestos del sistema y el conocimiento local.

- **La validación.** Alineación de los valores de los datos con respecto a los puntos de referencia externos relevantes.

2.1.4 Evaluación de la calidad de datos por medio de indicadores

Un indicador de gran relevancia en la evaluación de la DQ durante el proceso de ETL es el indicador de Confianza [9], que mide la proporción de registros que ingresan sin errores en comparación con el total de registros entrantes. Su cálculo se realiza de manera individual para cada atributo relevante en la detección de errores y se expresa en forma de porcentaje (ver Fórmula 1), lo que facilita la interpretación de la DQ.

$$\boldsymbol{Confianza} = \left(\frac{Registros\ sin\ errores}{Total\ de\ registros\ entrantes}\right) x\ 100$$

Fórmula 1 Indicador de Confianza

En términos de interpretación, un alto valor de Confianza indica que la mayoría de los registros en una variable son precisos y confiables; sugiriendo una alta calidad en la recopilación, entrada y procesamiento de datos. Por otro lado, un valor bajo de Confianza puede indicar que se requieren mejoras en estos procesos, lo que implica la necesidad de abordar posibles problemas en la DQ. El indicador de Confianza es una herramienta para evaluar y mejorar la calidad de los datos en una organización.

2.2 Estado del Arte

Se realizó una revisión sistemática siguiendo los modelos propuestos por [18] y [19], que consta de tres etapas: Planificación, Ejecución y Análisis de Resultados.

2.2.1 Etapa de Planificación

En esta etapa se definieron aspectos como: (i) las preguntas de investigación, (ii) la estrategia de búsqueda y (iii) los criterios de inclusión/exclusión.

a) Definición de las preguntas de investigación

Para obtener un conocimiento más detallado del tema que se aborda en este proyecto con respecto a las medidas de DQ en procesos ETL, se plantearon dos preguntas de investigación, las cuales se presentan en la Tabla 1, junto con su motivación.

Tabla 1 Preguntas de investigación

Preguntas de investigación	Motivación
P01 ¿Qué tipos de estrategias existen en el proceso ETL para mejorar la DQ a cargar en las bodegas de datos de salud?	Entender que métodos, modelos o marcos de trabajos de BI han sido utilizados para la DQ durante el proceso ETL para el sector salud.

Preguntas de investigación	Motivación
P02 ¿En qué países fueron desarrollados los estudios de ETL en el sector salud?	Conocer el estado y manejo de la información del sector salud en el mundo.

Fuente: Elaboración propia

b) Estrategia de búsqueda

Identificando las palabras clave para esta investigación, se utilizaron los conectores lógicos "AND" y "OR", para obtener la cadena de búsqueda definitiva (ver Tabla 2). Además, se seleccionaron las bases de datos para aplicar dicha cadena como: Scopus, Science Direct y IEEE Xplore.

Tabla 2 Cadena de búsqueda

("data warehouse" OR "data warehousing" OR "business intelligence") **AND** ETL **AND** health

c) Criterios inclusión/exclusión

Para evaluar la información encontrada en las bases de datos, se definió que los estudios debían cumplir con los siguientes criterios: (i) escritos inglés o español y (ii) publicados entre los años 2015 y 2023. De estos estudios se revisó el título, resumen y palabras claves, descartando los que cumplieron al menos uno de los siguientes criterios: (i) no presentan información de medidas de calidad de datos, (ii) muestran información de medidas de calidad de datos de forma general, (iii) resúmenes u opiniones, (iv) duplicados encontrados en dos o más base de datos, (v) comparaciones de herramientas.

2.2.2 Etapa de ejecución

En esta etapa se ejecutó la cadena de búsqueda definida, obteniendo como resultado la cantidad de estudios por cada base de datos que se relaciona en la Tabla 3**.** Al aplicar los criterios de inclusión y exclusión definidos, quedaron seleccionados seis estudios primarios con los cuales se respondieron las preguntas de investigación.

Tabla 3 Resultados búsqueda

Cadena de búsqueda	Bases de Datos			Estudios primarios seleccionados
	IEEE	SCIENCE DIRECT	SCOPUS	
("data warehouse" OR "data warehousing" OR "business intelligence") AND ETL AND health	16	6	27	6

Fuente: Elaboración propia

2.2.3 Análisis de resultados

Con la revisión de los estudios seleccionados, se obtiene respuesta a las preguntas de investigación planteadas y luego se presenta el análisis realizado con estos resultados.

- **Respuesta a las preguntas de investigación**

P01. ¿Qué tipos de estrategias existen en el proceso ETL, para mejorar la DQ a cargar en un DW de salud?

En el año 2016 se propone un modelo de optimización para el proceso ETL en un repositorio de datos consta de tres fases [9]: *(i) Fase de Prerrequisitos.* Se realiza el proceso de "traducción" para estandarizar la estructura de los datos, asegurándose de que los datos estén en un formato coherente y uniforme. *(ii) Fase Principal*. Se divide en dos tareas**:** *(a) filtrado detectivo*: detecta errores atípicos e inconsistencias en los datos, con la ayuda de expertos en el tema que identifican posibles problemas en estos. *(b) Filtrado correctivo*: se registran en una tabla de variables los problemas encontrados y se organizan para realizar las correcciones correspondientes; luego se realiza un proceso de migración de datos, en el cual, se cargan los datos corregidos en la DW para su posterior uso. *(iii) Fase Alternativa.* Se llevan a cabo tres actividades: *(a) Historial de errores*: se almacenan las descripciones de los errores encontrados y las correcciones realizadas. *(b) Administración de variables***:** gestiona los filtros utilizados en el proceso de ETL, permitiendo crear nuevos filtros, consultar los existentes, actualizar y eliminar los que no sean relevantes. (*c) Cálculo de calidad del proceso*: se evalúa la calidad del proceso de ETL mediante el uso de indicadores de confianza y soporte.

El enfoque propuesto en 2018 para validar los procesos de ETL utilizando pruebas de equilibrio automatizadas [10] consta de cinco fases principales: *(i) Definición de propiedades genéricas*. Establece las propiedades de integridad, coherencia y validez sintáctica que deben cumplir los datos tanto en el origen como en el destino. *(ii) Identificación de asignaciones de origen a destino*. Se definen las reglas de transformación que mapean los datos del origen a los datos del destino. *(iii) Asignaciones de pruebas*. Se realizan pruebas para verificar la coincidencia del recuento de registros desde el origen hasta el destino, comprobando si los datos extraídos del origen y transformados de acuerdo con las reglas establecidas coinciden con los datos cargados en el destino. *(iv) Evaluación del enfoque*. Se valida la efectividad del enfoque propuesto, por medio de scripts para confirmar que el

proceso de ETL cumple con las propiedades definidas y que se mantienen las asignaciones de origen a destino*. (v) Prueba de mutación automatizada*. Se selecciona aleatoriamente un registro de la tabla de destino y se muta el valor de un atributo específico, para detectar posibles fallas en el proceso de ETL al verificar cómo responde el sistema ante cambios inesperados en los datos.

Más adelante, en el año 2019, se propone el desarrollo de un proceso de garantía de la calidad [13], enfocado en las categorías de Conformidad y Completitud de Kahn, aplicado en todas las etapas del proceso de ETL por medio de un proceso iterativo hasta obtener resultados de calidad aceptables. *(i) Extracción*. Se realiza una copia de las dimensiones y la tabla de hechos de la fuente actual, se descargan los metadatos actualizados del Asociación para la Observación de los Resultados Médicos (OMOP, por sus siglas en inglés) y se aplican criterios de exclusión para eliminar duplicaciones de filas y rangos de fechas no válidos, luego, en la categoría de completitud se compara fila por fila para validar que los criterios de inclusión y exclusión se apliquen correctamente en todas las dimensiones y tablas de hechos. *(ii) Transformación*. Se identifican los registros creados, actualizados o eliminados utilizando una combinación de clave primaria y el ID de procesamiento de autoincremento del ETL, se cuentan las adiciones y actualizaciones en el Modelo de Datos de OMOP (MDL, por sus siglas en inglés), que incluye conceptos o relaciones nuevos o modificados; se aplica la categoría de Conformidad relacional para garantizar que se realicen actualizaciones y eliminaciones en cascada en todas las tablas relevantes del modelo OMOP. Además, se realizan comprobaciones para evitar que existan identificadores huérfanos en las tablas de hechos, es decir, que apunten a campos inexistentes. *(iii) Carga*. Los datos transformados se asignan a los metadatos de OMOP y se cargan en las tablas de dominio correspondientes, aplicando la Conformidad de valor, que se refiere al grado en que los valores de los datos transformados se ajustan a las restricciones definidas en el MDL de OMOP.

En el mismo año, se propone una evaluación de calidad de datos a través del proceso de curación de datos que consiste en un conjunto de actividades analíticas y de consulta para evaluar la calidad de los datos [15], basándose en las categorías de calidad de datos de Kahn y consta de cinco pasos: *(i) Respuesta a una consulta de diagnóstico,* en este paso el socio de la red respondía a una consulta de diagnóstico, que evaluaba la conformidad con

el MDL a nivel de tabla. *(ii) Actualización del diccionario de datos*, en este paso, los socios de la red completan un diccionario de datos que recopila información técnica sobre los DataMarts, el estado de implementación de tablas y campos, y otros detalles relacionados con la ETL. Luego, ejecutan una consulta de caracterización de datos que genera frecuencias, estadísticas descriptivas y tabulaciones cruzadas. *(iii) Informe de caracterización de datos empíricos*, a partir de los resultados de la consulta de caracterización de datos, los analistas del Centro de Coordinación crean un informe resumido llamado "Informe de caracterización de datos empíricos" y lo distribuyen al socio de la red. *(iv) Revisión del resultado de la consulta de caracterización de datos,* a partir de los resultados de la consulta de caracterización de datos, los analistas del Centro de Coordinación crean un informe resumido llamado "Informe de caracterización de datos empíricos" y lo distribuyen al socio de la red. *(v) Comentarios y oportunidades de mejora*: Si se identifican oportunidades de mejora en los datos, se documentan en un plan de mitigación. Esta evaluación fue aplicada a seis dominios: datos demográficos, diagnósticos, encuentros, inscripción, procedimientos y signos vitales.

En el mismo año, se propone un framework automatizado para la curación de datos por medio de una arquitectura compuesta de tres módulos [11]*: (i) Evaluación de datos*. Extrae los metadatos del conjunto de datos (como información sobre las variables y su tipo de datos) y calcula estadísticas descriptivas para comprender la distribución de los datos; *(ii) Control de calidad de datos*. Se detectan los valores faltantes en el conjunto de datos ayudando a identificar posibles problemas de integridad; se utilizan métodos univariados y multivariados para detectar anomalías en los datos, identificando valores atípicos o inesperados; se utilizan métricas de similitud y técnicas de detección de duplicados para identificar características similares o registros duplicados en los datos; *(iii) Estandarización de datos.* Se centra en garantizar la coincidencia del conjunto de datos sin procesar con los modelos de referencia o estándares establecidos.

Más recientemente, en el año 2020 [14] propone una arquitectura inspirada en el enfoque de Kimball para garantizar la calidad de los datos en el flujo de trabajo ETL, compuesta por cinco componentes principales que son: *(i) Auditoría de calidad de datos*, mide la calidad de los datos durante el flujo de trabajo ETL mediante filtros de diagnóstico que generan eventos de error. Estos eventos se almacenan en una tabla de hechos de un esquema en estrella

vinculados a una dimensión de auditoría para facilitar el análisis de eventos. (ii) *Sistema de auditoría,* evalúa el conjunto de eventos generados y proporciona métodos para supervisar los problemas de calidad de los datos por medio de consultas SQL. (iii) *Registro de eventos de error*, proporciona una API para que los desarrolladores del proceso ETL registren problemas de calidad de datos en el almacén de eventos. (iv) *Almacén de eventos,* contiene la información detallada sobre problemas de calidad. (v) *Sistema de control,* Proporciona una interfaz de usuario donde se visualizan y analizan los resultados de las métricas de DQ aplicadas*.*

P02 ¿En qué países fueron desarrollados los estudios de ETL en el sector salud?

Los estudios encontrados fueron desarrollados en Estados unidos [13], Alemania [14], Italia [10] y Colombia [9], este último es aplicado en caso de estudio de datos ambientales.

- **Análisis de los resultados obtenidos**

Se observa que los estudios [9], [10] y [11] abordaron la DQ en el sector de la salud, centrándose en aspectos como la estructura y el dominio de los datos, así como en la estandarización y la detección de anomalías para corregir errores en las variables.

Kahn plantea las categorías de calidad de datos de *Conformidad, completitud y Plausibilidad* que son ampliamente aceptadas en el sector salud porque involucraron la participación de aproximadamente cien profesionales de diversas disciplinas a nivel internacional, quienes colaboraron en el desarrollo y revisión de la terminología de calidad de datos armonizada. Además, se basaron en un conjunto de datos con más de 540 millones de registros de pacientes, respaldando la solidez y relevancia de estas categorías en el ámbito de la salud.

Los estudios [13], [14] y [15] aplicaron las categorías planteadas por Khan para evaluar la DQ. Sin embargo, es importante señalar que [13] no incorporó la categoría de *Plausibilidad.* Esta categoría es importante porque se encarga de verificar la coherencia y la lógica de los datos dentro del contexto médico, asegurando así su veracidad.

CAPITULO 3

3 METODOLOGÍA

Esta investigación utilizó la metodología de Patrón de Investigación Iterativa (PII, por sus siglas en inglés) propuesto por Pratt [20], la cual consta de cuatro etapas: observación, identificación del problema, desarrollo de la solución y prueba de la solución.

3.1 Ciclo 1 Adaptación de las categorías de calidad

En este ciclo, se da inicio con el estudio de las DQ planteadas por Kahn y de la adaptación de estas que proponen los estudios primarios seleccionados. Posteriormente, se procede a la identificación de los problemas que se presentan en los reportes de asignación de citas médicas y entrega de medicamentos generados por las IPS. Luego, se enfoca en la adaptación de las DQ a las particulares de estos reportes y finalmente, se lleva a cabo la validación de esta adaptación por medio de un grupo focal de expertos.

3.1.1 Observación

En esta etapa se estudió la adaptación de las categorías DQ de Kahn que presentaban los estudios primarios seleccionados [13],[15],[14], encontrando en algunos de estos estudios lo siguiente:

- Uno de ellos se centraba en la categoría de *Conformidad* y *Completitud*, sin tener en cuenta la categoría de *Plausibilidad*, mientras que los otros dos abordaban todas las categorías, pero en un orden diferente al sugerido por el autor.
- Se enfocaban en mostrar los resultados de la aplicación de estas categorías en lugar de describir en detalle cómo las aplicaban.

Además, se estudiaron cada una de las categorías DQ planteadas por Kahn y sus respectivas subcategorías, para comprender el objetivo de cada una de ellas y así poder definir su aplicación en el contexto específico de los reportes de entrega de medicamentos y asignación de citas médicas. En esta actividad se identificó lo siguiente:

- **Subcategoría Conformidad de valor.** Permite verificar que los valores de los atributos se adecuan a dos tipos de restricciones, por un lado, las restricciones de formato internas como: (i) *Tipos de datos*, se refiere al tipo de datos que se espera en un campo,

como texto, número, fecha, etc. (ii) *Formatos de datos*, indica el formato específico que debe seguir un dato, como un número de teléfono con guiones o una fecha en el formato "dd/mm/aaaa". (iii) *Longitud*, hace referencia a la longitud o cantidad de caracteres permitidos en un campo de datos. (iv) *Valores o rangos permitidos,* establece los valores o intervalos aceptables para un campo, como los valores válidos para un campo de "sexo" ("M" o "F").

- **Subcategoría Conformidad relacional.** Ayuda a determinar en los reportes las restricciones de obligatoriedad y las restricciones de integridad referencial como: (i) *Clave foránea*, se refiere a la verificación de que los valores en un campo coincidan con las claves foráneas de otras tablas. (ii) *Clave primaria,* verifica si los valores en un campo son únicos en una tabla.

- **Subcategoría Conformidad Computacional.** En los registros de los reportes no se encontraron datos derivados computacionalmente, por lo tanto, esta subcategoría no se consideró en la adaptación.

- **Categoría Completitud.** Se encarga de garantizar que los atributos que presentan ausencia de datos cumplan con definiciones específicas establecidas para cada uno de ellos. De esta forma, en un conjunto de datos, se verifica que no haya valores faltantes en los atributos y que, en caso de que existan, estos sean consistentes con las definiciones y requisitos particulares de cada atributo.

- **Subcategoría Plausibilidad de Unicidad.** Se encarga de identificar y evitar la duplicación de atributos que están destinados a identificar de manera única un objeto en un conjunto de datos.

- **Subcategoría Plausibilidad Atemporal.** Se enfoca en garantizar que los valores y distribuciones de datos, así como las restricciones lógicas entre estos valores o mediciones repetidas, sean coherentes con el conocimiento interno o información proveniente de fuentes externas. Su principal función es verificar que los datos no solo sean precisos, sino también que estén en línea con las expectativas lógicas y el contexto específico en el que se utilizan.

- **Subcategoría de Plausibilidad Temporal.** Garantiza que los valores observados o derivados, así como las secuencias de valores o las mediciones de densidad de valor, sean consistentes, coherentes y se ajusten a las propiedades temporales.

3.1.2 Identificación del Problema.

Se revisó el consolidado de los reportes generados por las IPS, que se basan en la estructura establecida por el Ministerio de Salud y Protección Social (MSPS) mediante la Resolución 1604 de 2013 [2], la cual define los requisitos para la entrega de medicamentos. Asimismo, se consideró la Resolución 1552 de 2013 [3] que establece los lineamientos para los reportes de asignación de citas médicas. Estas resoluciones establecen los estándares y procedimientos que deben seguir las IPS al generar los reportes que envían a las EPS, las cuales realizan la consolidación, seguimiento y control, de la entrega de medicamentos y la asignación de citas médicas.

Los problemas identificados en la información generada por las IPS para cada uno de los atributos de los reportes se muestran en la Tabla 4. Los atributos que presentaron los cinco problemas identificados son *Servicio Solicitado* y *Concentración*, mostrando inconsistencias como campos vacíos (VA), datos no normalizados (NN), datos erróneos (ER), datos incompletos (IM), incoherentes (IH) e inconsistentes (IC). Otros atributos como: C*antidad pendiente, Cantidad prescrita, Cantidad entregada, Número total de citas asignadas* se identificaron sólo tres problemas en VA, ER y IC.

Tabla 4 Problemas de DQ en reportes de Asignación de citas y Entrega de medicamentos

Nombre del Reporte	Atributo del Reporte	VA	NN	ER	IM	IH	IC
Asignación de citas Médicas y Entrega de medicamentos	Departamento del afiliado	X	X	X	X		X
	Departamento IPS	X	X	X	X		X
	Municipio del afiliado	X	X	X	X		X
	Municipio IPS	X	X	X	X		X
	Nit prestador (sin dv)	X		X	X	X	X
	Nombres y apellidos del afiliado	X		X	X		X
	Número de identificación	X		X	X	X	X
	Razón social del proveedor	X		X	X		X
	Régimen	X	X	X	X		X
	Tipo de documento	X	X	X	X		X

Nombre del Reporte	Atributo del Reporte	VA	NN	ER	IM	IH	IC
	Genero	X	X	X	X		X
	Teléfono del afiliado	X					
Asignación de citas Médicas	Fecha en que el usuario solicita la cita	X		X	X	X	X
	Fecha en que el usuario solicita le sea asignada la cita.	X		X	X	X	X
	Fecha para la cual se asigna la cita	X		X	X	X	X
	Número Total de Citas Asignadas	X		X		X	
	Servicio Solicitado	X	X	X	X	X	X
	Vía administración	X	X	X	X		X
Entrega de medicamentos	Cantidad pendiente	X		X		X	
	Cantidad prescrita	X		X		X	
	Cantidad entregada	X		X		X	
	Código cum del medicamento	X		X	X	X	X
	Concentración	X	X	X	X	X	X
	Diagnóstico principal	X	X	X	X		X
	Diagnóstico relacionado	X	X	X	X		X
	Días de tratamiento	X		X		X	
	Fecha de autorización	X		X	X	X	X
	Fecha de registro de entrega	X		X	X	X	X
	Fecha de prescripción	X		X	X	X	X
	Fecha de radicación de la fórmula médica por parte del usuario	X		X	X	X	X
	Forma farmacéutica	X	X	X	X		X
	Nombre del medicamento	X		X	X		X
	Fecha de registro de entrega	X		X	X	X	X
	¿Tipo de entrega del medicamento, domiciliaria o en el punto de entrega?	X	X				
	Número de entrega a satisfacción	X		X	X	X	X

Nomenclatura: (VA) Vacío, que señala la presencia de registros sin datos; (NN) No normalizado, que indica que los datos no se presentan en un formato normalizado o estándar; (ER) Erróneo, que refleja que el valor del atributo es incorrecto o inexacto; (IM) Incompleto, que se aplica en los casos en los que falta información o el dato se encuentra incompleto; (IH) Incoherente, que indica que el valor del atributo no concuerda con otros datos o no es coherente con la información general; y (IC) Inconsistente que se usa cuando no se encuentra el dato en absoluto.

3.1.3 Desarrollo de la Solución

Se realizó una adaptación general de las categorías y subcategorías de DQ planteadas por Kahn por medio de dos estrategias:

- **Verificación.** Se centra en confirmar si los valores de los atributos cumplen con las expectativas predefinidas y/o el conocimiento local, lo cual implica la revisión y confirmación de que los datos estuvieran en conformidad con las normas y requisitos establecidos, así como con el contexto específico de su aplicación.

- **Validación.** Se orienta hacia la revisión de los valores de los atributos para garantizar que sean coherentes con fuentes externas de referencia.

La adaptación general realizada para cada subcategoría de DQ se muestra en la Tabla 5, en la cual se especifican las características identificadas de cada una de ellas, además, se detallan los criterios identificados para la verificación y validación.

Tabla 5 Adaptación general de las categorías y subcategorías de DQ

Categoría	Subcategoría	Características	Criterios de Verificación	Criterios de Validación
Conformidad	Conformidad de valor	Restricciones de formato internas	Tipo de dato, formato de dato, longitud	NA
		Valores o rangos permitidos.	Dominio de datos	NA
	Conformidad de relacional	Cumple con las restricciones integridad referencial	Clave foránea	Bases de datos externas
		Cumple con la restricción de unicidad	Clave primaria	NA
		Cumple con las restricciones de nulabilidad	Nulabilidad	NA
	Conformidad computacional	Los valores calculados se ajustan a las especificaciones computacionales o de programación.	NA	NA
Completitud	Completitud	La ausencia de valores en un solo momento en el tiempo está de acuerdo con las expectativas locales o comunes	Atributos cumplen con la ausencia del valor	NA
Plausibilidad	Plausibilidad de unicidad	Los valores de los datos que identifican un solo objeto no se duplican	Atributos que cumplen la condición	NA
	Plausibilidad Atemporal	Busca determinar si los valores, las distribuciones o las densidades de los datos observados concuerdan con	Valores y distribuciones	Bases de datos externas
			Restricciones lógicas entre valores	Bases de datos externas

Categoría	Subcategoría	Características	Criterios de Verificación	Criterios de Validación
		el conocimiento local o "común".	Valores de medición repetida	Bases de datos externas
	Plausibilidad temporal	Los valores observados o derivados se ajustan a las propiedades temporales esperadas	Valores Observados o Derivados	NA
			Secuencias de valores	NA
			Medidas de densidad de valor	NA

Luego se procedió a aplicar la adaptación de estos criterios para cada uno de los atributos presentes en los reportes de entrega de medicamentos y asignación de citas médicas, la cual, se explica en detalle en el Capítulo 4 ADAPTACIÓN DE LAS CATEGORÍAS.

3.1.4 Prueba de la Solución

La validación de la adaptación de las subcategorías de DQ se realizó a través de la participación de un grupo focal compuesto por expertos altamente cualificados en el campo de la salud. Estos expertos cuentan con un extenso conocimiento especializado y más de dos décadas de experiencia en el sector, lo que garantiza que la adaptación cumple con las necesidades de los reportes de entrega de medicamentos y la asignación de citas médicas.

En el Capítulo 6 se presenta la metodología utilizada para llevar a cabo el grupo focal, la selección de expertos, la evaluación realizada y los resultados obtenidos, así como las recomendaciones resultantes.

3.2 Ciclo 2 Incorporación y evaluación de la DQ en el proceso ETL

En este ciclo, se define la incorporación de las DQ adaptadas a los reportes en las fases del proceso ETL, luego se implementan paquetes de transformaciones de ETL para aplicar las DQ adaptadas y por último se evalúa la DQ por medio del indicador de confianza.

3.2.1 Observación

En esta etapa se estudiaron las fases del proceso ETL y la incorporación en este proceso de las categorías DQ planteadas por Kahn que presentaban los estudios primarios seleccionados [13], [14] y [15], encontrando en algunos de estos lo siguiente:

- Mostraban diagramas detallados que representaban la aplicación de estas categorías, mientras que otros no proporcionaban una visualización tan clara.

- Algunos estudios detallaban la aplicación de las categorías de DQ. Sin embargo, otros solo mencionaban las categorías DQ de manera superficial y no proporcionaban detalles específicos.

En el marco de cada fase del proceso ETL, en estos estudios primarios [13], [14] y [15] se encontró lo siguiente:

- **Extracción.** Se observó que la mayoría de los orígenes de datos procedían de sistemas transaccionales.
- **Transformación.** La mayoría de los estudios incluían un proceso inicial de limpieza, normalización y enriquecimiento de los datos, y algunos hacían referencia a la aplicación de criterios de inclusión o exclusión de registros. Además, los registros inconsistentes eran eliminados durante la transformación.
- **Carga.** En algunos estudios, la información final se almacena en bases de datos antes de ser transferida a los componentes de monitoreo web, mientras que, en otros casos, la información se guarda directamente en archivos planos.

3.2.2 Identificación del problema

Se determinó en que fases del proceso ETL se podían incorporar los criterios adaptados a las categorías de DQ, para los reportes de entrega de medicamentos y asignación de citas médicas:

- **Extracción.** Se identificó que la fuente de origen de los datos para los reportes de entrega de medicamentos y asignación de citas médicas, son archivos en formato Excel proporcionados por las EPS, que se generan tras la consolidación de los reportes diligenciados previamente por las IPS.

- **Transformación.** Inicialmente es necesario realizar un proceso *limpieza y normalización de datos*, para asegurar que los datos estén en un formato adecuado para su posterior procesamiento. Luego se procede a la aplicación de las categorías de DQ adaptando los criterios definidos para cada una de ellas en el siguiente orden: (i) *Conformidad*, verificar que cumplan con las restricciones sintácticas o estructurales (ii) *Completitud*, para garantizar que no falte información esencial, y (iii) *Plausibilidad,* evaluar la coherencia y consistencia de los datos.

- **Carga.** Se completa el proceso de incorporación de los datos transformados en el sistema final de destino, que puede ser una base de datos, un archivo en Excel, un almacén de datos u otra plataforma de almacenamiento. Además, se procede a la incorporación de los archivos de registro (logs) que son generados durante el proceso de transformación de los datos, permitiendo la identificación del error detallado en cada etapa y categoría del proceso, lo que resulta fundamental para la trazabilidad y seguimiento de los errores.

Luego de identificada la incorporación de los criterios de las categorías adaptadas de DQ en las fases del proceso ETL para los dos reportes, se procedió a definir una arquitectura para presentar de forma gráfica esta incorporación, la cual se muestra en detalle en el Capítulo 5 en la sección 5.1, denominado “Arquitectura del proceso ETL”.

3.2.3 Desarrollo de la solución

Con la arquitectura de incorporación de la DQ en las fases del proceso ETL, se procedió a realizar la implementación del proceso ETL específico (paquetes y transformaciones) para los reportes de entrega de medicamentos y asignación de citas médicas, lo cual se muestra en detalle en el Capítulo 5 en la sección 5.2 Proceso ETL.

3.2.4 Prueba de la solución

Una vez finalizado el proceso ETL se procede a realizar la evaluación de los resultados de las subcategorías de DQ aplicadas a los reportes por medio del indicador de Confianza de acuerdo con la Fórmula 1 definida previamente en el Capítulo 2 sesión 2.1.4.

CAPITULO 4

4 ADAPTACIÓN DE LAS CATEGORÍAS

En esta sesión se presenta en detalle la adaptación de las categorías y subcategorías de DQ para los reportes de entrega de medicamentos y asignación de citas médicas.

4.1 Adaptación Categorías al Reporte de Entrega de Medicamentos

La adaptación de las categorías de DQ *Conformidad de Valor y Conformidad Relacional* para el reporte de entrega de medicamentos, se presenta en la Tabla 6 (notaciones utilizadas se explican al final de la tabla). Esta adaptación se llevó a cabo considerando diferentes aspectos de cada atributo, como:

- **ID:** Identificación numérica única para cada atributo.
- **Atributo**: Nombre del atributo o campo de datos.
- **Conformidad de Valor**: Indica la adaptación realizada a los criterios identificados para esta categoría, garantizando que los valores de los datos cumplan con estándares específicos, incluyendo detalles sobre el tipo de dato, longitud, formato y dominio.
- **Conformidad Relacional**: Detalla la adaptación realizada a los criterios identificados para esta categoría, con respecto a obligatoriedad y las relaciones entre los datos como: Claves primarias y Foráneas.
- **Validación**: Permite identificar si los valores de los atributos deben cumplir con datos almacenados en alguna base de datos externa.
- **Verificación**: Facilita la comprobación de si los valores de los atributos coinciden con el conocimiento local o común.

Como se puede observar en la Tabla 6, para los valores de los atributos de este reporte con respecto a la *Conformidad de valor*, no se aplica (NA) la adaptación de criterios de validación, debido a que no es necesario contrastar el cumplimiento de estos criterios con alguna base de datos externa. Con respecto a los criterios de verificación, para el atributo *Número de identificación* se especifica: que el Tipo de dato (TD) debe ser un Texto corto (TC) debido a que algunas identificaciones como los pasaportes pueden ser alfanuméricas, se establece una longitud (LO) de 16 caracteres, se indica que este atributo no cumple con

un formato (FO) por lo que se coloca No aplica (NA), tampoco se dispone de un dominio (DO) de valores establecidos para este atributo, por lo que se coloca también como NA.

Tabla 6 Adaptación de Conformidad de Valor y Relacional – Entrega de medicamentos

ID	Atributo	Conformidad Valor					Conformidad Relacional			
		Adaptación Criterios					Adaptación Criterios			
		Verificación				Validación	Verificación			Validación
		TD	LO	FO	DO		OB	CP	CF	
1	Tipo de documento	TC	2	NA	NA	NA	NN	NA	NA	BDUA
2	Número de identificación	TC	16	NA	NA	NA	NN	NA	BDA	BDUA
3	Nombres y apellidos	TL	80	NA	NA	NA	NN	NA	NA	NA
4	Genero	TC	9	NA	EJ5	NA	NN	NA	NA	BDUA
5	Régimen	TC	12	NA	NA	NA	NN	NA	NA	BDUA
6	Departamento del afiliado	NM	2	NA	NA	NA	NN	NA	BDA	BDUA
7	Municipio del afiliado	NM	5	NA	NA	NA	NN	NA	BDA	BDUA
8	Teléfono del afiliado	NM	20	NA	NA	NA	NL	NA	NA	NA
9	Código cum del medicamento	NM	11	NA	NA	NA	NN	NA	NA	BDIN
10	Nombre del medicamento	TL	200	NA	NA	NA	NN	NA	NA	NA
11	Concentración	TL	30	NA	EJ1	NA	NN	NA	NA	NA
12	Forma farmacéutica	TL	180	NA	EJ2	NA	NN	NA	NA	NA
13	Vía administración	TC	40	NA	EJ3	NA	NN	NA	NA	NA
14	Días de tratamiento	NM	4,0	NA	NA	NA	NN	NA	NA	NA
15	Cantidad prescrita	NM	4,0	NA	NA	NA	NN	NA	NA	NA
16	Cantidad entregada	NM	4,0	NA	NA	NA	NL	NA	NA	NA
17	Cantidad pendiente	NM	4,0	NA	NA	NA	NN	NA	NA	NA
18	Diagnóstico principal	TC	4	NA	NA	NA	NN	NA	NA	CIE
19	Diagnóstico relacionado	TC	4	NA	NA	NA	NL	NA	NA	CIE
20	Fecha de prescripción	FE	10	FF	NA	NA	NN	NA	NA	NA
21	Fecha de autorización	FE	10	FF	NA	NA	NN	NA	NA	NA
22	Fecha de radicación de la fórmula médica por parte del usuario	FE	10	FF	NA	NA	NN	NA	NA	NA
23	Fecha de registro de entrega	FE	10	FF	NA	NA	NL	NA	NA	NA
24	Nit prestador (sin dv)	NM	16	NA	NA	NA	NN	NA	BDI	REPS
25	Razón social del proveedor	TL	100	NA	NA	NA	NN	NA	NA	NA
26	¿Tipo de entrega del medicamento, domiciliaria o en el punto de entrega?	BO	NA	NA	EJ4	NA	NL	NA	NA	NA
27	Número de entrega a satisfacción	TL	50	NA	NA	NA	NL	NA	NA	NA

Nomenclatura: Texto Corto (TC), Numérico (NM), Texto largo (TL), Booleano (BO), Fecha (FE), NA (0), Formato fecha DD-MM-YYYY (FF), Nulo (NL), No Nulo (NN), 200 mg, 300 mg, 600mg/100000 u.i. (EJ1), Solución inyectable, Tableta (EJ2), Oral, Intramuscular, Infiltrativa – local (EJ3), 1: Domiciliaria, 2: Punto de entrega (EJ4).

En esta misma tabla, en cuanto a la *Conformidad relacional*, se aplica la verificación (fuentes internas) con las Bases de datos de Afiliados EPS (BDA) e IPS Contratadas (BDI) y la validación (fuentes externas) con la Base de Datos Única de Afiliados (BDUA) del MSPS, Base de Datos del INVIMA (BDIN), Clasificación Internacional de Enfermedades CIE10 (CIE) y el Registro Especial de Prestadores de Servicios de Salud (REPS). Para el atributo *Número de identificación* se establece en la columna de obligatoriedad (OB) que es No Nulo (NN) porque es un campo obligatorio para poder identificar al afiliado que corresponde la fórmula médica, la clave primaria (CP) no aplica porque un mismo afiliado puede tener diferentes prescripciones, pero es clave foránea (CF) porque el valor de este atributo se verifica que exista en la BDA y se valida que exista en la BDUA.

La adaptación de la categoría de *Completitud* se encarga de evaluar que no falten datos en los atributos específicos, y que la ausencia de datos cumpla con las regulaciones establecidas por el MSPS. En este contexto, se evaluaron que tres atributos cumplieran con el criterio de verificación: *Cantidad entregada*, *Fecha de registro de entrega*, y *Diagnóstico relacionado*. Como se puede observar en la Tabla 7, la *Cantidad Entregada* puede ser nula o contener un cero, indicando que el medicamento no se entregó, pero en este caso, la *fecha de registro de entrega* del medicamento también debe ser nula. También, cuando el atributo *Fecha de registro de entrega* es nula, el atributo *Cantidad entregada* también deber ser nulo o cero. Con respecto al *diagnóstico relacionado* este puede ser nulo, siempre y cuando exista el *diagnóstico principal*. Para este reporte no fue necesario adaptar criterios de validación.

Tabla 7 Adaptación categoría Completitud – Entrega de medicamentos

Atributo	Adaptación Criterio Verificación	Adaptación Criterio Validación
Cantidad entregada	Si *Cantidad entregada* es nula o igual a cero: la *Fecha de registro de entrega* debe ser nula.	NA
Fecha de registro de entrega	Si Fecha de registro de entrega es nula: la *Cantidad entregada* debe ser nula o cero.	NA
Diagnóstico relacionado	Si *Diagnóstico relacionado* es nulo: *Diagnóstico principal* no debe ser nulo.	NA

En la Tabla 8, se aborda la subcategoría *Plausibilidad de Unicidad*, cuyo propósito es identificar la existencia de registros duplicados en los registros del reporte. En este caso,

para el criterio de verificación, la combinación de los atributos: *Tipo de documento*, *Número de identificación*, *Código cum del medicamento* y *Fecha de prescripción*, deben ser únicos en el conjunto de datos del reporte. No se adaptan criterios de validación para esta subcategoría.

Tabla 8 Adaptación subcategoría Plausibilidad de Unicidad – Entrega de medicamentos

Atributos	Adaptación Criterio Verificación	Adaptación Criterio Validación
• Tipo de documento • Número de identificación • Código cum del medicamento • Fecha de prescripción	No deben existir registros duplicados: *Tipo de documento* **+** *Número de identificación* **+** *Código cum del medicamento* **+** *Fecha de prescripción*	NA

En cuanto a la subcategoría *Plausibilidad Atemporal*, como se observa en la Tabla 9 los valores de los atributos *Tipo de documento y Número de identificación* se concatenan para verificar que los valores coincidan con la BDA (verificación) y con la BDUA (validación). Lo mismo ocurre con las filas 2 y 3 de esta misma tabla. En relación con el atributo *Días de tratamiento* se aplica una restricción lógica de verificación, ya que no deben existir días de tratamiento iguales a 0, porque incluso los medicamentos de dosis única tienen al menos un día mínimo de tratamiento; pero no es necesario aplicar ningún criterio de validación. Para los atributos *Cantidad entregada, Cantidad pendiente y Cantidad prescrita*, también se estableció una restricción lógica de verificación con el propósito de revisar la coherencia de las C*antidades registradas*, lo que implica que la suma de la *Cantidad entregada y la Cantidad pendiente* debe ser menor o igual a la *Cantidad prescrita*. No se adaptan criterios de validación para estos atributos.

Tabla 9 Adaptación subcategoría Plausibilidad Atemporal – Entrega de medicamentos

Atributos	Adaptación Criterio Verificación	Adaptación Criterio Validación
Tipo de documento **+** Número de identificación	**Coinciden** con la BD Afiliados.	**Coinciden** BDUA
Régimen afiliado **+** Tipo de identificación **+** Número de identificación	**Coinciden** con la BD Afiliados.	**Coinciden** BDUA
Departamento del afiliado **+** Municipio del afiliado **+** Tipo de identificación **+** Número de identificación	**Coinciden** con la BD Afiliados.	**Coinciden** BDUA

Atributos	Adaptación Criterio Verificación	Adaptación Criterio Validación
Días de tratamiento	No deben existir días de tratamiento en 0, medicamentos de dosis única se maneja en un día.	NA
Cantidad entregada, Cantidad pendiente y Cantidad prescrita	*Cantidad entregada* **+** *Cantidad pendiente* **<=** *Cantidad prescrita.*	NA

Por último, en la Tabla 10 para la subcategoría *Plausibilidad Temporal*, se verifica que la *Fecha de registro de entrega* sea posterior o igual a la *Fecha prescripción* del medicamento, de la misma forma se verifica que el atributo *Fecha de autorización* sea posterior o igual a la *Fecha prescripción.* Para el atributo *Cantidad entregada* y *Fecha de Prescripción***,** se realiza una verificación teniendo en cuenta dos momentos, el registro actual *(t)* y el registro del mes anterior *(t-1)*, cuando los dos registros tengan los mismos valores en ciertos atributos. En el caso de *Cantidad entregada* los valores de los registros *(t)* y *(t-1)*, que tengan los mismos valores en los atributos *Tipo de documento, Número de identificación, Código cum del medicamento, Cantidad prescrita y Fecha de prescripción*, deben cumplir que el atributo *Cantidad pendiente (t-1) menos la Cantidad entregada (t)* debe ser mayor o igual *Cantidad pendiente (t).* En el caso de *Fecha de Prescripción* los valores de los registros *(t)* y *(t-1)*, que tengan los mismos valores en los atributos *Tipo de documento, Número de identificación, Código cum del medicamento, Cantidad prescrita*, deben cumplir que el atributo *Fecha de Prescripción (t)* debe ser mayor a la suma entre *Fecha de Prescripción (t-1)* y *Días de tratamiento (t-1)*.

Tabla 10 Adaptación subcategoría Plausibilidad Temporal – Entrega de medicamentos

Atributos	Adaptación Criterio Verificación	Adaptación Criterio Validación
Fecha de registro de entrega	*Fecha de registro de entrega* **>=** *Fecha de prescripción*.	NA
Fecha de autorización	*Fecha de autorización* **>=** *Fecha de prescripción*.	NA
Cantidad entregada (t)	Para dos registros en el tiempo **t y t-1,** que tengan el mismo: *Tipo de documento* **+** *Número de identificación* **+** *Código cum del medicamento* **+** *Cantidad prescrita* **+** *Fecha de prescripción.* Se revisa: *Cantidad pendiente* **(t-1) -** *Cantidad entregada* **(t) >=** *Cantidad pendiente* **(t)**	NA

Atributos	Adaptación Criterio Verificación	Adaptación Criterio Validación
Fecha de Prescripción (t)	Para dos registros en el tiempo **t y t-1,** que tengan el mismo: *Tipo de documento* **+** *Número de identificación* **+** *Código cum del medicamento* **+** *Días de Tratamiento*. Se revisa: *Fecha de Prescripción* **(t) >** *Fecha de Prescripción* **(t-1)** *+ Días de tratamiento* **(t-1)**	NA

4.2 Adaptación categorías de DQ al reporte asignación de citas médicas

La Tabla 11, que presenta esta adaptación tiene la misma estructura que la tabla presentada para la adaptación al reporte de entrega de medicamentos (Tabla 6), por esto no se explica en esta sección el significado de las columnas.

Para los valores de los atributos de este reporte con respecto a la *Conformidad de valor*, no se aplica la adaptación de criterios de validación, debido a que no es necesario contrastar el cumplimiento de estos criterios con alguna base de datos externa. Con respecto a los criterios de verificación, al igual que para la adaptación realizada al reporte de entrega de medicamentos por cada uno de los atributos se especifica el tipo de dato, la longitud, formato y dominio (si aplica). Por ejemplo, para el atributo *Servicio solicitado*, se define como un dato numérico con una longitud de cuatro dígitos, no aplica un formato o dominio. Además, en la adaptación de los criterios de conformidad relacional, se establece que el dato no puede ser nulo, no se considera clave primaria, pero sí tiene relaciones como clave foránea con la fuente interna BDI y con la fuente externa REPS.

En cuanto a la Conformidad relacional, se aplica la verificación (fuentes internas) con las bases BDA, BDI y la validación (fuentes externas) con BDUA y Registro Especial de Prestadores de Servicios de Salud (REPS).

Tabla 11 Adaptación Conformidad de Valor y Relacional - Asignación de citas médicas

ID	Atributo	Conformidad valor					Conformidad Relacional			
		Adaptación Criterios					Adaptación Criterios			
		Verificación				Validación	Verificación			Validación
		TD	LO	FO	DO		OB	CP	CF	
1	Tipo de Documento	TC	2	NA	NA	NA	NN	NA	BDA	BDUA
2	Número de Identificación	NM	16	NA	NA	NA	NN	NA	BDA	BDUA
3	Nombres y apellidos	TL	80	NA	NA	NA	NN	NA	NA	NA

ID	Atributo	Conformidad valor					Conformidad Relacional			
		Adaptación Criterios					Adaptación Criterios			
		Verificación				Validación	Verificación			Validación
		TD	LO	FO	DO		OB	CP	CF	
4	Régimen	TC	12	NA	NA	NA	NN	NA	BDA	BDUA
5	Departamento del afiliado	NM	2	NA	NA	NA	NN	NA	BDA	BDUA
6	Municipio del afiliado	NM	5	NA	NA	NA	NN	NA	BDA	BDUA
7	Teléfono	NM	30	NA	NA	NA	NN	NA	NA	NA
8	Fecha en que el usuario solicita la cita o cirugía.	FE	10	FF	NA	NA	NN	NA	NA	NA
9	Fecha en que el usuario solicita le sea asignada la cita o cirugía	FE	10	FF	NA	NA	NN	NA	NA	NA
10	Fecha para la cual se asigna la cita o cirugía	FE	10	FF	NA	NA	NN	NA	NA	NA
11	Departamento IPS	NM	2	NA	NA	NA	NN	NA	BDI	REPS
12	Municipio IPS	NM	5	NA	NA	NA	NN	NA	BDI	REPS
13	Nit prestador (sin dv)	NM	12	NA	NA	NA	NN	NA	BDI	REPS
14	Código de Habilitación	NM	12	NA	NA	NA	NN	NA	BDI	REPS
15	Razón social del proveedor	TL	100	NA	NA	NA	NN	NA	NA	NA
16	Servicio solicitado	NM	4	NA	NA	NA	NN	NA	BDI	REPS
17	Otra especialidad o Cirugía	TL	100	NA	NA	NA	NU	NA	NA	NA
18	Número Total de citas asignadas	NM	2	NA	NA	NA	NN	NA	NA	NA

Nomenclatura: Tipo de Dato (TD), longitud (LO), Formato (FO), Dominio (DO), Obligatoriedad (OB), Clave primaria (PR), Clave foránea (FO), Texto Corto (TC), Numérico (NM), Texto Largo (TL), Fecha (FE), No Aplica (NA), Formato Fecha DD-MM-YYYY (FF), Nulo (NU), No Nulo (NN), BD Afiliados EPS(BDA), BD Ips Contratadas EPS(BDI), BD Base de Datos Única de Afiliados - BDUA(BDUA),

En la Tabla 12, se presenta la adaptación del criterio de verificación para la categoría de *Completitud* en relación con los atributos *Servicio solicitado* y *Otra especialidad o Cirugía,* donde se establece que, si el atributo *Servicio solicitado* es nulo, entonces se espera que el atributo *Otra especialidad o Cirugía* no este vacío, y viceversa. No fue necesario adaptar criterios de validación para esta subcategoría.

Tabla 12 Adaptación categoría Completitud - Asignación de citas médicas

Atributo	Adaptación Criterio Verificación	Adaptación Criterio Validación
Servicio solicitado	Si *Servicio solicitado* es nulo: *Otra especialidad o Cirugía* no debe ser nulo.	NA

Otra especialidad o Cirugía	Si *Otra especialidad o Cirugía* es nulo: *Servicio solicitado* no debe ser nulo.	NA

Con respecto a la subcategoría *Plausibilidad de Unicidad*, en la Tabla 13 se muestra la adaptación del criterio de verificación que establece que no deben existir más de dos registros duplicados, que tengan los mismos valores para los atributos *Tipo de documento, Número de identificación, Fecha para la cual se asigna la cita o cirugía, Nit prestador,* y *Servicio Solicitado.* No fue necesario adaptar criterios de validación para esta subcategoría.

Tabla 13 Adaptación subcategoría Plausibilidad de Unicidad - Asignación de citas médicas

Atributos	Adaptación Criterio Verificación	Adaptación Criterio Validación
• Tipo de documento • Número de identificación • Fecha para la cual se asigna la cita o cirugía • Nit prestador • Servicio Solicitado	No deben existir más de 2 registros duplicados: *Tipo de documento* **+** *Número de identificación* **+** *Fecha para la cual se asigna la cita o cirugía* **+** *Nit prestador* **+** *Servicio Solicitado* >=2	NA

En la Tabla 14, para la subcategoría *Plausibilidad Atemporal*, se establecen criterios específicos de verificación y validación para cada conjunto de atributos. En el caso de los atributos *Tipo de documento y Número de identificación*, se revisa que coincidan con las fuentes interna BDA (verificación) y con la fuente externa BDUA (validación). De manera similar, para los atributos de las filas dos y tres.

Tabla 14 Adaptación subcategoría Plausibilidad Atemporal - Asignación de citas médicas

Atributos	Adaptación Criterio Verificación	Adaptación Criterio Validación
Tipo de documento **+** Número de identificación	**Coinciden** con la BD Afiliados.	BDUA
Régimen **+** Tipo de Identificación **+** Número de identificación	**Coinciden** con la BD Afiliados.	BDUA
Departamento del afiliado **+** Municipio del afiliado **+** Tipo de Identificación **+** Número de identificación	**Coinciden** con la BD Afiliados.	BDUA

Por último, en la Tabla 15, se presenta para la subcategoría *Plausibilidad Temporal*, los criterios de verificación dado que no existen fuentes externas para su comprobación, por lo que no se adaptan criterios de validación. Para el atributo *Fecha de asignación cita o cirugía* se verifica que sea mayor o igual a *Fecha en que el usuario solicita la cita o cirugía,* de manera similar, se aplica la adaptación de verificación del atributo *Fecha en que el usuario*

solicita le sea asignada la cita. No fue necesario adaptar criterios de validación para esta subcategoría.

Tabla 15 Adaptación subcategoría Plausibilidad Temporal - Asignación de citas médicas

Atributos	Adaptación Criterio Verificación	Adaptación Criterio Validación
Fecha de asignación cita o cirugía	*Fecha para la cual se asigna la cita o cirugía* **>=** *Fecha en que el usuario solicita la cita o cirugía*.	NA
Fecha en que el usuario solicita le sea asignada la cita	*Fecha en que el usuario solicita le sea asignada la cita* **>=** *Fecha en que el usuario solicita la cita o cirugía*.	NA

CAPITULO 5

5 INCORPORACIÓN CATEGORIAS EN EL PROCESO ETL

La incorporación de las categorías de DQ en el proceso de ETL se divide en dos partes principales: (i) la arquitectura del proceso ETL y (ii) el propio proceso ETL, el cual plantea la ejecución de los paquetes que tienen relación con todas las subcategorías de DQ, en el siguiente orden: Conformidad de Valor, Conformidad Relacional, Completitud, Plausibilidad de Unicidad, Plausibilidad Atemporal y por último Plausibilidad Temporal.

5.1 Arquitectura del proceso ETL

La arquitectura de este proceso se divide en tres fases: *Extracción*, de los archivos fuente; *Transformación*, limpieza y aplicación de las categorías de calidad de datos adaptadas; y *Carga*, de los archivos finales con la verificación, validación y log de errores como se muestra en la Figura 1.

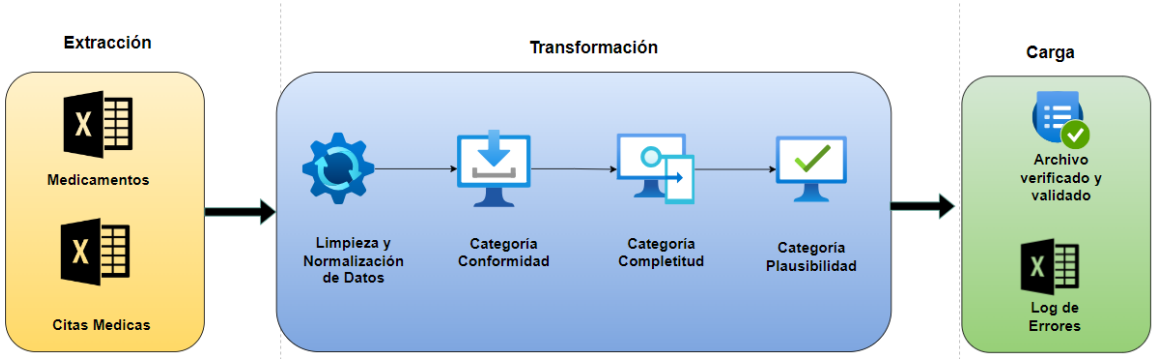


Figura 1 Diagrama categorías de DQ en el proceso ETL

5.1.1 Extracción.

En esta fase (ver Figura 1) se obtienen los datos de dos fuentes principales, los archivos consolidados mensuales de los reportes de entrega de medicamentos (*Medicamentos*) y asignación de citas médicas (*Citas médicas*).

5.1.2 Transformación

Como se observa en la Figura 2, esta fase está compuesta por cuatro procesos: Limpieza y normalización de datos, para garantizar su integridad y el formato adecuado; Categoría de Conformidad, Categoría de Completitud y la Categoría de Plausibilidad.

a) Limpieza y normalización de datos

En la fase de transformación se inicia con un proceso de limpieza y luego la normalización de los datos (ver Figura 2).

Figura 2 Limpieza y normalización de datos

Esta fase implica cuatro actividades antes de avanzar hacia la aplicación de la categoría de Conformidad:

1. **Detección y manejo de valores nulos.** Identificación y tratamiento de valores nulos o faltantes en los datos mediante su eliminación o rellenándolos con valores predeterminados.

2. **Corrección de errores tipográficos y de formato.** Rectificación de posibles errores tipográficos o de formato que puedan afectar la consistencia de los datos.

3. **Estandarización de valores.** Conversión de los valores de los datos a un formato estándar para garantizar coherencia, incluyendo la normalización de fechas, unidades de medida, códigos, etc.

4. **Manejo de mayúsculas/minúsculas y acentos.** Homogeneización de la presentación del texto mediante la conversión a mayúsculas o minúsculas, y la eliminación o normalización de acentos.

b) Categoría de Conformidad

Este proceso está compuesto por dos subprocesos para las subcategorías de: Conformidad de Valor y de Conformidad Relacional.

1. Subcategoría de Conformidad de Valor

En la Figura 3 se muestra la incorporación de los criterios de verificación adaptados a la subcategoría de *Conformidad de Valor*, que buscan que cada atributo cumpla con las restricciones de formato definidas, tales como *CV_Tipo de Dato, CV_Longitud y CV_Formato* y las restricciones de valores o rangos permitidos como *CV_Dominio*, para

asegurarse de que estén en conformidad con la metadata de la base de BDA y de las BDI. No fue necesario aplicar criterios de validación (No Aplica).

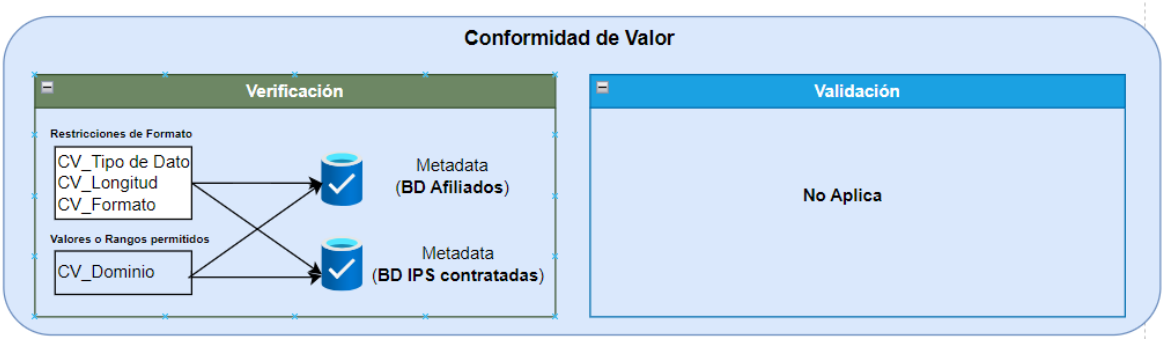


Figura 3 Conformidad de Valor

2. Subcategoría Conformidad Relacional

En la Figura 4, se detalla la incorporación de los criterios de verificación adaptados a la subcategoría *Conformidad Relacional*: *CR_Llave primaria* y *CR_Clave foránea* en cada uno de los reportes, lo que permite establecer relaciones entre las bases de datos internas, como la base de datos de Afiliados (BDA) y las bases de datos de IPS contratadas (BDI). Además, se considera la *CR_Obligatoriedad*, para establecer las relaciones entre la metadata de BDA y BDI. En los criterios de validación se utilizan las fuentes de datos externas, como la Base de Datos Única de Afiliados (BDUA) del MPSP, el Instituto Nacional de Vigilancia de Medicamentos y Alimentos (INVIMA), el Registro Especial de Prestadores de Servicios de Salud (REPS) y la Clasificación Internacional de Enfermedades (CIE10).

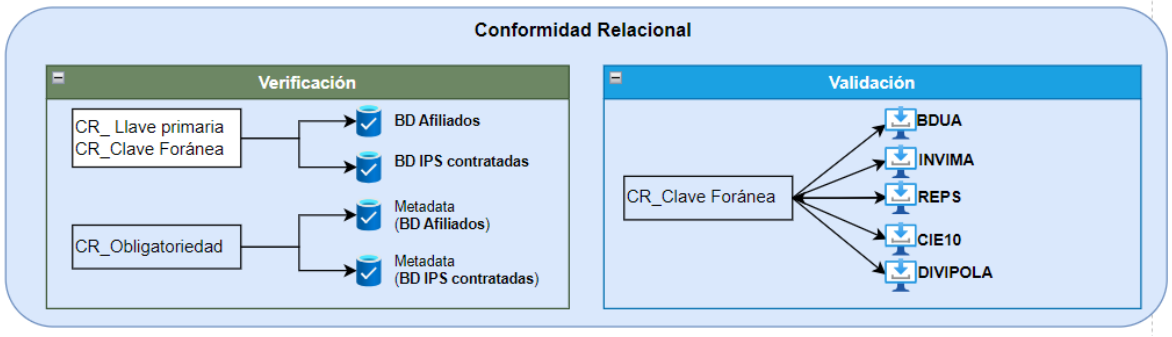


Figura 4 Conformidad Relacional

c) Categoría Completitud

En la Figura 5 se presentan los criterios de verificación para la *Categoría Completitud*, la cual se encarga de evaluar la ausencia de datos en atributos específicos, siguiendo los

parámetros definidos en la Tabla 7 para el reporte de entrega de medicamentos y en la Tabla 12 para el reporte de asignación de citas médicas. Para los criterios de validación se utiliza la fuente externa del Instituto Nacional de Vigilancia de Medicamentos y Alimentos (Invima), para conformar que el *Código Cum* se encuentre efectivamente desabastecido.

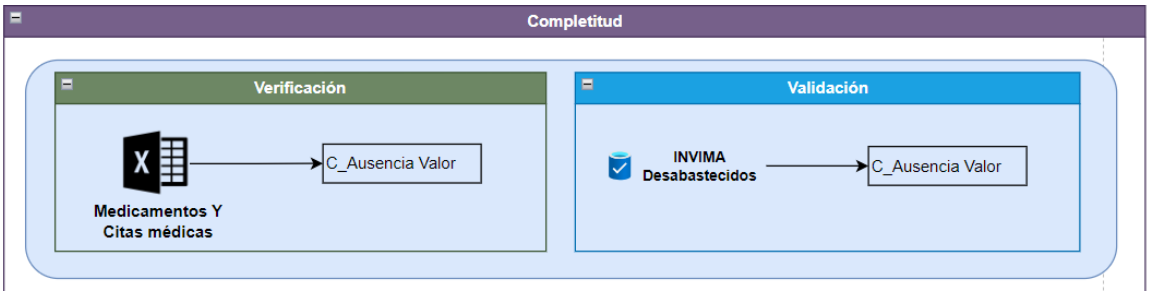


Figura 5 Completitud

d) Plausibilidad

Este proceso está compuesto por tres subprocesos para las subcategorías de: Plausibilidad de Unicidad, Plausibilidad Atemporal y Plausibilidad Temporal.

1. Plausibilidad de Unicidad

En la Figura 6 se presentan los criterios de verificación para la *Subcategoría Plausibilidad de Unicidad*, la cual garantiza que los valores de los datos que identifican un solo objeto (*PU_Objetos*) no se dupliquen, siguiendo los parámetros definidos en la Tabla 8 (Medicamentos) Tabla 13 (Citas Médicas). Para esta subcategoría no se aplica ninguna validación.

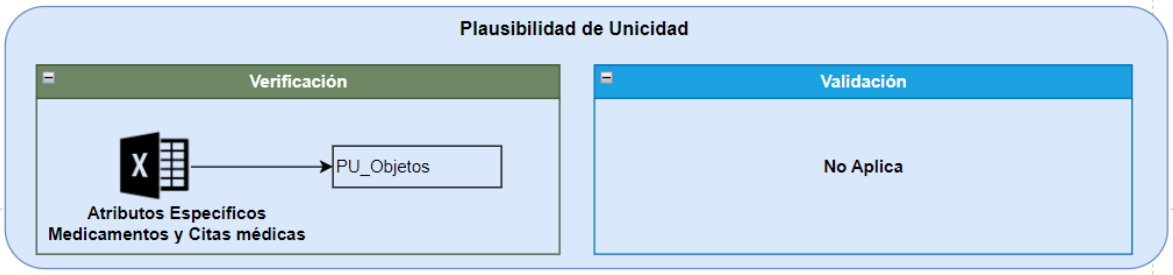


Figura 6 Plausibilidad de Unicidad

2. Plausibilidad Atemporal

En la verificación para la subcategoría *Plausibilidad Atemporal*, como se visualiza en la Figura 7, se comprueba si los valores (*PA_Valores*) o distribuciones (*PA_Distribucciones*) de ciertos atributos de los reportes concuerdan con la información de otros atributos y con

el conocimiento local (BDA). En validación, se revisa si los valores o distribuciones concuerdan con las fuentes externas (BDUA) de acuerdo con los criterios definidos en las Tabla 9 (Medicamentos) y Tabla 14 (Citas Médicas).

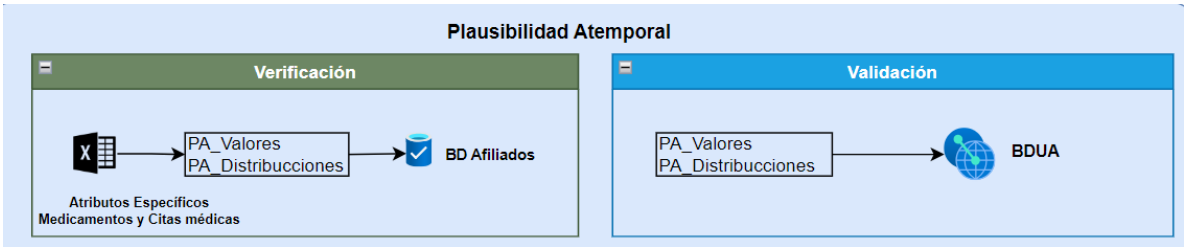


Figura 7 Plausibilidad Atemporal

3. Plausibilidad Temporal

En la verificación de la subcategoría *Plausibilidad Temporal,* como se muestra en la Figura 8, comprueba si los valores observados (*PT_Valores_Observados*) o derivados (*PT_Valores_Derivados*) de los de los reportes tienen una secuencia temporal coherente, por ejemplo, si existen dos fechas las cuales de forma lógica una no debe ser menor a la otra, esto debe ser verificado. Los criterios para este reporte se encuentran detallados en la Tabla 10 (Medicamentos) y en la Tabla 15 (Citas Médicas). Por otro lado, no se aplica ninguna validación para esta subcategoría.

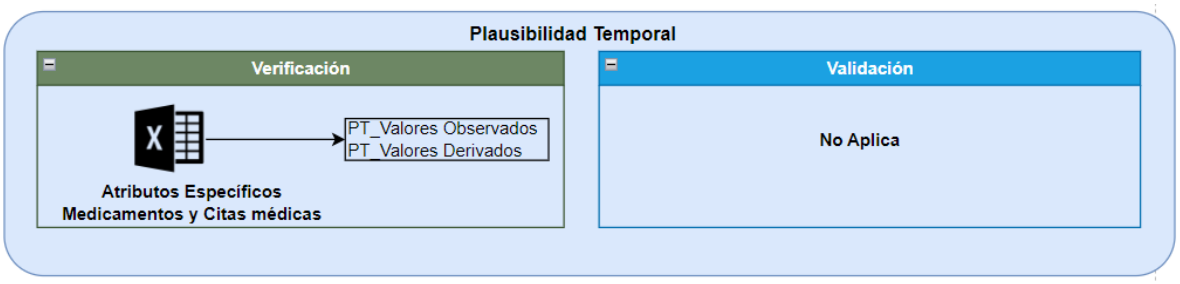


Figura 8 Plausibilidad Temporal

5.1.3 Carga

Esta es la última fase del proceso ETL (ver Figura 1), que culmina con la exportación de un archivo (*Archivo verificado y validado*) para cada uno de los reportes, cuya DQ de los atributos ha sido evaluada por medio de la adapatación de los criterios de verificación y validación de acuerdo con la fase anterior.

Además de este archivo final, se genera un registro detallado (*log Errores*) que recoge los errores resultantes de la aplicación de cada una de las adaptaciones por cada categoría.

En este registro se incluye el código específico del error, el atributo que presenta la inconsistencia, una descripción detallada del error identificado, junto con la categoría y subcategoría a la que corresponde el error.

5.2 Proceso ETL

En esta sesión, se presenta la implementación de un proceso ETL con la incorporación de las categorías de calidad de datos para los reportes de entrega de medicamentos y asignación de citas médicas, por medio de la herramienta Pentaho.

Estos reportes fueron proporcionados por Asmet Salud EPS SAS, comprendiendo los meses de enero y febrero de 2021 para el reporte de medicamentos, y abril de 2021 para el reporte de citas médicas.

5.2.1 Entrega de medicamentos

El flujo de trabajo de la implementación del proceso ETL para el reporte de entrega de medicamentos, como se muestra en la Figura 9, consta de cinco transformaciones que son: (i) Limpieza y Normalización de los datos, (ii) Conformidad y Completitud, (iii) Plausibilidad de Unicidad y Atemporal, (iv) Plausibilidad Temporal y (v) Consolidado de log de errores. Cada una de estas transformaciones está identificada por etiquetas que ofrecen información específica: (i) *Etiquetas Amarillas*, identifican el atributo al que están relacionadas dentro de la transformación. (ii) *Etiquetas Rojas*, indican la función donde se almacena el consolidado de errores. (iii) *Etiquetas Verde Claro,* indican el archivo final que concluye la transformación. (iv) *Etiquetas Verde Oscuro*, representan la categoría o subcategoría que se está aplicando en cada proceso.

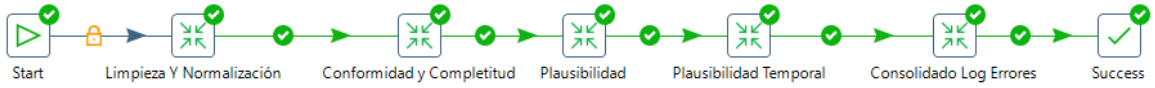


Figura 9 Proceso ETL Medicamentos

1. **Limpieza y Normalización de los datos**

El flujo de esta transformación se visualiza en la Figura 10, y se compone de cuatro procesos: Corrección de errores tipográficos y de formato, Detección y manejo de valores nulos, Estandarización y organización de valores y Manejo de mayúsculas/minúsculas y acentos.

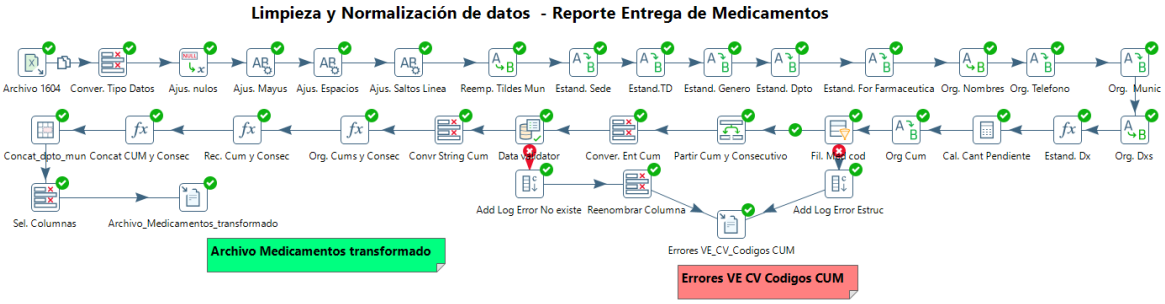


Figura 10 Limpieza, normalización de datos Medicamentos

- **Corrección de errores tipográficos y de formato:** se verifican y/o ajustan los atributos al tipo de datos establecidos (*Conver. Tipo Datos*) de acuerdo con la adaptación propuesta como se muestra en la Tabla 6. Por ejemplo, se asegura que los *diagnósticos, nombres y apellidos, razón social del prestador, genero, departamento, municipio, sede* estén en formato string, las fechas estén en el tipo de dato "date" y en un formato de fecha ‘dd-mm-yyyy’ y que los dias de tratamiento y cantidades estén en tipo de dato numérico.

- **Detección y manejo de valores nulos:** se realizaron ajustes en los valores nulos (*Ajust. Nulos*). Por ejemplo, en los campos como *días de tratamiento, diagnóstico principal, forma farmacéutica, o teléfono* están vacíos, se reemplazan por el valor "0". Las fechas de entrega que carecen de un valor válido se asignan a la fecha "01-01-3000".

- **Manejo de mayúsculas/minúsculas, acentos, espacios en blanco y saltos de línea:** todos los registros se convirtieron a mayúsculas (*Ajus. Mayus*) para lograr uniformidad en los datos. Además, se procedió a eliminar espacios en blanco (*Ajus. Espacios)*, saltos de línea (*Ajus. Saltos Linea*) en atributos de campo de texto, como nombres, apellidos, municipios y departamentos y ajuste en las tildes del atributo municipios (*Reemp. Tildes Mun).*

- **Estandarización y organización de valores:** se realizaron ajustes para estandarizar ciertos atributos durante el proceso. Por ejemplo:
    - Sede departamental del reporte (*Estand. Sede*) ajustando *N.Santander* en *Norte de Santander*.

- Tipo de documento (*Estand.TD*) modificando los registros como *cedula* a *CC*, *Registro Civil* a *RC.*
- Género (*Estand. Genero*) se categorizo para mostrar M (masculino) y F (femenino).
- Nombres del departamento (*Estand. Dpto*), identificados como *Valle* o con el código *76001* se ajustaron a *Valle del* Cauca.
- Forma farmacéutica *(Estand. For farmacéutica)* se ajusta *solución* para que fuera representada como *solución oral*, *tab* como *tableta*, y *sub* como *subcutánea.*

Además, otros procesos como:

- Se eliminaron registros numéricos innecesarios en los atributos nombres y apellidos completos (*Org. Nombres*).
- Para los registros de teléfono (*Org. Teléfono*) que se mostraban como N/A, S/N, no tiene, sin dato o no suministra, se ajustaron a un valor numérico de "0".
- Los datos de municipios (*Org. Munic*) se normalizaron para que coincidieran con la base de datos de departamentos y municipios del DANE (Departamento Administrativo Nacional de Estadística).
- Para los diagnósticos (*Org. Dxs)* se eliminaron tildes y caracteres especiales, en la función *(Estand. Dx*) se redujo el campo a cuatro dígitos, alineándose así con la estructura estándar del CIE10.
- La función de cálculo (*Cal. Cant Pendiente*) se implementó para analizar la diferencia entre la *cantidad prescrita* y la *cantidad entregada*, con el fin de verificar la coherencia de la cantidad pendiente.
- Se organizó la codificación CUMS (Códigos Únicos de Medicamentos) para cumplir con un formato específico de 13 dígitos (*Org Cum)* sustituyendo los registros que estén como *N/A, no aplica, No especificado* por un valor 0.
- La función *(Fil. Med cod)* separara los códigos CUMS que no contienen el carácter guion ("-") que indica la separación entre el código CUM y su respectivo código de consecutivo.
- Los registros que no cumplen esta condición se almacenan en un archivo consolidado de errores (*Errores VE_CV_Códigos CUM*), donde se detallan las

características del error, su descripción, categoría y subcategoría correspondiente (*Add Log Error Estruc*).

- Usó de funciones*: (Partir Cum)* para dividir los registros en código cum y código Consecutivo; (*Conver. Ent Cum),* convierte el código cum en formato texto y el consecutivo en entero; (*Data validator*) para verificar que los códigos cum no excedan los diez dígitos, según la base de datos del Invima. Aquellos que superaron esta longitud se registraron en el archivo de errores consolidado (*Errores VE_CV_Codigos CUM*) a través de las funciones (*Add Log Error No existe)* que identifica las características del error*,* (*Reenombrar Columna*) renombra el atributo *Cum Código* a *Código cum del medicamento* para que coincida con la estructura del consolidado de errores.
- Se emplearon funciones adicionales para ajustar los códigos CUM y sus consecutivos a una longitud específica de ocho dígitos y tres dígitos respectivamente, completando con "0" cuando fuera necesario (*Rec. Cum y Consec*). La función (*Concat CUM y Consec*) concatena estos dos registros junto con el separador guion (-). La función (*Concat_dpto_mun*), concatena los atributos departamento y municipio del afiliado, la función (*Sel. Columnas*), selecciona las columnas finales que pasan al archivo final (*Archivo_Medicamentos_transformado*).

2. Conformidad y Completitud

El flujo de este proceso se visualiza en la Figura 11, y se compone de siete procesos: (i) Caso especial Conformidad Relacional; (ii) Conformidad de Valor; (iii) Conformidad Relacional atributo CUMS; (iv) Conformidad Relacional atributo Nit del prestador*;* (v) Conformidad Relacional atributo Diagnostico Principal; (vi) Conformidad Relacional atributo Diagnostico Relacionado; (vii) Completitud.

- **Caso especial *Conformidad relacional:*** dado que existe una discrepancia en los atributos *departamento y municipio del afiliado* debido a la falta de normalización en los registros, donde algunos están identificados con la descripción del municipio o departamento, mientras que otros están registrados mediante el código correspondiente a la División Político-Administrativa (Divipola). Para ellos, se realiza un

proceso previo para unificar y homologar estos atributos para que concuerden con el código Divipola, así:

- Se renombran los atributos en minúscula (*Sel. Minus*).
- Se ordena la concatenación de los atributos departamento y municipio del afiliado (*Ord. Municipio*).
- Se accede a la base de datos externa de Divipola (*BD Dpto y Munic*).
- Se renombran los atributos con el fin de facilitar su identificación (Sel. *Cod_dpto_mun*).
- Se ordenan los registros por el atributo concatenado (*Ord. Mun Bd*).
- Se emplea la operación Left Outer Join con el propósito de convertir las descripciones de departamentos y municipios a códigos Divipola (*Norm_Municipio*).
- Se filtran los registros que no coincidieron con el join realizado (*Fil. Error Munc*).
- Se detalla el código, descripción, categoría y subcategoría asociados a los errores identificados (*Log Error Munic*).
- Se seleccionan los atributos para que coincidan con el archivo consolidado de errores (*Sel.Cam Munic*).
- Se centraliza la consolidación de errores (*Cons_Errores_VE_Conformidad*).

- ***Conformidad de Valor**,* se encarga de asegurar que el tipo de dato, la longitud, dominio y el formato de cada atributo sean los correctos (*VE_CV_Td_Form_Long_Oblig*). Este proceso, se muestra en detalle en la Figura 11, donde se aplican las adaptaciones definidas en la Tabla 6 por medio de los siguientes pasos:
  - Se verifican los criterios de conformidad de valor definidos por cada atributo (*VE_CV_Td_Form_Long_Oblig*).
  - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes a las inconsistencias identificadas (*Log Error Conformidad Valor*)
  - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel. VE_Conformidad*).

- Se almacenan los errores detectados en un archivo consolidado de errores denominado (*Cons_Errores_VE_Conformidad*).

- ***Conformidad Relacional atributo CUMS**,* para este atributo (CUMS) resaltado en color amarillo (funciones que se agrupan), se ejecutan los siguientes pasos:
  - Se ordenan los registros por el atributo Cums (*Ord. Cum Med*).
  - Se extrae la información de la fuente externa del Invima (*BD Invima*).
  - Los datos se ordenan de acuerdo con el código CUMS (*Ord. Bd Invima*).
  - Se lleva a cabo una validación de correspondencia mediante un Left Outer Join entre los dos atributos (*VA_CR_FK_CUMS*).
  - Se aplica un filtro para separar los atributos que coinciden, permitiendo que continúe el proceso ETL (*Fil. Cum Erróneos*).
  - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error detectado (*Log Error Cum Inexistentes*).
  - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel. Campos Cums*).
  - Se almacenan los errores detectados en un archivo consolidado de errores denominado (*Cons_Errores_VE_Conformidad*).

- ***Conformidad Relacional* atributo *Nit del prestador***, para este atributo (*NITS*) resaltado en color amarillo, se realizan los siguientes pasos:
  - Se ordenan los registros según el atributo *Nit_Prestador* (*Ord. Nit IPS*).
  - Se extraen datos de la fuente interna de la Base de Datos de Ips Contratadas (*IPS contratadas*).
  - Los registros de esta fuente se ordenan según el atributo *NIT del prestador* (*Ord. Nit IPS BD*).
  - Se valida la correspondencia entre ambos conjuntos de datos mediante la aplicación de un Left Outer Join (*VE_CR_FK_NIT*).
  - Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*S. Nit Inexistentes*).
  - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error detectado (*Log Error Nit Inexistentes*).

- Se seleccionan los atributos que se mostrarán en el resumen consolidado de errores (*Sel. Campos Nits*).
- Los errores identificados se almacenan en un archivo consolidado de errores denominado (*Cons_Errores_VE_Conformidad*).

- ***Conformidad Relacional*** **atributo** ***Diagnostico Principal***, para este atributo (*DX PPAL*) resaltado en color amarillo, se realizan los siguientes pasos:
  - Se ordenan los registros por el atributo *Diagnóstico principal* (*Ord. Dx Ppal*).
  - Se extraen los registros de la fuente externa de la Base de Datos de CIE10 (*Códigos CIE10*).
  - Los datos de esta fuente externa se ordenan según el Código de diagnóstico CIE10 (*Ord. Dx bd*).
  - Se valida la correspondencia entre ambos conjuntos de datos mediante la aplicación de un Left Outer Join (*VE_CR_FK_ DX Ppal*).
  - Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (Fil. Dx PPal).
  - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Errores_VA_CR_FK_ DX Ppal y Log Error Cod_Dx_Ppal).*
  - Se Seleccionan los atributos que se mostraran en el consolidado de errores (*Sel. Campos Dx_Ppal)*.
  - Los errores detectados se almacenan en un archivo consolidado de errores *(Cons_Errores_VE_Conformidad).*

- ***Conformidad Relacional*** **atributo** ***Diagnostico Relacionado*****,** para este atributo (*DX RELACIONADO*) resaltado en color amarillo, se realizan los siguientes pasos:
  - Se ordenan los registros por el atributo *Diagnóstico Relacioado* (*Ord. Dx relacionado*).
  - Se extraen los registros de una fuente externa de la Base de Datos de CIE10 (*Códigos CIE10 2*).
  - Los datos de esta fuente externa se ordenan según el Código de diagnóstico CIE10 (*Ord. Dx bd 2*).

- Se valida la correspondencia entre ambos atributos mediante un Left Outer Join (*VE_CR_FK_ DX Relacionado*).
- Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*Fil. Dx relacionado*).
- Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Errores_VA_CR_FK_ DX Relacionado y Log Error Cod_Dx_Relacionado*).
- Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel. Campos Dx_Relacionado*).
- Los errores detectados se almacenan en un archivo consolidado de errores (*Cons_Errores_VE_Conformidad*).

Dado de que los atributos *diagnostico principal* y *diagnostico relacionado* no son objeto de exclusión del reporte principal, los errores que se identificaron en el proceso anterior se consolidan (*Cons_Errores_DX*) y se devuelven los registros al flujo de datos para continuar con la transformación de *Completitud.*

- ***Completitud***, mediante la función (*VE_COMPLETITUD*), se aplican las adaptaciones definidas en la Tabla 8, por medio de los siguientes pasos:
  - Se filtran los errores del atributo *Cantidad total entregada* (*Cant. Total Entregada*), *Fecha de entrega de los medicamentos* (*Fecha Entrega*) y *Diagnóstico relacionado* (*Diagnostico*).
  - Se detalla el código, descripción, categoría y subcategoría asociados a los errores identificados (*Log Error Cant Med, Log Error Fec Entrega, Log Error DX*)
  - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel. Cant, Sel. Fec_Entrega y Sel. DX Col*).
  - Los errores detectados se almacenan en un archivo consolidado de errores *(Cons_Errores_VE_Conformidad*).

3. Plausibilidad de Unicidad, Atemporal y Temporal

La transformación de Plausibilidad (ver Figura 12), se compone de nueve procesos: Plausibilidad de Unicidad, Plausibilidad Atemporal para tipo de documento, departamento y municipio del afiliado y cantidad entregada de medicamento, y Plausibilidad Temporal

para fecha de entrega de medicamento, fecha de autorización, Cantidad de fórmulas en el mismo tiempo duración del tratamiento y Cantidades Entregadas y Pendientes Totales.

- ***Plausibilidad de Unicidad***, en este proceso se realizan los siguientes pasos:
    - Se verifica el tipo de dato de cada uno de los atributos del archivo plano generado en la transformación anterior (*Conv String Td y ID Arc*).
    - Se concatenan los atributos según la adaptación definida en la Tabla 8 (*Concat Unicidad*).
    - Se realiza un conteo de los registros duplicados (*Conteo Reg Unicidad*).
    - Se ordenan en función de la cantidad de registros duplicados encontrados, clasificándolos de mayor a menor (*Ord Unicidad*).
    - Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*Sin_Duplicados*).
    - Se ordenan por la concatenación realizada (*Ord Unicidad Dupl*).
    - Los registros que no son duplicados continúan el flujo
    - Se ejecuta un left outer join entre los dos atributos concatenados para identificar los registros que continúan en la transformación (*VE_Plaus_Unicidad*).
    - Los registros duplicados se almacenan en un archivo independiente del consolidado de errores (*Errores_Unicidad*).
    - Se detalla el código, descripción, categoría y subcategoría asociados a los errores identificados (*Log Error Unicidad*).
    - Se seleccionan los atributos que se mostrarán en el archivo de errores (*Sel. Unicidad_error).*
    - Los errores detectados se almacenan en un archivo de errores *(Errores_Unicidad).*

- ***Plausibilidad Atemporal para tipo de documento*****,** en este proceso se realizan los siguientes pasos de acuerdo con las adaptaciones definidas en la Tabla 9:
    - Se ordenan los registros del flujo por la unión de los atributos Tipo de documento y Número de identificación (*Ord_Td_Id_Arc*).
    - Se extraen los registros de una fuente interna de la Base de Datos de afiliados *(Bd_afiliados).*
    - Se eliminan los espacios en blanco de la fuente interna (*Reemp. Espacios BD*).

- Se convierten los atributos para que coincidan con el tipo de datos del archivo principal (*Conv String Td y ID BD).*
- Se ordenan los registros de la fuente interna por los atributos *Tipo de documento y Número de identificación* (*Ord. TI_ID_Bd*).
- Se valida la correspondencia entre ambos atributos mediante un Left Outer Join (*VE_TI_ID_Afiliados*).
- Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*Fil. Afiliados Cruce*).
- Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Log Error Afiliados Inexistentes*).
- Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel. Afiliado*).
- Los errores detectados se almacenan en un archivo consolidado de errores *(Cons. Errores_plausibilidad*).

- ***Plausibilidad Atemporal para departamento y municipio del afiliado*****,** en este proceso se realizan los siguientes pasos de acuerdo con las adaptaciones definidas en la Tabla 9:
  - Se utiliza la misma extracción de los registros de una fuente interna de la Base de Datos de afiliados *(Bd_afiliados*).
  - Se valida la correspondencia entre ambos atributos mediante un Left Outer Join (*VE_Cod_Mun_Afiliado*).
  - Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*Fil. Error Cod Mun*).
  - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Log Error Mun Inexistentes*).
  - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel Municipios*).
  - Los errores detectados se almacenan en un archivo consolidado de errores *(Cons. Errores_plausibilidad*).

- ***Plausibilidad Atemporal para cantidad entregada de medicamento*****,** en este proceso se realizan los siguientes pasos de acuerdo con las adaptaciones definidas en la Tabla 9:
    - Se realiza la verificación del tipo de datos entero a los atributos *cantidad_prescrita, cantidad_entregada, rect_cant_pendiente* en *(Bd_afiliados)*
    - Se verifica por medio de la función formula las adaptaciones definidas por cada registro (*VE_P.Atemp_Cant Entregada*).
    - Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*Fil. CantEntregadas*).
    - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Log Error Cant Entregada*).
    - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel Cant. Entregada*).
    - Los errores detectados se almacenan en un archivo consolidado de errores *(Cons. Errores_plausibilidad).*

- ***Plausibilidad Temporal para fecha de entrega*****,** en este proceso se realizan los siguientes pasos de acuerdo con las adaptaciones definidas en la Tabla 10:
    - Se verifica por medio de la función formula las adaptaciones definidas por cada registro (*VE_Fec_Entrega*).
    - Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*Filter VE_Fec Entrega*).
    - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Log Error Fecha Entrega).*
    - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel Fecha Entrega).*
    - Los errores detectados se almacenan en un archivo consolidado de errores *(Cons. Errores_plausibilidad).*

- ***Plausibilidad Temporal para fecha de autorización*****,** en este proceso se realizan los siguientes pasos de acuerdo con las adaptaciones definidas en la Tabla 10:
    - Se verifica por medio de la función formula las adaptaciones definidas por cada registro (*VE_Fec_Autorización*).

- Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*Filter VE_Fec Autorización*).
- Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Log Error Fecha Autorización*).
- Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel Fecha Autorización*).
- Los errores detectados se almacenan en un archivo consolidado de errores *(Cons. Errores_plausibilidad*).

- ***Plausibilidad Temporal para Cantidad de fórmulas en el mismo tiempo duración del tratamiento***, en este proceso se realizan los siguientes pasos de acuerdo con las adaptaciones definidas en la Tabla 10:
  - Se realiza la concatenación de los atributos definidos (*Concat Td_ID_Cum_Diastto).*
  - Se realiza la sumatoria (*Sum_fecPres_+_Dias_tto*) de los atributos (*fecha_de_prescripción y dias de tratamiento*).
  - Se ordenan los registros concatenados (*Ord.Concatenado*).
  - Se realiza el conteo de los registros duplicados (*Conteo Cantidad Reg Duplicados*).
  - Se ordenan los registros por el atributo *cantidad total pendiente*. (*Ord_Cant_Duplicados*).
  - Se verifican los registros duplicados mediante formula (*VE_mayor_dos_Registros*).
  - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Log Error Cantidad Formulas*).
  - Se seleccionan los atributos que se mostrarán en el archivo de errores (*Sel Cantidad Formulas*).
  - Los errores detectados se almacenan en un archivo de errores *(Cons. Errores_plausibilidad Temporal*).

Los registros que cumplieron con los criterios continúan el flujo de la transformación asi:

- Se realiza una copia de los registros del flujo desde el momento que se aplicó la sumatoria (*Reg Sin duplicados*).

- Se ordenan los registros por el atributo concatenado (*Ord. Concatenado_dias_tto*).
- Se ejecuta un left outer join entre los dos atributos para que los registros que no tienen inconsistencias continúen el flujo (*VE_PA_dias_tto*).
- Se seleccionan los atributos que se mostrarán en el archivo de errores (*Sel Cantidad Formulas).*
- Los errores detectados se almacenan en un archivo de errores (*Archivo final).*

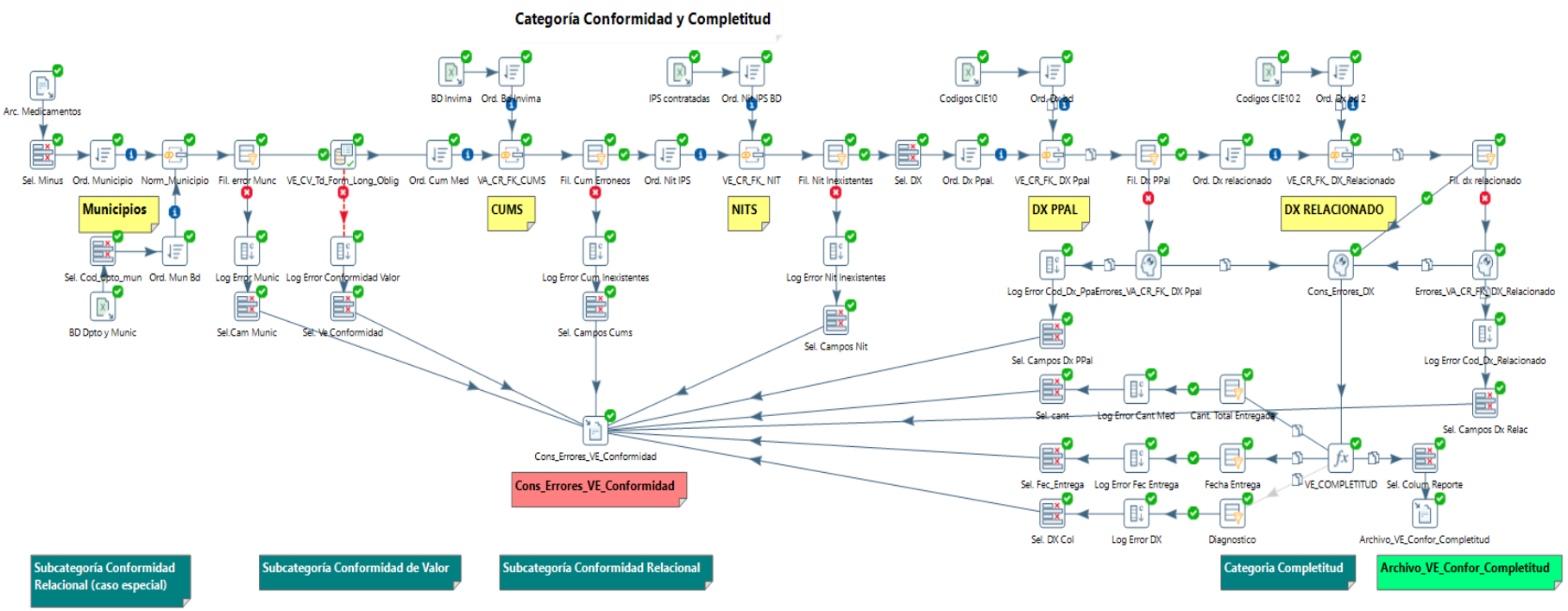


Figura 11 Conformidad de Valor, Relacional y Completitud Medicamentos

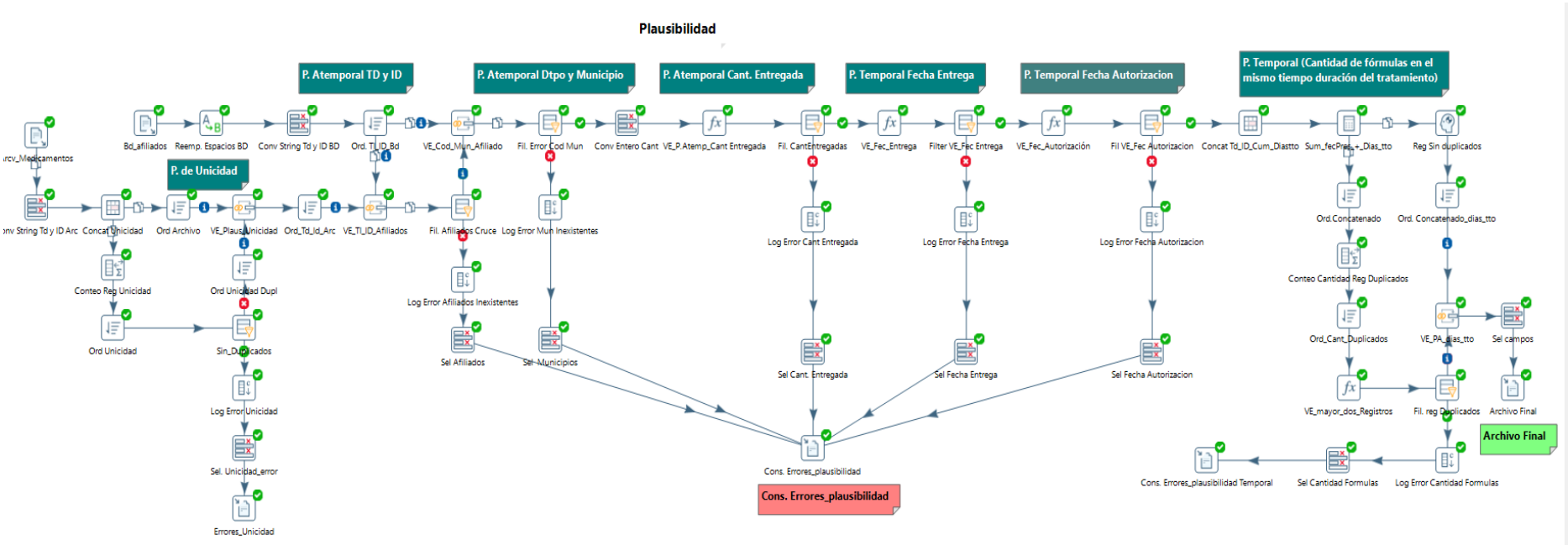


Figura 12 Plausibilidad Medicamentos

- ***Plausibilidad Temporal Cantidades Entregadas y Pendientes Totales,*** en la Figura 13, se muestra el detalle la transformación realizada de acuerdo a la adaptación definida en la Tabla 10. Esta transformación independiente fue necesaria debido a las limitaciones de espacio de memoria en la herramienta Pentaho, lo que llevó a la creación de un paquete adicional para gestionarla, esta transformación implica la comparación de dos reportes mensuales (enero y febrero de 2021) como se detalla en los siguientes pasos:
    - Se extrae los registros de los archivos planos de medicamentos de los meses enero (*1604 enero*) y de febrero (*1604 febrero*).
    - Se concatenan los registros de ambos archivos por los atributos *tipo de documento, número de identificación, Código cum y fecha de prescripción (Concat Td_id_cum_fpres_Febrero)* y *(Concat Td_id_cum_fpres_enero)* respectivamente.
    - Se ordenan los registros por la concatenación realizada para ambos archivos (*Ord Febrero*) y (*Ord Enero*) respectivamente.
    - Se ejecuta un Left Join (*Ve_Pt_Td_Id_Cum_Fecpres*) con los atributos concatenados para identificar registros coincidentes.
    - Se seleccionan las columnas que continúan en el flujo de transformación (*Select Col*).
    - Se filtran los registros que no coinciden (*Fil Cant Presc*) y se almacenan en el archivo final (*Archivo Final*).
    - Los registros que coincidieron se realiza la verificación de cantidades entre los atributos *cantidad pendiente* del archivo del flujo y *cantidad entregada* del reporte comparativo (*VE_Cant_pend_2reportes*), si el registro cumple se registra en la columna que el registro esta OK, de lo contrario, se registra un mensaje de error.
    - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Log Error P. Temporal Cantidades*).
    - Se realiza un filtro (*Fil. Ok*) de los registros con mensaje de error y se almacenan en un archivo de errores (*VE_PT_Errores_Cantidades_Entregadas*).
    - Los registros que no presentaron inconsistencias, se seleccionan los atributos que se mostrarán en el archivo de errores (*Select Col Def*).
    - Los errores detectados se almacenan en un archivo de errores (*Archivo final*).

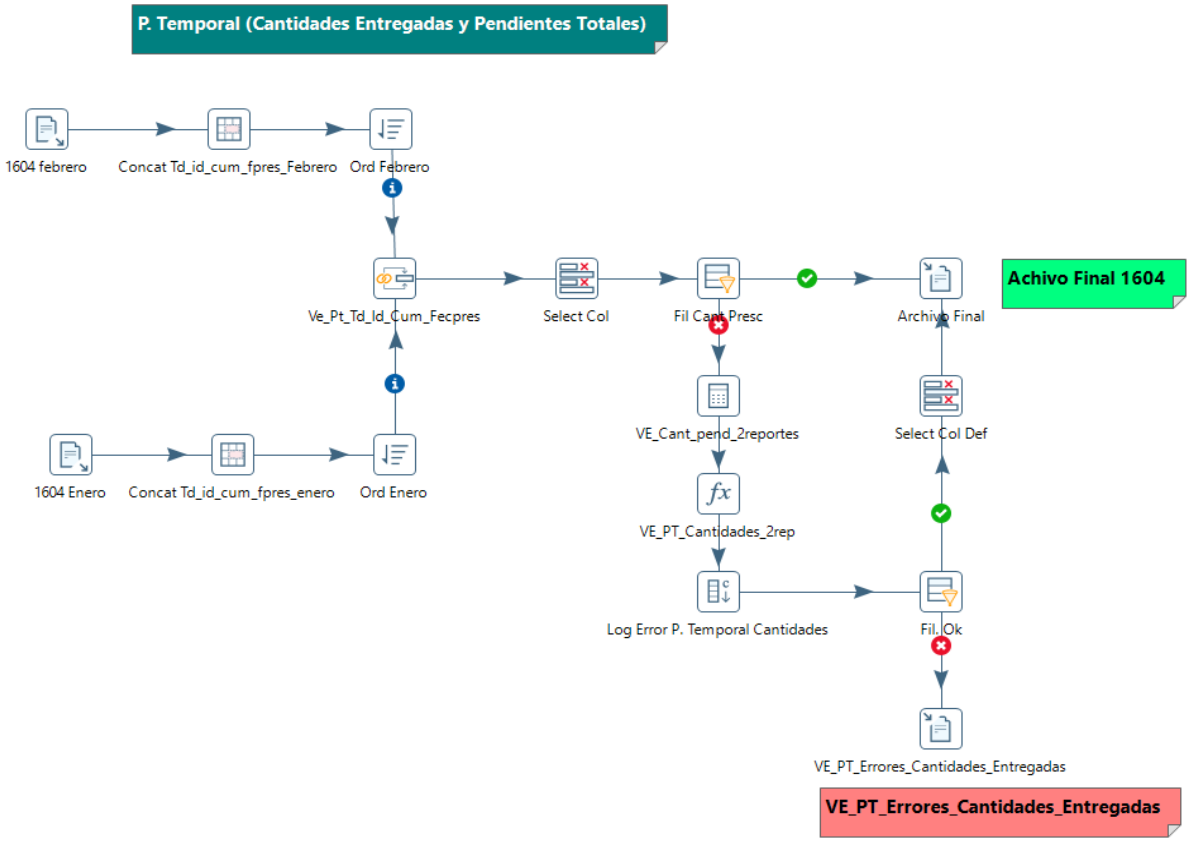


Figura 13 Plausibilidad Temporal Medicamentos

4. Consolidado de log de errores

Finalmente, como parte integral de todo el proceso de ejecución para cada categoría y subcategoría del proceso ETL, (ver Figura 14) se lleva a cabo la consolidación de los registros de errores de cada transformación en un paquete independiente. Este paso se realiza con el propósito de poder agrupar, visualizar, analizar y evaluar los resultados.

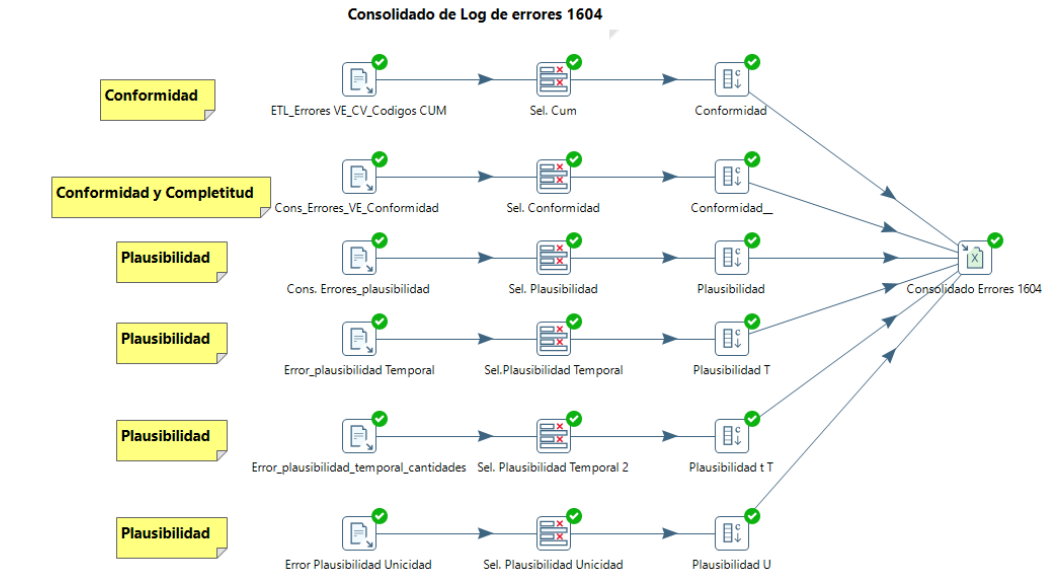


Figura 14 Consolidado log de errores Medicamentos

5.2.2 Asignación de citas médicas

En la implementación del proceso ETL para el reporte de asignación de citas médicas (ver Figura 15), el flujo de trabajo de la implementación consta de cuatro transformaciones que son: (i) Limpieza y Normalización de los datos, (ii) Conformidad y Completitud, (iii) Plausibilidad y (iv) Consolidado de log de errores.

Al igual que para el reporte de entrega de medicamentos, cada una de estas transformaciones está identificada por etiquetas que ofrecen información específica: (i) *Etiquetas Amarillas*, el atributo al que están relacionadas dentro de la transformación. (ii) *Etiquetas Rojas*, la función donde se almacena el consolidado de errores. (iii) *Etiquetas Verde Claro,* el archivo final que concluye la transformación. (iv) *Etiquetas Verde Oscuro*, representan la categoría o subcategoría que se está aplicando en cada proceso.

Figura 15 Proceso ETL Citas Médicas

1. Limpieza y Normalización de los datos

El flujo de esta transformación se visualiza en la Figura 16, y se compone de cuatro procesos: Corrección de errores tipográficos y de formato, Detección y manejo de valores

nulos, Estandarización y organización de valores y Manejo de mayúsculas/minúsculas y acentos.

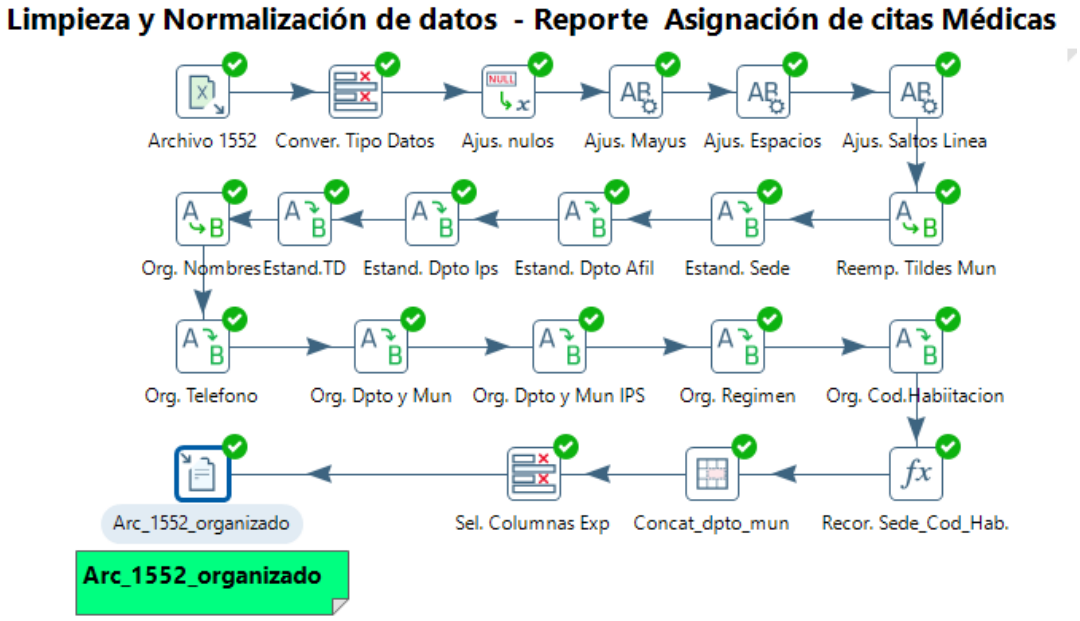


Figura 16 Limpieza, normalización de datos Citas Médicas

- **Corrección de errores tipográficos y de formato**: se verifican y/o ajustan los atributos al tipo de datos establecidos (*Conver. Tipo Datos*) de acuerdo con la adaptación propuesta como se muestra en la Tabla 11. Por ejemplo, se asegura que *Número de identificación, departamento del afiliado, servicio solicitado* estén en tipo de dato String y los atributos como *fecha en el que el usuario solicita la cita o cirugía* estén en un tipo de datos date y en un formato de fecha 'dd-mm-yyyy'.

- **Detección y manejo de valores nulos:** se realizaron ajustes en los valores nulos (*Ajus. nulos*), por ejemplo, *departamento del afiliado, municipio del afiliado, tipo de documento* se reemplaza por "No registra", el Código de servicio se pone "0".

- **Manejo de mayúsculas/minúsculas:** todos los registros se convirtieron a mayúsculas (*Ajus. Mayus*) para lograr uniformidad en los datos. Además, se procedió a eliminar espacios en blanco (*Ajus. Espacios)*, saltos de línea (*Ajus. Saltos*

Linea) en atributos de campo de texto, como nombres, apellidos, municipios y departamentos y ajuste en las tildes del atributo municipios (*Reemp. Tildes Mun).*

- **Estandarización y organización de valores,** se realizaron ajustes para estandarizar ciertos atributos durante la transformación. Por ejemplo:
    - Sede departamental del reporte (Estand. Sede) transformando *Valle* en *Valle del cauca*.
    - Tipo de documento (*Estand.TD*) modificando los registros como *cedula* a *CC*, *Registro Civil* a *RC.*
    - Código de habilitación de la sede (*Recor. Sede_Cod_Hab*.) recorta a 10 dígitos el código para que coincida posteriormente con la validación con una fuente externa (REPS).

  Además, otras transformaciones como:
    - Se realiza la concatenación del departamento y municipio para que en la siguiente transformación, coincida con la validación con la bd externa Dane (*Concat_dpto_mun*).
    - Se seleccionan las columnas que conformarán la estructura del archivo final (Sel. Columnas Exp)
    - Se almacenan los registros finales del archivo (*Arc_1552_organizado*).

1. Conformidad y Completitud.

El flujo de esta transformación se muestra en detalle en la Figura 17, y se compone de siete procesos: (i) Caso especial Conformidad Relacional; (ii) Conformidad de Valor; (iii) Conformidad Relacional atributo CUMS; (iv) Conformidad Relacional atributo Nit del prestador*;* (v) Conformidad Relacional atributo habilitación; (vi) Conformidad Relacional atributo Código de servicio; (vii) Completitud.

- **Caso especial *Conformidad relacional:*** al igual que el reporte de medicamentos este archivo presenta una discrepancia en los atributos *departamento y municipio del afiliado* debido a la falta de normalización en los registros, donde algunos están identificados con la descripción del municipio o departamento, mientras que otros están registrados mediante el código correspondiente a la División Político-Administrativa

(Divipola), para este reporte se realiza el mismo proceso realizado para el reporte de entrega de medicamentos.

- ***Conformidad de Valor**,* en este proceso, se aplican las adaptaciones definidas en la Tabla 11, por medio de los siguientes pasos:
    - Se verifica el tipo de datos de los atributos del archivo (*Sel. Col*).
    - Se verifican los criterios de conformidad de valor definidos por cada atributo (*VE_CV_Td_Form_Long_Oblig*).
    - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error detectado (*Log Error CV*).
    - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel. VE_Conformidad*).
    - Se almacenan los errores detectados en un archivo consolidado de errores denominado (*Consolidado Errores*).

- ***Conformidad Relacional atributo Nit**,* para este atributo (*Nit*) resaltado en color amarillo (funciones que se agrupan), se ejecutan los siguientes pasos:
    - Se ordenan los registros del flujo por el atributo *Nit* (*Ord. Nit IPS*).
    - Se extrae la información de la fuente interna IPS contratadas (*IPS contratadas*).
    - Se ordenan los registros de la fuente interna por el atributo *Nit* (*Ord. Nit IPS BD*).
    - Se lleva a cabo una validación de correspondencia mediante un Left Outer Join entre los dos atributos (*VE_CR_FK_ NIT*).
    - Se aplica un filtro para separar los atributos que coinciden, permitiendo que continúe el proceso ETL (*Fil. Nit Inexistentes*).
    - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error detectado (*Log Error Nit Inexistentes).*
    - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel. Campos Nit)*.
    - Se almacenan los errores detectados en un archivo consolidado de errores denominado *(Consolidado Errores).*

- ***Conformidad Relacional atributo habilitación**,* para este atributo (*Habilitación*) resaltado en color amarillo (funciones que se agrupan), se ejecutan los siguientes pasos:
    - Se ordenan los registros por el atributo *Código de habilitación sin sede* (*Ord_Cod_hab_Arc*).
    - Se extrae la información de la fuente externa REPS (*Reps_Sedes*).
    - Se ordenan los registros de la fuente externa por el atributo *Código de habilitación* (*Ord_Cod_Sede*).
    - Se realiza la validación de correspondencia mediante un Left Outer Join entre los dos atributos (*VE_CR_FK_ HAB_SEDE*).
    - Se aplica un filtro para separar los atributos que coinciden, permitiendo que continúe el proceso ETL (*Fil. Hab Inexistente*).
    - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error detectado (*Log Error Sede Inexistentes).*
    - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel. Campos Sede)*.
    - Se almacenan los errores detectados en un archivo consolidado de errores denominado *(Consolidado Errores).*

- ***Conformidad Relacional atributo Código de servicio**,* para este atributo (*Código Servicio*) resaltado en color amarillo (funciones que se agrupan), se ejecutan los siguientes pasos:
    - Se ordenan los registros por el atributo *Código de habilitación sin sede* (*Ord Cod Servicio*).
    - Se extrae la información de la fuente externa REPS (*Cod Servicio Reps*).
    - Se ordenan los registros de la fuente externa por el atributo *Código de servicio* (*Ord Cod Servicio Reps*).
    - Se lleva a cabo una validación de correspondencia mediante un Left Outer Join entre los dos atributos (*VE_CR_FK_ COD_SERVICIOS*).
    - Se aplica un filtro para separar los atributos que coinciden, permitiendo que continúe el proceso ETL (*Fil. Cod_Serv Inexistente*).

- Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error detectado (*Log Error Cod_Serv_Inexistente*).
- Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel. Campos Servicio*).
- Se almacenan los errores detectados en un archivo consolidado de errores denominado (*Consolidado Errores*).

- ***Completitud***, mediante la función (*VE_COMPLETITUD*), se aplican las adaptaciones definidas en la Tabla 12, por medio de los siguientes pasos:
  - Se filtran los errores del atributo *Cantidad total entregada* (*Cant. Total Entregada*).
  - Se filtran los registros que no cumplieron con la adaptación (*Fecha Entrega*).
  - Se detalla el código, descripción, categoría y subcategoría asociados a los errores identificados (*Log Error Completitud*).
  - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel. Campos Completitud*).
  - Los errores detectados se almacenan en un archivo consolidado de errores (*Consolidado Errores*).
  - Para los registros que no presentaron inconsistencia, se seleccionan los atributos que se mostraran en el archivo (*Sel. Colum Reporte*) y por último se almacenan en el archivo final (*archivo final*).

2. Plausibilidad

La transformación de Plausibilidad (ver Figura 18), se compone de cinco procesos: Plausibilidad de Unicidad, Plausibilidad Atemporal para tipo de documento, departamento y municipio del afiliado y Plausibilidad Temporal para fecha de asignación de cita y fecha que solicita sea asignada.

- ***Plausibilidad de Unicidad***, en esta transformación se realizan los siguientes pasos:
  - Se ordenan los registros por el atributo *Numero de identificación* (*Ord. Id Arc*).
  - Se verifican los registros de acuerdo con la adaptación definida por cada atributo (*P. Unicidad*).

- ***Plausibilidad Atemporal para tipo de documento*****,** en este proceso se realizan los siguientes pasos de acuerdo con las adaptaciones definidas en la Tabla 13:
    - Se verifican que los atributos identificados en la adaptación estén acordes con el tipo de dato (*Conv String Td y ID Arc*).
    - Se ordenan los registros del flujo por la unión de los atributos Tipo de documento y Número de identificación (*Ord_Td_Id_Arc*).
    - Se extraen los registros de una fuente interna de la Base de Datos de afiliados *(Bd_afiliados).*
    - Se eliminan los espacios en blanco de la fuente interna (*Reemp. Espacios BD*).
    - Se convierten los atributos para que coincidan con el tipo de datos del archivo principal (*Conv String Td y ID BD).*
    - Se ordenan los registros de la fuente interna por los atributos *Tipo de documento y Número de identificación* (*Ord. TI_ID_Bd*).
    - Se valida la correspondencia entre ambos atributos mediante un Left Outer Join (*VE_TI_ID_Afiliados*).
    - Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*Fil. Afiliados Cruce*).
    - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Log Error Afiliados Inexistentes*).
    - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel. Afiliado*).
    - Los errores detectados se almacenan en un archivo consolidado de errores *(Cons. Errores_plausibilidad*).

- ***Plausibilidad Atemporal para departamento y municipio del afiliado*****,** en este proceso se realizan los siguientes pasos de acuerdo con las adaptaciones definidas en la Tabla 14:
    - Se utiliza la misma extracción de los registros de una fuente interna de la Base de Datos de afiliados *(Bd_afiliados*).
    - Se valida la correspondencia entre ambos atributos mediante un Left Outer Join (*VE_Cod_Mun_Afiliado*).

- Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*Fil. Error Cod Mun*).
- Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Log Error Mun Inexistentes*).
- Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel Municipios*).
- Los errores detectados se almacenan en un archivo consolidado de errores *(Cons. Errores_plausibilidad).*

- ***Plausibilidad Temporal para fecha de asignación*****,** en este proceso se realizan los siguientes pasos de acuerdo con las adaptaciones definidas en la Tabla 15:
  - Se verifica por medio de la función formula las adaptaciones definidas por cada registro (*VE_Fec_Asig_Cita*).
  - Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*Filter Fecha_asig*).
  - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Log Error Fec_Asignacion*).
  - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel Columnas Fecha Asigno Cita*).
  - Los errores detectados se almacenan en un archivo consolidado de errores *(Cons. Errores_plausibilidad).*

- ***Plausibilidad Temporal para fecha de solicita sea asignada*****,** en este proceso se realizan los siguientes pasos de acuerdo con las adaptaciones definidas en la Tabla 15:
  - Se verifica por medio de la función formula las adaptaciones definidas por cada registro (*VE_Fec_Sea_Asig_Cita*).
  - Se aplica un filtro para separar los atributos que coinciden de los que no coinciden (*Filt. Fec Sea Asignada*).
  - Se detalla la descripción, campo, código, categoría y subcategoría correspondientes al error encontrado (*Log Error Fec_Sea_Asignada_cita*).
  - Se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel Columnas Fecha Sea Asignada*).

- Los errores detectados se almacenan en un archivo consolidado de errores *(Cons. Errores_plausibilidad*).
- Para los registros que no presentaron inconsistencias, se seleccionan los atributos que se mostrarán en el consolidado de errores (*Sel Columnas*), y se almacenan en un archivo final *(Arc_Final_1552*).

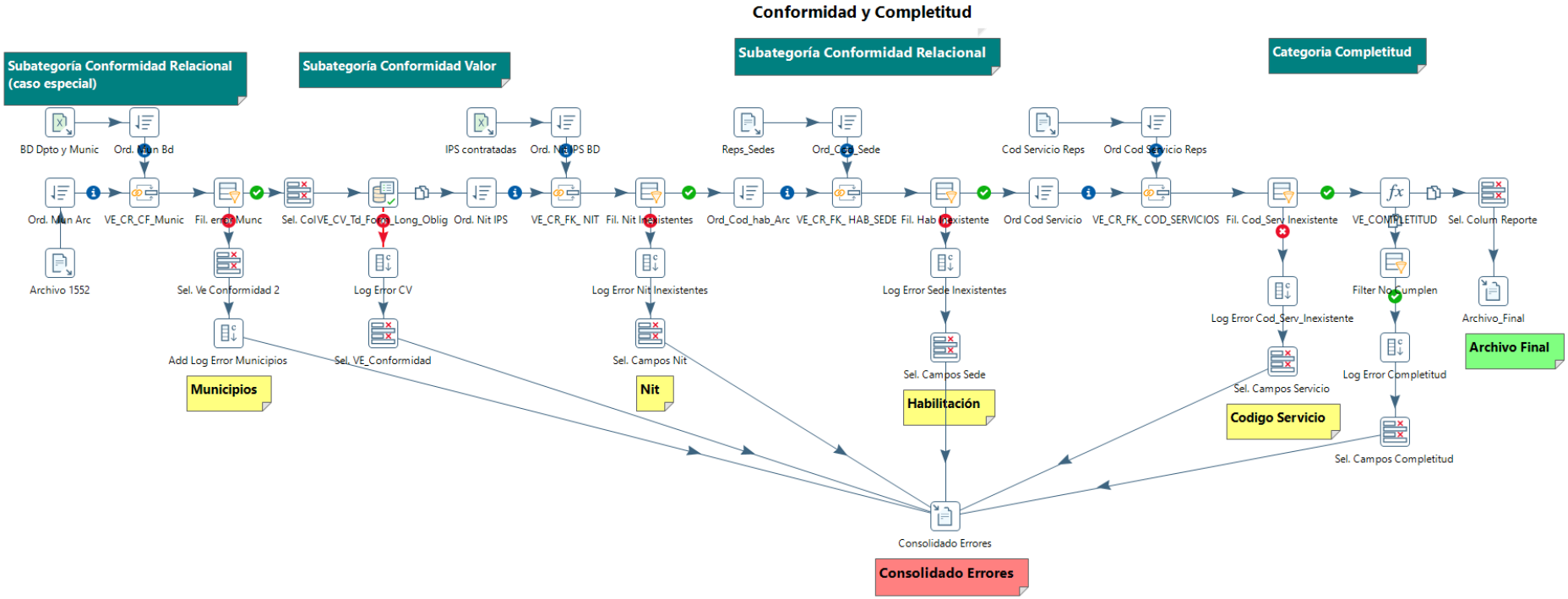


Figura 17 Conformidad de Valor, Relacional y Completitud Citas Médicas

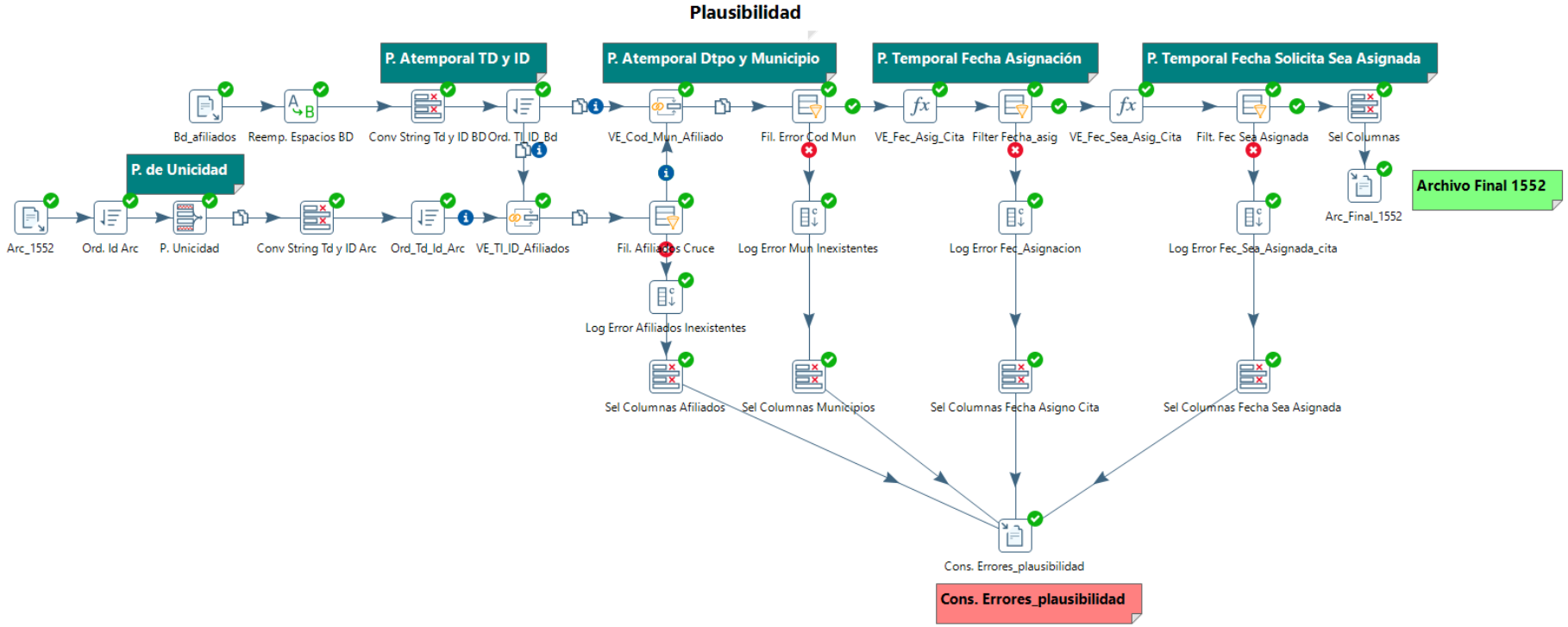


Figura 18 Plausibilidad Citas Médicas

3. Consolidado de log de errores

Finalmente, como parte integral de todo el proceso de ejecución para cada categoría y subcategoría del proceso ETL, (ver Figura 19) se lleva a cabo la consolidación de los registros de errores de cada transformación en un paquete independiente. Este paso se realiza con el propósito de poder agrupar, visualizar, analizar y evaluar los resultados.

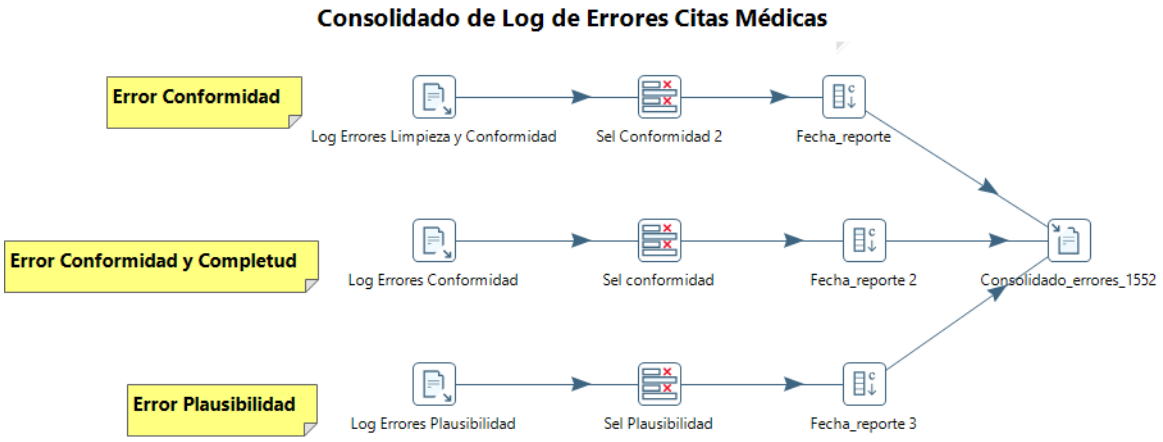


Figura 19 Consolidado log de errores Citas Médicas

CAPITULO 6

6 EVALUACIÓN

En este capítulo se presenta la validación de la adaptación propuesta por medio de un grupo focal y la evaluación de las DQ en el proceso ETL por medio del indicador de Confianza.

6.1 Evaluación grupo focal de expertos

El proceso de validación de la adaptación de las categorías de DQ planteadas por Khan se llevó a cabo a través de un grupo focal de expertos, el cual permite obtener percepciones detalladas y enriquecedoras de un grupo específico en relación con un área de interés [21]. Los pasos para realizar un grupo focal, según autores como [21] y [22] están definidos en cuatro pasos que son:

1. Planificación del grupo focal.
2. Selección de participantes
3. Preparación de materiales y diseño del cuestionario
4. Conducción de la sesión del Grupo Focal
5. Análisis de la información y reporte de resultados.

6.1.1 Planificación del grupo focal

En este paso se aborda la planificación del grupo focal, identificando los recursos clave que se requieren para su ejecución, como son: el objetivo del grupo focal, criterios para la selección de los participantes y la preparación de los materiales necesarios junto con los procedimientos que guiarán la realización del grupo focal.

1. Objetivo del grupo focal

Conocer la opinión de expertos en el sector Salud respecto a la adaptación de las categorías de DQ de Kahn a los reportes de entrega de medicamentos y asignación de cita medicas generados por las IPS.

2. Criterios de selección de los participantes

Para la selección de los expertos que participarán en el grupo focal, se consideraron los siguientes criterios:

- Tener un sólido conocimiento en el sector de la salud, con énfasis en medicamentos y citas médicas.
- Estar activos y en ejercicio en el sector de la salud.
- Contar con una experiencia laboral de más de 10 años en el campo de la salud.

6.1.2 Selección de los participantes

Una vez identificado el perfil de los expertos, se procedió a enviar la invitación formal vía correo electrónico, adjuntando el material de lectura que detallaba la adaptación realizada, sus objetivos y la relevancia del estudio. En la Tabla 16 se proporciona información de los expertos que fueron seleccionados para participar en esta sesión.

Tabla 16 Expertos en salud

Experto ID	Ocupación	Estudios
E1	Vicepresidente de servicios de salud, Docente Catedrático de postgrados.	Médico cirujano, Especialista Gerencia de Servicios de, Especialista Gerencia de Calidad y auditoria en Salud.
E2	Asesor experto en desarrollo de productos de salud, vicepresidente de salud.	Médico cirujano, Especialista Gerencia de Servicios de Salud, Especialista Gerencia de Calidad y auditoria en Salud
E3	Profesional Servicios de Salud Nacional, Medico General	Médico y cirujano, Especialista en Administración Hospitalaria, Especialista en Auditoria y Garantía de la Calidad con Énfasis en Epidemiología EAN, Conferencista a Nivel Nacional en temas de Interés en el Campo de la Administración en Salud, Asesor de Hospitales y E.S.E.S. en el Cauca y Valle del Cauca.

Fuente: Elaboración propia

6.1.3 Preparación de materiales y diseño del cuestionario

Se detalla la preparación de los materiales y diseño del cuestionario que guiarán el desarrollo del grupo focal.

a) Estructura del protocolo del grupo focal

Con el objetivo de asegurar coherencia y un enfoque apropiado durante la sesión del grupo focal, se estableció una estructura de protocolo como se muestra en la Tabla 17.

Tabla 17 Protocolo del grupo focal elemento descripción

Elemento	Descripción
Agenda de trabajo	Secuencia de actividades a realizar por cada participante durante la sesión.

Elemento	Descripción
Propuesta a evaluar	Documento que describe la adaptación propuesta de las categorías de DQ para los reportes de entrega de medicamentos y asignación de citas médicas.
Cuestionario	Preguntas diseñadas para extraer información relevante para la validación de la adaptación.

Fuente: Elaboración propia

b) Elementos empleados

La planificación del grupo focal incluye detalles clave sobre la sesión, como se muestra en la Tabla 18:

Tabla 18 Elementos para realizar el grupo focal

Elemento	Información
Fecha de realización:	06 Julio 2023
Hora de Inicio:	06:00 pm
Duración:	2 horas
Lugar:	Online – Vía Google Meets
Tema:	Validación de la adaptación de las categorías de DQ para los reportes de entrega de medicamentos y asignación de citas medicas
Moderadora:	Daisy Yisel Meneses Lopez – Conectada desde Popayán
Relatora:	PhD. Martha Eliana Mendoza Becerra – conectada desde Popayán
Expertos:	E1: Dr. Hugo Moreno Reales – conectado desde Bogotá E2: Dr. Xavier Navarro – conectado desde Popayán E3: Dr. Elkin Javier Idárraga – conectado desde Popayán

Fuente: Elaboración propia

c) Métodos de captura y registro de información

La ejecución del grupo focal se llevará a cabo utilizando los siguientes métodos de captura y registro:

- ✓ Cuestionario de evaluación para recopilar información relevante.
- ✓ Grabación en video de toda la reunión para capturar las interacciones y los comentarios de los expertos.

d) Métodos de análisis de la información

Después del grupo focal, se realizará un análisis detallado de la información recopilada utilizando dos métodos:

- ✓ Análisis estadístico de la información extraída de los cuestionarios.
- ✓ Análisis cualitativo de las observaciones y oportunidades de mejora identificadas durante la sesión.

e) Definición del cuestionario

En la Tabla 19, se presenta el cuestionario definido para validar la adaptación de la adaptación de las subcategorías de DQ en los reportes de entrega de medicamentos y asignación de citas médicas. Este cuestionario se diseñó usando la escala de Likert para una evaluación más precisa y se estructura en dos sesiones: la primera incluye preguntas cerradas dirigidas a las categorías y subcategorías con el fin de valorar el grado en que adecua la adaptación realizada a cada reporte, mientras que la segunda consta de preguntas abiertas diseñadas para recopilar información adicional y detallada. Estas últimas permiten expresar comentarios, ofrecer sugerencias o compartir cualquier otra observación relevante.

Tabla 19 Preguntas cuestionario de evaluación de expertos

	Preguntas Cerradas					
ID	**¿Considera adecuada la adaptación de las subcategorías de DQ realizada para el reporte?**	**5**	**4**	**3**	**2**	**1**
2	Conformidad de valor	O	O	O	O	O
3	Conformidad relacional	O	O	O	O	O
4	Completitud	O	O	O	O	O
5	Plausibilidad de Unicidad	O	O	O	O	O
6	Plausibilidad de Atemporal	O	O	O	O	O
7	Plausibilidad de Temporal	O	O	O	O	O
8	**Preguntas Abiertas**					
9	¿Considera que se deben agregar criterios a la adaptación de alguna de las subcategorías de DQ al reporte? En caso afirmativo ¿cuáles y por qué?					
10	¿Considera que se deben eliminar criterios a la adaptación de alguna de las subcategorías de DQ al reporte? En caso afirmativo ¿cuáles y por qué?					
11	¿Tiene algún comentario adicional acerca de la adaptación de las categorías de DQ presentado?					
12	¿Tiene alguna sugerencia adicional acerca de la adaptación de las categorías de DQ presentado?					

Nomenclatura: Completamente de Acuerdo (5), De Acuerdo (4), Ni de Acuerdo Ni en Desacuerdo (3), En Desacuerdo (2), Completamente en Desacuerdo (1)

6.1.4 Conducción de la sesión del Grupo Focal

La creación de la sesión del grupo focal sigue una estructura ordenada, diseñada para asegurar una interacción fluida y efectiva entre los participantes. El proceso se ilustra en la Figura 20, organizada de acuerdo con las siguientes fases:

- **Introducción**: se inicia con una bienvenida a los participantes, se presenta la agenda y se delinean los objetivos de la sesión.

- **Adaptación Reporte Medicamentos**: se presenta en detalle la adaptación propuesta de las DQ de Kahn para los reportes de entrega de medicamentos.
- **Evaluación a través de Cuestionario para el reporte de Medicamentos:** se brinda la oportunidad a los expertos de completar un cuestionario web que proporcionará datos cuantitativos y percepciones sobre la adaptación.
- **Adaptación de Reportes de Asignación de Citas Médicas**: se realiza un proceso similar de adaptación con respecto a los reportes de asignación de citas médicas.
- **Evaluación a través de Cuestionario para el reporte de Asignación de Citas Médicas:** de la misma forma para el reporte de Medicamentos se realiza la evaluación por parte de los expertos.
- **Discusión y Comentarios Finales**: se abre un espacio para que los participantes compartan sus opiniones, hagan preguntas y brinden comentarios adicionales.

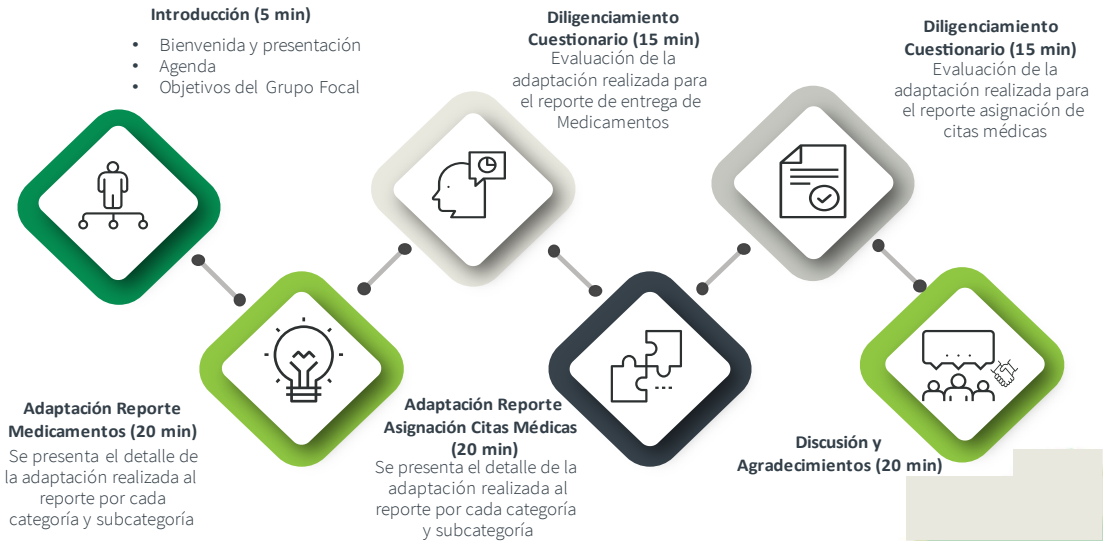


Figura 20 Sesión grupo focal

6.1.5 Análisis de la información y reporte de resultados

Después de la realización del grupo focal, se procedió a analizar los aportes obtenidos tanto de los cuestionarios completados por los expertos como de las discusiones llevadas a cabo durante la sesión. Esta combinación de datos proporciona una comprensión integral de las

percepciones y opiniones de los expertos en relación con la adaptación de las categorías de DQ a los reportes de entrega de medicamentos y asignación de citas médicas.

En esta sección, se analizaron las respuestas proporcionadas por los expertos destacando los aspectos más relevantes que surgieron durante la validación. Estos resultados proporcionaron una comprensión más profunda de cómo las categorías de DQ se adaptan a los informes de entrega de medicamentos y asignación de citas médicas, lo que, a su vez, contribuirá de manera positiva a la mejora de la calidad de los reportes.

a) Validación de la adaptación para el reporte entrega de medicamentos

Los resultados obtenidos del grupo focal de expertos con respecto a la adaptación de cada subcategoría se presentan en la Figura 21, la subcategoría de *Plausibilidad Atemporal* destacó con una alta adaptación, ya que el 100% de los expertos manifestó estar "Completamente de Acuerdo" con la adaptación propuesta. En cuanto a *Conformidad de Valor, Conformidad Relacional, Completitud y Plausibilidad Temporal*, estas subcategorías también demostraron una adaptación, con el 66.7% de los expertos indicando estar "Completamente de Acuerdo" y el 33.3% "De Acuerdo". La subcategoría de *Plausibilidad de Unicidad* obtuvo una respuesta favorable, con un 66.7% de los expertos "De Acuerdo" y un 33.3% "Completamente de Acuerdo". Estos resultados señalan una receptividad generalizada y positiva hacia la adaptación propuesta en los reportes.

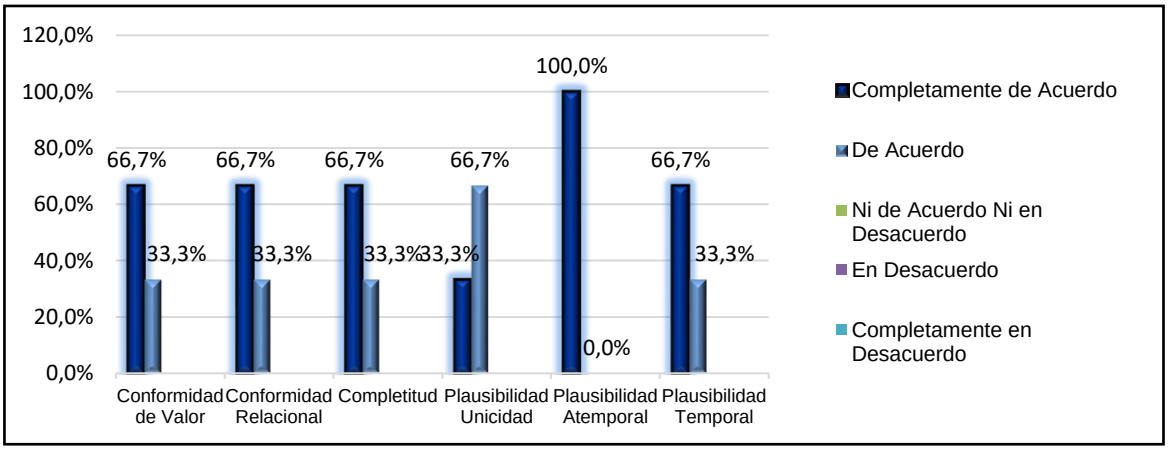


Figura 21 Resultados de la evaluación grupo focal al reporte de entrega de medicamentos

En la Tabla 20 se presenta uno de los comentarios de los expertos a las preguntas abiertas con su respectiva acción de mejora.

Tabla 20 Comentarios - Entrega de medicamentos

Id Experto	Comentario	Acción de mejora /Justificación
E3	Tener en cuenta que la resolución 521 fue derogada, al terminar la pandemia	Se excluye el atributo Grupo al que pertenece el afiliado Res.521 (Grupo 1,2,3,4 o ninguno).

b) Validación de la adaptación para el reporte asignación de citas medicas

Los resultados del grupo focal de expertos en relación con la adaptación de cada subcategoría se presentan en la Figura 22. Las subcategorías que destacan por su alta adaptación son *Conformidad de Valor y Plausibilidad Atemporal*, con un 100% de los expertos indicando estar "Completamente de Acuerdo". *Conformidad Relacional, Completitud y Plausibilidad Temporal* también muestran una sólida adaptación, con el 66.7% de los expertos expresando estar "Completamente de Acuerdo" y el 33.3% manifestando estar "De Acuerdo". Por otro lado, la subcategoría *Plausibilidad de Unicidad* obtuvo un 66.7% de respuestas en la categoría "De Acuerdo" y un 33.3% en "Completamente de Acuerdo". Estos resultados subrayan la favorable acogida y aceptación general de la adaptación propuesta a estos reportes por parte de los expertos.

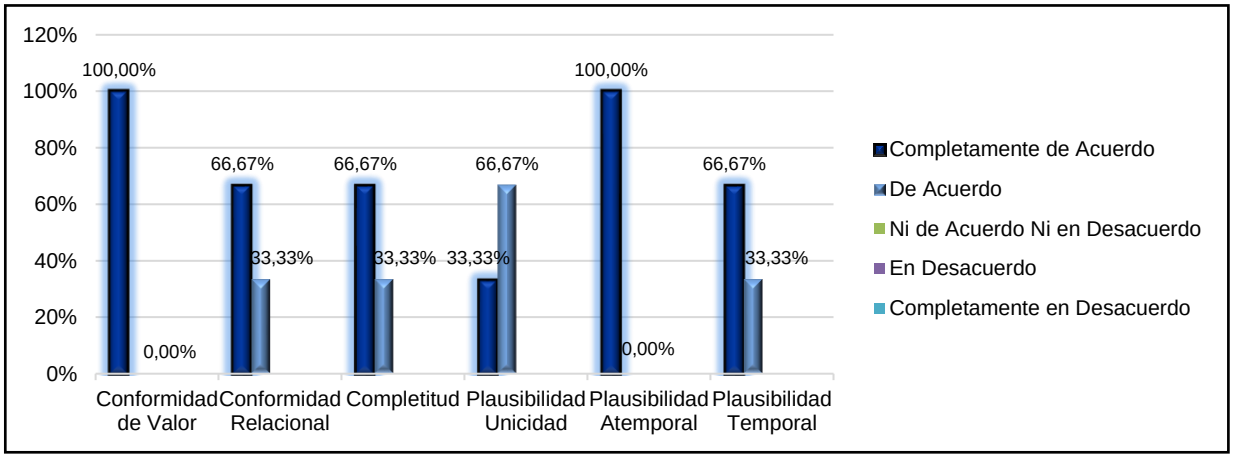


Figura 22 Resultados de la evaluación grupo focal al reporte de asignación de citas médicas

En la Tabla 21, se presenta uno de los comentarios de los expertos a las preguntas abiertas con su respectiva acción de mejora.

Tabla 21 Comentarios - Asignación de citas médicas

Id Experto	Comentario	Acción de mejora / Justificación
E1	Necesidad de establecer una codificación específica para las supra especialidades	Su implementación depende de la IPS, dado que no existe una clasificación estandarizada por parte del MSPS.

6.2 Evaluación de la DQ en el proceso ETL

Los resultados de la Evaluación de DQ realizada en el proceso ETL, fueron analizados mediante el indicador de Confianza, como se describe en la Fórmula 1. Este indicador se calcula con el objetivo de obtener información relevante en cuanto a la calidad real de los datos por cada categoría, subcategoría y atributos relevantes.

La visualización de estos resultados se realizó mediante paneles interactivos creados en la plataforma Power BI, los cuales proporcionan una representación gráfica y detallada de la calidad de los datos en los reportes relacionados con la entrega de medicamentos y la asignación de citas médicas, en esta sesión se presenta en detalle dos tableros realizados que corresponden de los resultados por cada reporte.

6.2.1 Resultados proceso ETL – Reporte Entrega de Medicamentos

Los resultados presentados en el tablero de la Figura 23, corresponden a la aplicación de la adaptación realizada al reporte de entrega de medicamentos con corte al mes de enero del año 2021 proporcionado por Asmet Salud EPS SAS. En la parte superior del tablero se muestra el resultado de tres métricas principales: *Total Registros Entrantes*, este indicador muestra el recuento total de los registros que contiene el reporte de medicamentos; *Registros sin Errores*, indica el número total de registros que no presentaron inconsistencias o errores; *% Indicador Confianza*, representa el resultado del indicador de confianza, calculado de acuerdo con los parámetros definidos en la Fórmula 1.

En el lado izquierdo del tablero, se presentan cuatro tipos de filtros para facilitar la navegación y análisis de los datos: *Fecha Reporte*, este filtro posibilita visualizar los resultados de un reporte específico, permitiendo la selección de una fecha concreta o un rango de fechas para analizar la información correspondiente a ese período en particular; *Categoría*, permite realizar la segmentación de los datos por una Categoría específica;

Subcategoría, permite la selección de una subcategoría específica dentro de la categoría elegida; *atributos,* permite segmentar la información según un atributo específico, lo que facilita un análisis más detallado y específico de los datos.

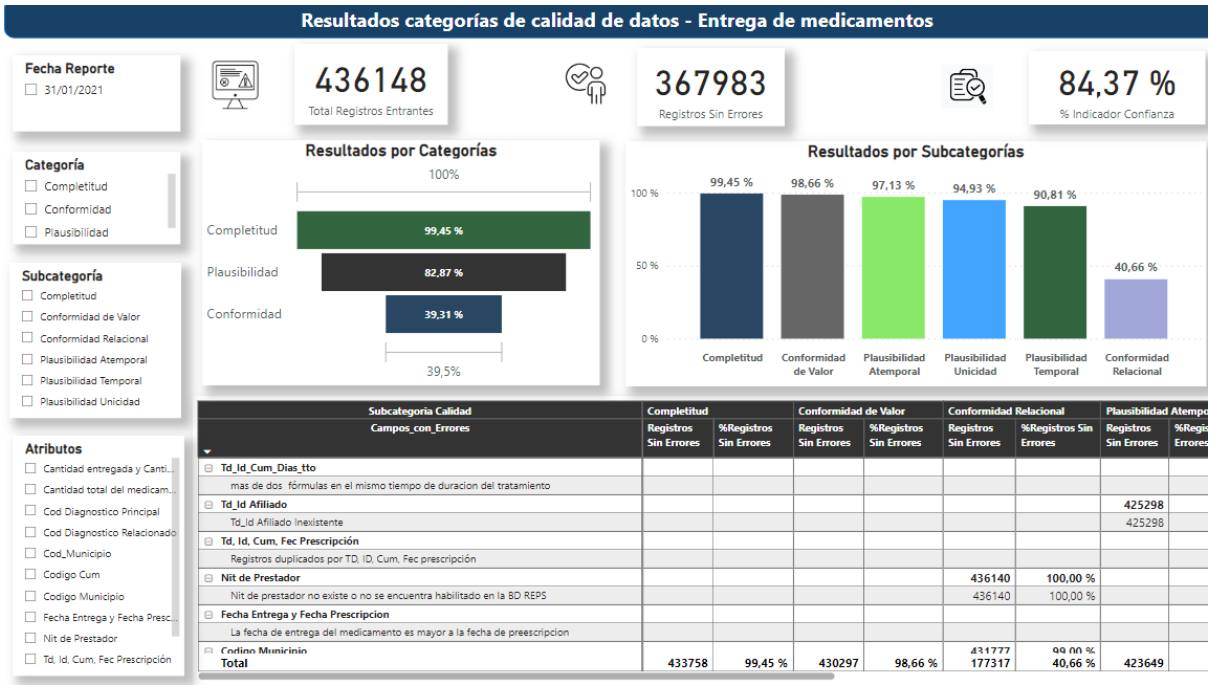


Figura 23 Tablero resultados proceso ETL - Entrega de medicamentos

En el gráfico tipo embudo (*Resultados por Categorías*), como se muestra en la Figura 24 presenta los resultados de las categorías *Completitud, Conformidad y Plausibilidad.* Para la Categoría *Completitud*, se evidencia un alto nivel de completitud de datos, con un promedio del 99.45% de confianza, lo que refleja un alto grado de cumplimiento de la adaptación propuesta, indicando que la mayoría de los datos esperados se encuentran presentes y disponibles para su análisis; por otro lado, la categoría *Plausibilidad* resalta un promedio general de 82.87% de confianza, indicando que existe coherencia y credibilidad de los datos en relación con aspectos temporales. Sin embargo, la categoría *Conformidad,* presenta un resultado del 39.31% de confianza, lo que señala un nivel insuficiente de cumplimiento de restricciones de formato interno, integridad referencial, la unicidad y la nulabilidad, respecto a lo esperado.

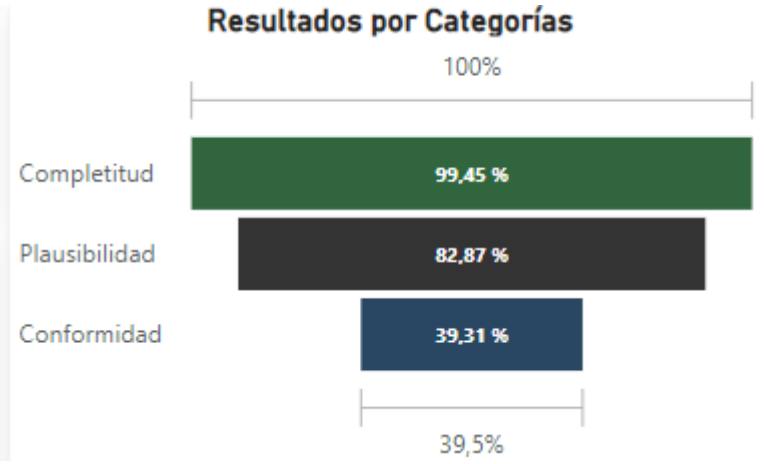


Figura 24 Resultado general por Categorías - Entrega de Medicamentos

En el gráfico de barras, representado en la Figura 25, se muestran los resultados para cada una de las categorías y subcategorías de DQ.

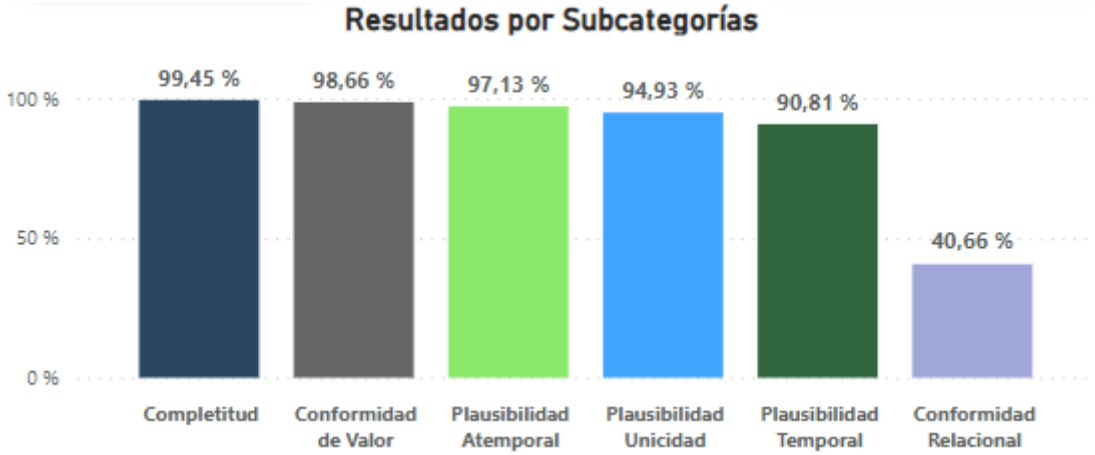


Figura 25 Resultado por Subcategorías - Entrega de Medicamentos

La Categoría *Completitud* (ver Figura 25) se destaca con el mayor porcentaje de confianza alcanzando un 99.45%, lo que indica la presencia de la gran mayoría de los datos esperados en el reporte, como la *Cantidad total entregada y la Fecha de entrega* de los medicamentos. Sin embargo, es fundamental mencionar que los registros relacionados con el *Diagnóstico principal* (ver Figura 26) presentan un 39.87% de inconsistencias, mientras que el *diagnóstico relacionado* muestra un 9.71% de errores en este informe. Dado que estos dos atributos no están sujetos a exclusión, se genera un archivo de registro de errores independiente para la revisión por parte de la EPS.

Campos_con_Errores	Registros Sin Errores	%Registros Sin Errores
Cod Diagnostico Relacionado	**393817**	**90,29 %**
Codigos de diagnostico Relacionado Inexistente	393817	90,29 %
Cod Diagnostico Principal	**262277**	**60,13 %**
Codigos de diagnostico principal Inexistente	262277	60,13 %
Total	**219946**	**50,43 %**

Figura 26 Inconsistencias Diagnósticos Completitud - Entrega de Medicamentos

En cuanto a la categoría *Conformidad de Valor* (ver Figura 25), alcanza el 98.66% de confianza, indicando que las restricciones aplicadas a la mayoría de los atributos del reporte cumplen con los criterios definidos. Sin embargo, el atributo *código CUM*, presenta dos tipos de inconsistencias como se muestra en la Figura 27.

Subcategoria Calidad	Conformidad de Valor	
Campos_con_Errores	Registros Sin Errores	%Registros Sin Errores
Codigo Cum	**430297**	**98,66 %**
El expediente del medicamento supera la cantidad de digitos permitidos (mayor a 8 digitos)	435881	99,94 %
La estructura del medicamento no concuerda con la definida por el invima, no contiene el codigo del consecutivo separado del guion (-)	430564	98,72 %
Total	**430297**	**98,66 %**

Figura 27 Inconsistencias identificadas Conformidad de Valor - Entrega de Medicamentos

En la subcategoría *Plausibilidad Atemporal* (ver Figura 25), se registra nivel de confianza del 97.13%, para los atributos definidos en la adaptación propuesta: *Tipo de documento, Número de identificación* (*Td_Id Afiliado*) su confianza es del 97.51%; para el atributo *Departamento y Municipio Afiliado* (*Código de Municipio Afiliado Inexistente*) es del 99.62%; para el atributo *Cantidad entregada de medicamento* (*Cantidad entregada y Cantidad pendiente*), su confianza es del 100%; para los atributos *Días de tratamiento y Régimen Afiliado* no presentaron inconsistencias, como se muestra en detalle en la Figura 28.

Subcategoria Calidad	Plausibilidad Atemporal	
Campos_con_Errores	Registros Sin Errores	%Registros Sin Errores
Td_Id Afiliado	**425298**	**97,51 %**
Td_Id Afiliado Inexistente	425298	97,51 %
Cod_Municipio	**434505**	**99,62 %**
Codigo de Municipio Afiliado Inexistente	434505	99,62 %
Cantidad entregada y Cantidad pendiente	**436142**	**100,00 %**
Cantidad entregadas mas la cantidad pendiente no coincide con la cantidad preescrita	436142	100,00 %
Total	**423649**	**97,13 %**

Figura 28 Inconsistencias identificadas Plausibilidad Atemporal - Entrega de Medicamentos

En la subcategoría de *Plausibilidad Unicidad* (ver Figura 25), obtuvo un nivel de confianza del 94.93%. Esto señala que, dentro del conjunto de datos evaluados, se identificaron un

total de 22.122 registros, lo que representa un 5.7% del total, como registros duplicados (ver Figura 29).

Subcategoria Calidad	Plausibilidad Unicidad	
Campos_con_Errores	Registros Sin Errores	%Registros Sin Errores
Td, Id, Cum, Fec Prescripción	414026	94,93 %
Registros duplicados por TD, ID, Cum, Fec prescripción	414026	94,93 %
Total	**414026**	**94,93 %**

Figura 29 Inconsistencias identificadas Plausibilidad Unicidad - Entrega de Medicamentos

En la subcategoría *Plausibilidad Temporal* (ver Figura 25), se obtiene un resultado del 90.81% de confianza, indicando que para el atributo *Cantidad de fórmulas en el mismo tiempo duración del tratamiento* (*Td_Id_Cum_Dias_tto*) presenta un resultado del 93.69% (ver Figura 30), lo que significa que existen más de dos fórmulas para el mismo afiliado con el mismo tiempo de duración del tratamiento. Para el atributo *Fecha de entrega de medicamento* (*Fecha Entrega y Fecha Prescripción*) presentan un nivel de confianza del 100%, lo que significa que los registros cumplen con el criterio definido; para el atributo *Cantidades Entregadas y Pendientes Totales* (*Cantidad entregada y Cantidad pendiente*), como se muestra en la Figura 30, se obtiene un resultado del 97.12%. Este resultado señala la presencia de registros específicos (un total de 12.561) que muestran la misma fórmula médica registrada durante un mismo periodo o rango de tratamiento.

Subcategoria Calidad	Plausibilidad Temporal	
Campos_con_Errores	Registros Sin Errores	%Registros Sin Errores
Td_Id_Cum_Dias_tto	408642	93,69 %
mas de dos fórmulas en el mismo tiempo de duracion del tratamiento	408642	93,69 %
Fecha Entrega y Fecha Prescripcion	436140	100,00 %
La fecha de entrega del medicamento es mayor a la fecha de preescripcion	436140	100,00 %
Cantidad entregada y Cantidad pendiente	423581	97,12 %
El registro supera las cantidades preescritas, revisar	423581	97,12 %
Total	**396067**	**90,81 %**

Figura 30 Inconsistencias identificadas Plausibilidad Atemporal - Entrega de Medicamentos

Por último, en la subcategoría *Conformidad Relacional* (ver Figura 25), se observa un resultado del 40.66% de confianza. Este porcentaje indica que los atributos (ver Figura 31) *Nit del prestador* (100%) y *Código de Municipio* (99.00%) presentaron una alta confiabilidad. Sin embargo, el atributo *Código Cum* alcanzo un nivel de confianza del 91.23%, con un total

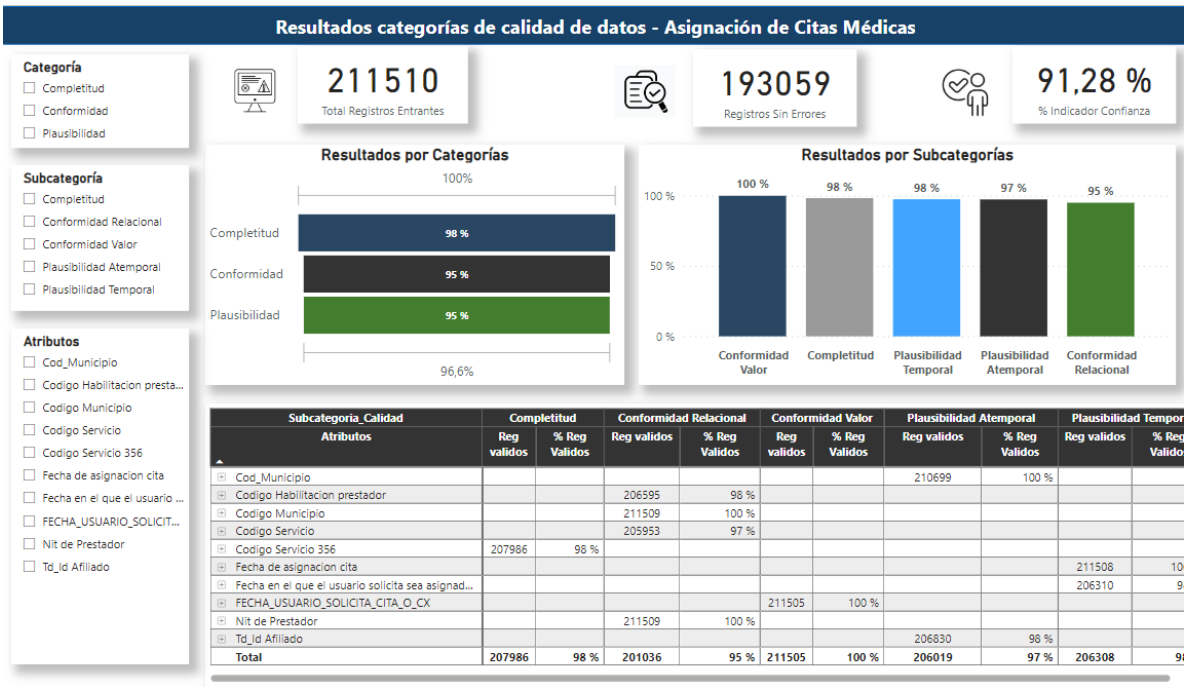
de 42.331 registros con *códigos CUM* inexistentes con la relación a la fuente externa del Invima; para el atributo *Cod Diagnóstico Relacionado* se obtuvo un nivel de confianza del 90.29%, mientras que el atributo *Cod Diagnóstico Principal*, mostro un nivel de confianza del 40.66%. Se identificaron errores recurrentes, especialmente relacionados con diagnósticos que no cumplían con los estándares de uso de acuerdo con la fuente externa CIE10.

Subcategoria Calidad	Conformidad Relacional	
Campos_con_Errores	Registros Sin Errores	%Registros Sin Errores
Nit de Prestador	**436140**	**100,00 %**
Nit de prestador no existe o no se encuentra habilitado en la BD REPS	436140	100,00 %
Codigo Municipio	**431777**	**99,00 %**
El codigo del municipio registrado no existe	431777	99,00 %
Codigo Cum	**397898**	**91,23 %**
El Codigo Cum del medicamento es Inexistente, conforme a la BD Invima	397898	91,23 %
Cod Diagnostico Relacionado	**393817**	**90,29 %**
Codigos de diagnostico Relacionado Inexistente	393817	90,29 %
Cod Diagnostico Principal	**262277**	**60,13 %**
Codigos de diagnostico principal Inexistente	262277	60,13 %
Total	**177317**	**40,66 %**

Figura 31 Inconsistencias identificadas Conformidad Relacional- Entrega de Medicamentos

6.2.2 Resultados proceso ETL – Reporte Asignación de citas Médicas

Los resultados presentados en el tablero como se muestran en la Figura 32, corresponden a la aplicación de la adaptación realizada al reporte de asignación de citas médicas con corte al mes de abril del año 2021 proporcionado por Asmet Salud EPS SAS. La estructura del panel del tablero tiene las mismas características presentadas en los resultados del proceso ETL para el reporte de entrega de medicamentos.



Subcategoria_Calidad / Atributos	Completitud		Conformidad Relacional		Conformidad Valor		Plausibilidad Atemporal		Plausibilidad Tempor	
	Reg validos	% Reg Validos	Reg validos	% Reg Validos	Reg validos	% Reg Validos	Reg validos	% Reg Validos	Reg validos	% Reg Valido
Cod_Municipio							210699	100 %		
Codigo Habilitacion prestador			206595	98 %						
Codigo Municipio			211509	100 %						
Codigo Servicio			205953	97 %						
Codigo Servicio 356	207986	98 %								
Fecha de asignacion cita									211508	10
Fecha en el que el usuario solicita sea asignad...									206310	9
FECHA_USUARIO_SOLICITA_CITA_O_CX					211505	100 %				
Nit de Prestador			211509	100 %						
Td_Id Afiliado							206830	98 %		
Total	**207986**	**98 %**	**201036**	**95 %**	**211505**	**100 %**	**206019**	**97 %**	**206308**	**9**

Figura 32 Tablero resultados proceso ETL – Asignación de citas médicas

En el gráfico tipo embudo (*Resultados por Categorías*), como se muestra en la Figura 33, se presentan los resultados de las categorías *Completitud, Conformidad y Plausibilidad*. La Categoría Completitud, presenta un promedio del 98% de confianza; y *Plausibilidad*, y *Conformidad*, presenta un resultado de 95% de confianza, lo que refleja que la mayoría de los registros de este reporte cumplen con los criterios de adaptación definidos.

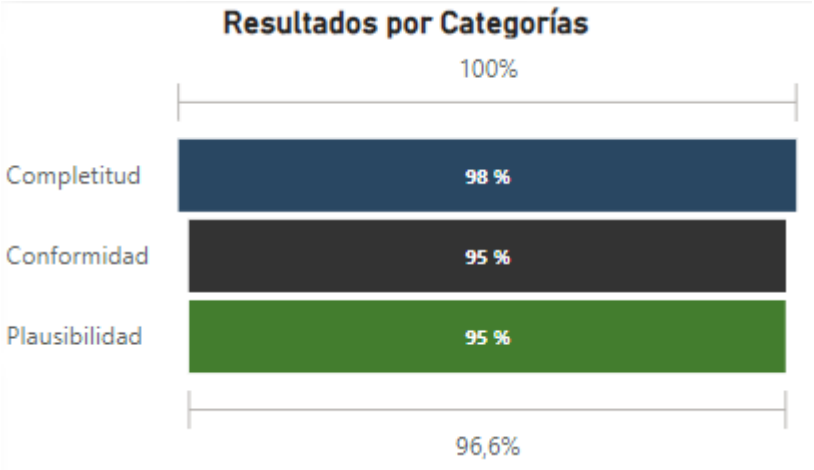


Figura 33 Resultado general por Categorías – Asignación de citas médicas

En el gráfico de barras (ver Figura 34) se presentan los resultados para cada una de las categorías y subcategorías de DQ. La Subcategoría *Conformidad de valor* se destaca con el mayor porcentaje de confianza alcanzando un 100%, este resultado indica que las restricciones aplicadas a los atributos del reporte cumplen con los criterios de adaptación definidos.

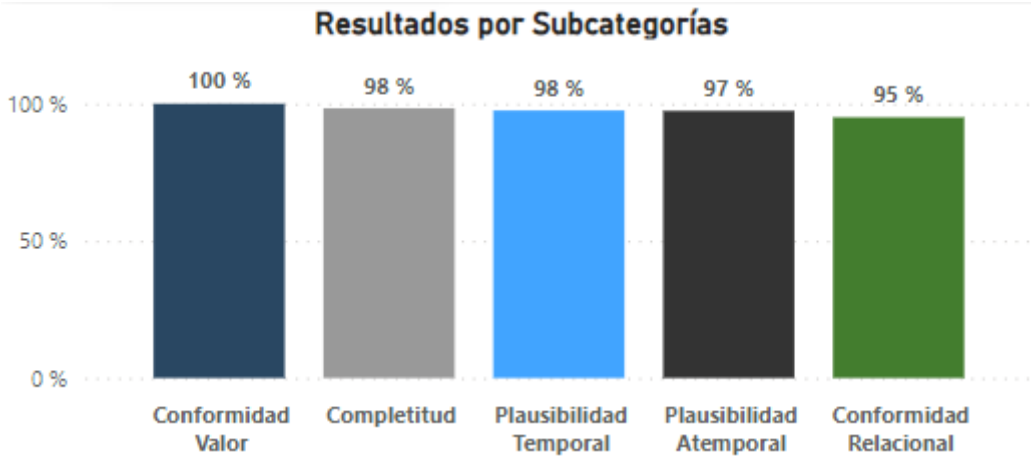


Figura 34 Resultado general por Subcategorías – Asignación de citas médicas

La categoría de *Completitud* (ver Figura 34) muestra un nivel de confianza del 98%, indicando que la gran mayoría de los datos esperados se encuentran presentes en el reporte de asignación de citas médicas, en concordancia con los criterios de adaptación definidos. Sin embargo, es relevante señalar que, al revisar el atributo *Otra especialidad o Cirugía* (ver Figura 35), se identificaron un total de (211510 – 207986 = 3524) registros inconsistentes. Estas discrepancias podrían requerir una revisión adicional por parte de la EPS para garantizar la coherencia de dichos datos.

Subcategoria_Calidad	Completitud	
Atributos	Reg validos	% Reg Validos
⊟ Otra especialidad o Cirugia	207986	98 %
El servicio (356) otra especialidad no registra el servicio solicitado	207986	98 %
Total	207986	98 %

Figura 35 Resultado Categoría Completitud – Asignación de citas médicas

La subcategoría *Plausibilidad Temporal* (ver Figura 34) presenta un nivel de confianza del 98%, lo que señala una notable coherencia temporal en los datos analizados. No obstante,

se ha detectado que existen inconsistencias en dos registros respecto al atributo *Fecha de asignación de cita o cirugía con dos registros inconsistentes* (211510 – 211508 = 2), así como en registros en lo que respecta al atributo *Fecha en que el usuario solicita que se le asigne la cita* (211510 – 206310 = 5200) con cinco mil doscientos registros inconsistentes, tal como se detalla en los resultados presentados en la Figura 36.

Subcategoria_Calidad	Plausibilidad Temporal	
Atributos	Reg validos	% Reg Validos
⊟ **Fecha de asignacion cita**	**211508**	**100 %**
La fecha de asignacion de la cita o cirugia es inferior a la fecha en que el usuario solicita la cita o cx, Asignan la cita antes de solicitarla?	211508	100 %
⊟ **Fecha en el que el usuario solicita sea asignada la cita**	**206310**	**98 %**
La fecha en la que el usuario solicita sea asignada la cita es menor a la fecha en que el usuario solicita la cita	206310	98 %
Total	**206308**	**98 %**

Figura 36 Resultado Subcategoría Plausibilidad Temporal– Asignación de citas médicas

La subcategoría *Plausibilidad Atemporal* (ver Figura 34), presenta un 97% de confianza, lo que indica que los criterios de verificación y validación concuerdan con las fuentes internas y externas definidas en la adaptación realizada como se muestra en la Figura 37. No obstante, se resalta la necesidad que la EPS revise en conjunto con los prestadores de servicios de salud los 5491 registros asociados al atributo *Tipo de documento + Número de identificación* (*Td_id Afiliado*), ya que presentan inconsistencias que requieren atención, corrección o verificación.

Subcategoria_Calidad	Plausibilidad Atemporal	
Atributos	Reg validos	% Reg Validos
⊟ **Cod_Municipio**	**210699**	**100 %**
Codigo de Municipio Afiliado Inexistente	210699	100 %
⊟ **Td_Id Afiliado**	**206830**	**98 %**
Td_Id Afiliado Inexistente	206830	98 %
Total	**206019**	**97 %**

Figura 37 Resultado Subcategoría Plausibilidad Atemporal– Asignación de citas médicas

La subcategoría *Conformidad Relacional* (Figura 34), presenta un nivel de confianza del 95%, siendo el porcentaje más bajo en comparación con las otras subcategorías, lo que sugiere que el 5% restante, que refleja inconsistencias en los atributos *Código Habilitación Prestador* con un 98%, *Código Municipio* con un 100%, *Código Servicio* con un 97%, y *Nit*

de prestador con un 95% de confianza como se muestra en la Figura 38, debe ser revisado por la EPS.

Subcategoria_Calidad	Conformidad Relacional	
Atributos	Reg validos	% Reg Validos
⊟ **Codigo Habilitacion prestador**	**206595**	**98 %**
El codigo de habilitacion del prestador no existe	206595	98 %
⊟ **Codigo Municipio**	**211509**	**100 %**
El codigo del municipio registrado no existe	211509	100 %
⊟ **Codigo Servicio**	**205953**	**97 %**
El codigo de servicio no existe en la bd REPS	205953	97 %
⊟ **Nit de Prestador**	**211509**	**100 %**
Nit de prestador Inexistente	211509	100 %
Total	**201036**	**95 %**

Figura 38 Resultado Subcategoría Conformidad Relacional– Asignación de citas médicas

CAPITULO 7

7 CONCLUSIONES Y TRABAJO FUTURO

7.1 Conclusiones

Al realizar un proceso previo de limpieza y normalización de datos permite mejorar la calidad del dato, y al adaptar los criterios de las categorías de calidad de datos propuestos a los reportes de entrega de medicamentos y asignación de citas médicas, permiten, identificar los atributos con más inconsistencias o más confiables. Siguiendo la secuencia sugerida por Kahn, se comienza con la categoría *Conformidad,* lo cual asegura que los datos se ajusten adecuadamente a las restricciones de formato y estándares establecidos. Luego se aborda la Categoría *Completitud*, que garantiza la presencia continua y consistente de datos en los reportes. Finalmente, la categoría *Plausibilidad,* para verificar que los datos sean creíbles, coherentes y acordes con el conocimiento local y temporal. Esta aplicación secuencial de las categorías de calidad de datos contribuye a la obtención de reportes más confiables, fomentando así la toma de decisiones informadas en el ámbito de la salud.

Durante la adaptación a los reportes de los criterios de la subcategoría de *Plausibilidad Atemporal*, no se consideró el criterio de *Valores de medición repetida*, debido a que no se identificaron registros que hubieran sido recolectados o medidos repetidamente en diferentes momentos temporales para un mismo afiliado dentro de los reportes analizados.

De la misma forma, en la subcategoría de *Plausibilidad Temporal*, no se consideraron tres criterios específicos: en primer lugar, *Valores derivados*, el cual implica la obtención de valores a través de cálculos o transformaciones de datos observados, lo cual no era relevante para la adaptación realizada. En segundo lugar, S*ecuencias de valores,* que se refiere al orden en que aparecen los datos dentro de una serie, y en este caso, no aplicaba a la naturaleza de los reportes analizados. Por último, *Medidas de densidad de valor*, que describe la distribución de los datos a lo largo de un rango determinado, aspecto que no fue relevante para la evaluación específica llevada a cabo en dichos reportes. Estos criterios, aunque importantes en otros contextos, no fueron pertinentes ni aplicables al análisis de la plausibilidad temporal dentro de la adaptación de los reportes en cuestión.

La adaptación de los criterios de las categorías de calidad de datos propuestos a los reportes de entrega de medicamentos y asignación de citas médicas se validó por un grupo focal, conformado por expertos del sector salud, cuyos resultados indicaron que la adaptación propuesta se ajusta a las necesidades de calidad de los datos en estos dos reportes, teniendo en cuenta que en todas las subcategorías se obtiene un porcentaje de 100% entre “Completamente de Acuerdo” y “De Acuerdo”. Esto respalda la relevancia de la implementación de los criterios de calidad de datos en estos reportes.

Aplicar las categorías de calidad de datos adaptadas a los reportes en la fase de transformación del proceso ETL permite la integración de diversas fuentes de información y facilita la consolidación de los reportes, ayudando a minimizar los errores y reduciendo la dependencia de tareas manuales, lo que contribuye a la generación de reportes confiables y precisos.

Por otra parte, la evaluación de la calidad de los datos en estos reportes se realizó por medio del indicador de confianza, cuyos resultados obtenidos han demostrado un alto nivel de confiabilidad en general, reflejando una consistencia significativa en la mayoría de las categorías y subcategorías evaluadas. Sin embargo, es importante resaltar que, a pesar del alto porcentaje general de confianza, las discrepancias detectadas en la subcategoría *Conformidad Relacional*, que fue la que obtuvo el más bajo porcentaje de confianza en especial para el reporte entrega de medicamentos con un 40.66% los atributos diagnóstico principal (60.13%) y diagnóstico relacionado (90.29%), fueron los que más inconsistencias presentaron, lo cual indica la necesidad de una revisión minuciosa por parte de las EPS y posiblemente la implementación de estrategias de mejora para fortalecer la precisión y la integridad de los atributos referenciales.

7.2 Trabajo futuro

Las inconsistencias identificadas en esta investigación en los reportes de entrega de medicamentos y asignación de citas médicas permiten vislumbrar acciones preventivas como:

- Crear un validador de los reportes**,** en el que se establezcan controles de entrada de datos, es decir, crear reglas de entrada de datos que eviten que los datos no válidos o inexactos ingresen a un sistema o un proceso ETL.
- Definir roles y responsabilidades claros que apliquen, supervisen, y realicen seguimiento al proceso ETL y la adaptación de las categorías de calidad de datos, con el objetivo de mejorar la toma de decisiones y una gestión efectiva de los datos e información.

La implementación de algoritmos de inteligencia artificial alineados con el proceso ETL para los reportes de entrega de medicamentos y asignación de citas médicas, que permitan:

- Detectar de forma automática las anomalías, corrigiendo posibles inconsistencias o datos atípicos de los reportes durante el proceso.
- Optimizar la calidad de los datos en tiempo real mediante validaciones continuas, garantizando que cumplan con los estándares establecidos desde su ingreso.
- Mejorar la eficiencia del proceso ETL con algoritmos de aprendizaje automático que identifiquen patrones y automaticen tareas, reduciendo el tiempo de procesamiento.

8 REFERENCIAS

[1] T. Dai, H. Hu, Y. Wan, Q. Chen, and Y. Wang, “A data quality management and control framework and model for health decision support,” in *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Zhangjiajie,China, Aug. 2015, pp. 1792–1796. doi: https://doi.org/10.1109/FSKD.2015.7382218.

[2] Ministerio de Salud y Protección Social (MSPS), “Resolución 1604 de 2013.” Accessed: Feb. 18, 2023. [Online]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/resolucion-1604-de-2013.pdf

[3] Ministerio de Salud y Protección Social (MSPS), “Resolución 1552 de 2013.” Accessed: Mar. 18, 2023. [Online]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/resolucion-1552-de-2013.pdf

[4] Instituto nacional de Vigilancia de Medicamentos y Alimentos INVIMA, “ABC Seguridad en el uso de medicamentos,” Ministerio de Salud y Protección Social (MSPS). Accessed: Feb. 09, 2022. [Online]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/CA/seguridad-en-la-utilizacion-de-medicamentos.pdf

[5] M. Hendayun, E. Yulianto, J. F. Rusdi, A. Setiawan, and B. Ilman, “Extract transform load process in banking reporting system,” *MethodsX*, vol. 8, p. 101260, 2021, doi: https://doi.org/10.1016/j.mex.2021.101260.

[6] L. Marco-Ruiz, D. Moner, J. A. Maldonado, N. Kolstrup, and J. G. Bellika, “Archetype-based data warehouse environment to enable the reuse of electronic health record data,” *Int J Med Inform*, vol. 84, no. 9, pp. 702–714, 2015, doi: https://doi.org/10.1016/j.ijmedinf.2015.05.016.

[7] Ralph Kimball- Jose Caserta, “The Data Warehouse ETL Toolkit,” *Wiley Publishing, Inc.*, 2004, doi: eISBN: 0-764-57923-1.

[8] Talend, “¿En qué consiste un proceso de ETL (Extraer, Transformar y Cargar)?” [Online]. Available: https://www.talend.com/es/resources/what-is-etl/

[9] N.Duque, E. Hernández, Á. Pérez, A. Arroyave, and D.Espinosa, “Modelo para el proceso de extracción, transformación y carga en bodegas de datos. Una aplicación con datos ambientales,” *Ciencia e Ingeniería Neogranadina*, vol. 26, no. 2, pp. 95–109, May 2016, doi: https://doi.org/10.18359/rcin.1799.

[10] I. Homayouni, H., Ghosh, S., & Ray, “An Approach for Testing the Extract-Transform-Load Process in Data Warehouse Systems,” *IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), Memphis, TN, USA*, pp. 158–161, Oct. 2018, doi: https://doi.org/10.1109/ISSREW.2018.000-6.

[11] V. C. Pezoulas *et al.*, “Medical data quality assessment: On the development of an automated framework for medical data curation,” *Comput Biol Med*, vol. 107, pp. 270–283, 2019, doi: https://doi.org/10.1016/j.compbiomed.2019.03.001.

[12] H. G. Kahn, Michael G.; Callahan, Tiffany J.; Barnard, Juliana; Bauck, Alan E.; Brown, Jeff; Davidson, Bruce N.; Estiri, “Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data,” *eGEMs*, vol. 4, no. 1, p. 1244, Sep. 2016, doi: https://doi.org/10.13063/2327-9214.1244.

[13] K. E. Lynch *et al.*, “Incrementally Transforming Electronic Medical Records into the Observational Medical Outcomes Partnership Common Data Model: A Multidimensional Quality Assurance Approach.,” *Appl Clin Inform*, vol. 10, no. 5, pp. 794–803, Oct. 2019, doi: https://doi.org/10.1055/s-0039-1697598.

[14] H. Spengler, I. Gatz, F. Kohlmayer, K. A. Kuhn, and F. Prasser, “Improving Data Quality in Medical Research: A Monitoring Architecture for Clinical and Translational Data Warehouses,” in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, Rochester, USA, Jul. 2020, pp. 415–420. doi: https://doi.org/10.1109/CBMS49503.2020.00085.

[15] L. G. Qualls *et al.*, “Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®),” *EGEMS (Wash DC)*, vol. 6, no. 1, p. 3, Apr. 2018, doi: https://doi.org/10.5334/egems.199.

[16] W. H. Inmon, *Building the Data Warehouse,3rd Edition*, 3rd ed. USA: John Wiley & Sons, Inc., 2002.

[17] Earley S. Henderson D. & Data Management Association., *Dama-dmbok : data management body of knowledge (Second)*, 2nd ed. Basking Ridge, NJ 07920 USA: Technics publications, 2017.

[18] M. Petersen, K., Feldt, R., Shahid, M., Mattsson, “Systematic mapping studies in software engineering,” *%B Proceedings of the 12th international conference on Evaluation and Assessment in Software Engineering, Italy*, pp. 68–77, 2008.

[19] B. Kitchenham, P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, *Systematic literature reviews in software engineering – A systematic literature review*, ISSN 0950-5849., vol. 51, Issue 1,. 2009. doi: https://doi.org/10.1016/j.infsof.2008.09.009.

[20] K. Pratt, “Design Patterns for Research Methods: Iterative Field Research,” *Assoc. Adv. Artif. Intell., no. 1994, 2009*, 2009, [Online]. Available: http://www.kpratt.net/wp-content/uploads/2009/01/research_methods.pdf.

[21] J. and L. L. Kontio Jyrki and Bragge, “The Focus Group Method as an Empirical Tool in Software Engineering,” in *Guide to Advanced Empirical Software Engineering*, J. and S. D. I. K. Shull Forrest and Singer, Ed., London: Springer London, 2008, pp. 93–116. doi: 10.1007/978-1-84800-044-5_4.

[22] M. M., . C., and . F. J., “Focus group como proceso en Ingeniería de Software: una experiencia desde la práctica,” *Dyna (Medellin)*, vol. 80, pp. 51–60, 2013, [Online]. Available: https://www.redalyc.org/articulo.oa?id=49628728006