

Metodología para la integración de datos clínicos del Centro de Inmunología Molecular

Methodology for the integration of clinical data from the Center for Molecular Immunology

MSc. Arasay Montes De Santis¹, MSc. Jorge Luis Palomino Herandez²

¹Centro de Inmunología Molecular, Calle 15 esq. 216 S/N, Siboney, Playa, La Habana, Cuba.

²Centro de Inmunología Molecular, Calle 15 esq. 216 S/N, Siboney, Playa, La Habana, Cuba.

*Autor de correspondencia: email (arasay@cim.sld.cu)

Resumen

Durante los 28 años de investigaciones clínicas en el Centro de Inmunología Molecular, se ha acumulado una gran cantidad de información en diferentes plataformas de entrada de datos, sobre la cual es importante realizar distintos tipos de análisis. Para ello se diseñó una metodología que permitiese la integración de las diferentes fuentes de datos. Como herramienta para implementar esta metodología se seleccionó Pentaho Data Integration, la cual permitió diseñar flujos reproducibles de transformación de datos.

Se validó esta metodología mediante la integración de 12 investigaciones de cáncer de pulmón. Se concluye que esta metodología facilita el proceso de captura, limpieza y almacenamiento de datos, utilizando un formato uniforme y consistente. Su uso permite establecer la trazabilidad y auditabilidad, lo cual mejora la calidad de los análisis y toma de decisiones, a la vez que reduce el tiempo de manejo de la información.

Palabras clave: integración de datos, limpieza de datos, análisis de datos.

Abstract

During the 28 years of clinical research at the Center for Molecular Immunology, a large amount of information has been accumulated in different data entry platforms, on which it is important to perform different types of analysis. For this, a methodology was designed that would allow the integration of different data sources. Pentaho Data Integration was selected as a tool to implement this methodology, which allowed designing reproducible data transformation flows.

This methodology was validated by integrating 12 lung cancer investigations. It is concluded that this methodology facilitates the process of capturing, cleaning and storing data, using a uniform and consistent format. Its use allows traceability and auditability to be established, which improves the quality of analysis and decision making, while reducing information handling time.

Keywords: data integration, data cleaning, data analysis.

1. Introducción

Las instituciones dedicadas a la investigación clínica a nivel mundial, manejan una gran cantidad de información¹. Esta información, en su gran mayoría corresponde a los datos generados como resultado de las investigaciones clínicas realizadas².

Una investigación clínica, es una investigación en la que se estudian personas, datos o muestras de tejido para comprender la salud y la enfermedad³. Los datos procedentes de las investigaciones

clínicas tienen gran importancia, ya que, al ser analizados, permiten obtener conclusiones sobre los productos médicos desarrollados.

El Centro de Inmunología Molecular es uno de los centros que conforman la industria biotecnológica en Cuba. Desde sus inicios ha desarrollado una serie de biomoléculas para el tratamiento de diferentes enfermedades. Los proyectos de investigación básica se centran en la inmunoterapia del cáncer y en la búsqueda de nuevos productos para el diagnóstico y tratamiento de enfermedades relacionadas con el sistema inmunológico⁴. Durante estos 28 años de desarrollo clínico de los productos del CIM, se han logrado 17 registros sanitarios de múltiples medicamentos para diferentes patologías.

Como resultado de este desarrollo, se ha generado una gran cantidad de información, la cual ha sido recogida usando distintos métodos. Inicialmente, los datos clínicos eran recopilados en Cuadernos de Recogida de Datos recogidos en papel (CRDp). Para su procesamiento y análisis, la información generada en el CRDp era transcrita en tablas de datos utilizando fundamentalmente el sistema Epi-Data, el cual permite además de la entrada de los datos, la comparación de dos tablas de datos⁵. De esta manera se recolectaron datos de más de 60 ensayos clínicos.

En el año 2012, los especialistas del CIM en conjunto con el Centro de Informática Médica (CESIM) de la Universidad de las Ciencias Informáticas (UCI), desarrollaron el Sistema de Gestión de Datos electrónicos de Ensayos Clínicos (alasClínicas), con el fin de acelerar e informatizar el proceso de gestión de datos clínicos.

Este nuevo sistema permitió la transmisión remota de datos entre la Red Nacional de Hospitales y el CIM. A este sistema posteriormente se le designó el nombre de XAVIA Clínicas⁶. En 2019, este sistema fue sustituido por el sistema Xavia Sidec. Este nuevo sistema constituyó una evolución lógica en el proceso de recogida de datos, resolviendo deficiencias presentadas por XAVIA Clínicas e introduciendo nuevas funcionalidades⁷.

Debido a la gran cantidad de información generada en todos estos años y a la variabilidad de los sistemas en los que se ha recopilado la misma, se vuelve muy engorroso hacer un análisis integrado de toda esta data. Hasta el momento, para realizar análisis sobre los datos integrados de algunos ensayos, los responsables del análisis estadístico han realizado la fusión de las bases de datos de manera manual. Este es un proceso complejo, propenso a errores y que consume gran cantidad de tiempo y esfuerzo. Debido a este problema se hace necesario desarrollar una metodología de integración de datos a gran escala que permita, con la ayuda de las herramientas informáticas adecuadas optimizar la integración y análisis de los datos clínicos en el CIM.

En el presente trabajo se propone y valida el diseño de una metodología para la integración de las diferentes fuentes de datos clínicos del CIM.

2. Materiales y Métodos

Para el desarrollo de la investigación se realizó un estudio del estado del arte relacionado a los procesos de integración de datos clínicos, metodologías y herramientas para realizarlos. Este estudio permitió desarrollar una metodología propia, que se adecuará a los datos clínicos almacenados en el Centro de Inmunología Molecular.

Esta metodología abarca como etapas fundamentales la caracterización de los datos, el proceso de diseño, la implementación y la validación. Durante la caracterización de los datos se analizan y evalúan los detalles prácticos de los conjuntos de datos, que dan como resultado una visión resumida de la estructura de los datos y sus valores. Esto permite destacar posibles problemas de calidad de los datos, decidir las áreas de mejora o mapear otro perfil de datos para el proyecto de integración.

En la etapa de diseño, se concibe el proceso de integración, teniendo en cuenta aspectos como el diseño arquitectónico, que muestra cómo se manejarán los datos en los sistemas, el modelo mejorado de los datos y las reglas de limpieza de datos, normalización y cotejo. En esta etapa

también se determina la tecnología que se utilizará para implementar, verificar, supervisar e iterar el proceso de integración.

Durante la etapa de implementación se efectúa la integración de las fuentes de datos. Esta implementación se producirá de forma incremental, integrando bajos volúmenes de datos procedentes de fuentes menos conflictivas, para ir posteriormente aumentando iterativamente los volúmenes y añadiendo más fuentes.

Finalmente, durante la validación, se comprueba que se ha realizado de manera adecuada la integración de los datos. Aquí se tienen en cuenta cuestiones fundamentales, tales como que la pérdida de datos es nula o mínima, que el significado de los datos no cambió durante la integración ni se deterioró la calidad de los datos, que la información se encuentra disponible en el momento en que se necesita y que el proceso de integración funciona siempre como se espera. El objetivo de esta metodología es posibilitar un proceso de integración reproducible, incremental, auditable y orientado a los análisis a realizar en el CIM.

Como herramientas informáticas fundamentales para la implementación de esta metodología se usó, el sistema de base de datos PostgreSQL para el almacenamiento centralizado de los datos y Pentaho Data Integration para realizar los procesos de transformación e integración de datos⁸. Para asegurar la auditabilidad y facilitar el trabajo en equipo se usó la herramienta de gestión de versiones Git⁹ y la herramienta de inteligencia de negocio Metabase para garantizar la exploración de los datos¹⁰.

La metodología desarrollada fue validada mediante la aplicación de un caso de estudio y por juicio de expertos. El caso de estudio consistió en la integración de los datos de 12 investigaciones de cáncer de pulmón, con el propósito de utilizar esta integración en el análisis de la supervivencia de los pacientes tratados con varios productos desarrollados en el CIM. Aunque en el CIM se han realizado estudios para combatir otros tipos de cáncer, predominan los estudios sobre esta localización, dada la prevalencia en Cuba y el mundo. Este caso de estudio permitió realizar una evaluación cualitativa del uso de esta metodología en este escenario.

El juicio de expertos es un método de validación útil para verificar la fiabilidad de una investigación, que se define como “una opinión informada de personas con trayectoria en el tema, que son reconocidas por otros como expertos cualificados en éste, y que pueden dar información, evidencia, juicios y valoraciones”¹¹. La validación por expertos se realizó a través de una encuesta con escala Likert con 7 preguntas para obtener y considerar sus opiniones con respecto a los beneficios de la integración de datos clínicos. La escala de Likert fue creada en 1932, es un modelo de encuesta que suele incluir de 5 a 7 opciones de respuesta que van desde muy de acuerdo a muy en desacuerdo, con una opción neutral en el medio. La escala puede utilizarse para medir una serie de temas de sentimiento diferentes, como la probabilidad, el acuerdo, la calidad, la frecuencia y la importancia¹².

3. Resultados y Discusión

Descripción del caso de estudio

En este caso de estudio se almacenaron, limpiaron e integraron utilizando la metodología propuesta, los datos clínicos de 12 investigaciones de cáncer de pulmón, con el propósito de utilizar esta integración en el análisis de la supervivencia de los pacientes tratados con varios productos desarrollados en el CIM.

De las 12 investigaciones seleccionadas, 1 corresponde a un estudio exploratorio, 2 son estudios patrocinados por el investigador, 2 son ensayos Fase II, 1 es Fase II/III, 4 son Fase III y 2 son Fase IV. En cada una de estas investigaciones se recogen por lo general más de 200 variables, correspondientes a una cantidad distinta de pacientes que va desde 20 en el estudio exploratorio, incrementándose hasta 1060 en uno de los Fase IV.

Del total de variables disponibles en cada una de las investigaciones escogidas para la integración, se seleccionaron inicialmente 27 variables (23 correspondientes a los datos demográficos), incluyéndose en un momento posterior 4 más, para un total de 31 variables.

Una vez seleccionadas las variables a partir de la caracterización de cada una de ellas, el equipo de manejo de datos, de conjunto con los especialistas en análisis estadístico incluyeron en un diccionario de términos estas variables, estableciendo un nombre único para cada una, su tipo de dato, y especificándose una descripción del significado de la variable para el negocio, además de los valores que puede contener la misma.

Se realizó el proceso de ingestión de los datos a partir de las fuentes originales, almacenándolos dentro de un esquema de la base de datos PostgreSQL, destinada a almacenar centralizadamente los datos en su estructura original, para luego realizar su limpieza. La ingestión y la limpieza se realizaron en paralelo por parte de los distintos especialistas del equipo de manejo de datos, utilizando para ello Pentaho Data Integration. Esta herramienta permitió construir flujos de transformación de datos. De manera general, estos flujos cargaron los datos desde el área de concentración, se renombraron las variables contenidas en el modelo que fueron seleccionadas para la integración, según el término especificado en el diccionario de términos del negocio, se cambiaron sus tipos de datos en caso de ser necesario, se creó la variable `codigo_paciente`, encargada de identificar de manera única a cada paciente y se almacenaron los datos limpios en un esquema de la base de datos dedicada a su almacenamiento.

En la Figura 1 se muestra un flujo de transformación creado usando Pentaho Data Integration para la limpieza de uno de los estudios.

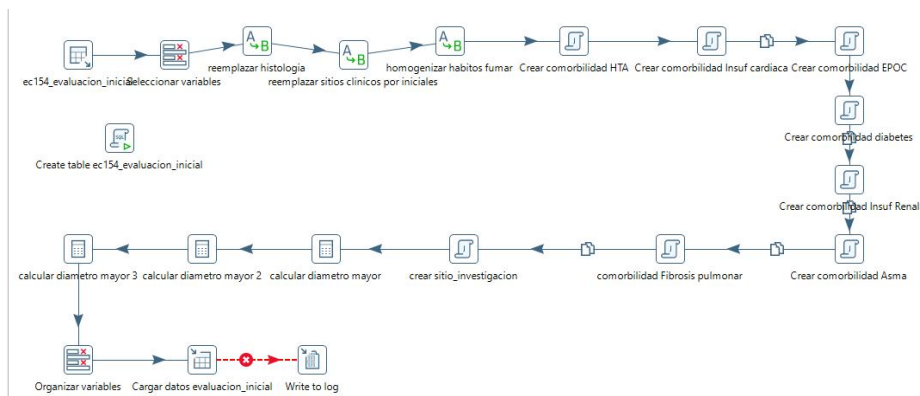


Fig.1 Flujo de transformación en Pentaho Data Integration

Como resultado de la ejecución de estos flujos quedaron almacenadas en el área de limpieza para cada investigación las variables seleccionadas homogeneizadas y completadas donde fue posible. Finalmente, se procedió a integrar las variables seleccionadas de cada una de las 12 investigaciones, usando una vez más un flujo de Pentaho Data Integration. Con el uso de esta herramienta se consiguió que el proceso de integración funcionara siempre como se esperaba, ya que en el momento que se decidió incluir las nuevas variables `localizacion_metastasis`, `numero_admin_tratamiento`, `habitos_toxicos` y `concentracion_egf`, se actualizaron los flujos de transformación correspondientes, se volvieron a ejecutar y se obtuvieron los nuevos resultados, con mínimos cambios.

Una vez que estuvieron los datos limpios de los distintos ensayos a integrar, se seleccionaron las variables según la estructura definida para esta integración de datos. Posteriormente se realizaron consultas SQL sobre los datos limpios almacenados, permitiendo determinar que variables se recogen en cada uno de los ensayos del conjunto de datos que se pretende integrar y en qué modelo se encuentra cada variable. Para esto se utilizó el editor de código fuente de Metabase.

Una vez identificadas y localizadas las variables que se definieron para la integración, mediante una consulta SQL de tipo inner join, teniendo como llave primaria el identificador único del paciente (variable `codigo_paciente`), se construyó una tabla para cada estudio que contiene estas variables.

Con la tabla construida para cada nuevo ensayo a integrar, se adicionaron estos datos al conjunto de datos de integración. Para esto se utilizó un proyecto de integración de Pentaho Data Integration.

En este proyecto se cargaron los datos de las tablas integradas para cada ensayo, se generaron las variables faltantes en cada uno de ellos para unificar su estructura y se realizó una integración vertical mediante una operación de unión, integrando en un único conjunto de datos toda la información de las distintas investigaciones. El conjunto de datos resultantes quedó almacenado en un área de la base de datos destinada a albergar los conjuntos de datos listos para ser analizados. Este proceso de integración se muestra en la Figura 2.

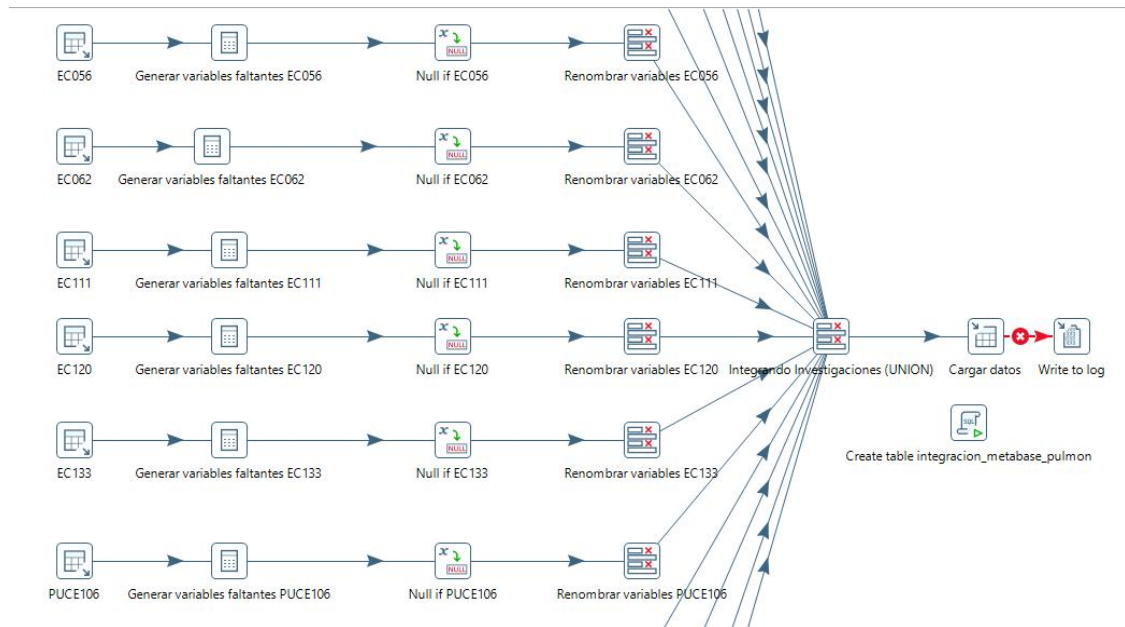


Fig.2 Construcción del conjunto final de datos

El uso de la base de datos PostgreSQL permitió sin problemas el almacenamiento centralizado de los modelos seleccionados de las 12 investigaciones a integrar. Esto fue posible gracias a que los datos clínicos son estructurados, por lo que pueden ser almacenados usando un gestor de bases de datos relacional.

En este contexto, el hecho de disponer una herramienta de almacenamiento que permita la consulta de los datos mediante el lenguaje SQL es muy beneficioso, dado que este es un lenguaje potente, ampliamente utilizado desde hace años y tiene una gran base de conocimiento, por lo que resultó a los especialistas de manejo de datos bastante fácil su asimilación y uso. En este sentido, Metabase fue de gran ayuda gracias a su editor de código SQL, el cual presenta completamiento de código. La posibilidad que brinda esta herramienta para guardar las consultas en un espacio de trabajo compartido ayudó a hacer el proceso de consultas reproducible y eficiente.

Pentaho Data Integration posibilitó realizar la transformación de datos de manera bastante sencilla, siendo su funcionamiento rápidamente asimilado por los especialistas de gestión de datos. Los componentes que proporciona la herramienta para la transformación de datos permitieron resolver todos los problemas de calidad encontrados. El desarrollo de flujos de transformación modulares en Pentaho Data Integration posibilitó que los especialistas pudieran trabajar organizadamente.

Dado que el proceso de transformación se realizó con esta herramienta quedó automatizado cada cambio realizado en los datos, haciendo este proceso reproducible y auditable. Cualquier especialista puede acceder a los flujos de transformación creados por el resto de los especialistas y modificar o adicionar nuevos cambios, como por ejemplo incluir nuevas variables en un proceso de integración. Luego de los cambios realizados se puede volver a ejecutar la transformación actualizándose las tablas de datos correspondientes.

El diccionario de términos del negocio fue decisivo en el proceso de integración de los datos. El mismo permitió que los especialistas en el manejo de datos, pudieran seguir un estándar común

para realizar las tareas de limpieza y homogeneización de los nombres de las variables, sus valores y tipo de datos, posibilitando la integración de estas variables.

Validación mediante criterio de expertos

Para la validación de la metodología de integración mediante el criterio de expertos, se seleccionó un panel de expertos variado, compuesto por 3 especialistas en estadística, 3 promotores de investigación clínica y 3 especialistas en manejo de datos. En la elección de los especialistas, se consideró una experiencia laboral de más de 5 años en investigaciones clínicas. Del total de expertos, 5 poseen la categoría científica de Máster en Ciencias, y 4 cuentan con el título de Doctor en Ciencias, lo que refleja un elevado nivel de competencia en el tema.

Se realizó una encuesta con escala Likert compuesta por 7 preguntas, para obtener las opiniones de los especialistas con respecto a los beneficios de la integración de datos clínicos. Cada experto expresó su valoración de cada indicador mediante la siguiente escala. Se representan por 5- Muy de acuerdo, 4- De acuerdo, 3- Ni de acuerdo ni en desacuerdo, 2- En desacuerdo y 1- Completamente en desacuerdo.

En la Tabla 1 se muestra la encuesta realizada:

Tabla 1. Encuesta aplicada por criterio de expertos

Preguntas	Muy de acuerdo	De acuerdo	Ni de acuerdo ni en desacuerdo	En desacuerdo	Completamente en desacuerdo
¿Cree usted que la integración de datos mediante esta metodología supera a la integración de manera aislada?					
¿Considera oportuna esta nueva metodología para la integración de datos?					
¿Cree que sea importante la implementación de estas tecnologías para el plan estratégico del CIM?					
¿Cree usted que la integración de datos clínicos ayuda a comprender, limpiar y entregar datos confiables?					
¿Considera usted que la arquitectura tecnológica es lo suficientemente flexible para absorber los cambios?					
¿Piensa que se identificaron correctamente las necesidades de integración de datos y se alinean con los objetivos del CIM?					
¿Cree que sería factible extender la integración hacia más bases de datos?					

La aplicación de esta encuesta, tuvo como objetivo obtener una valoración de los especialistas en cuanto a la efectividad de la metodología propuesta, a través de la medición de las siguientes variables:

- Identificación de necesidades de integración de datos.
- Importancia de la definición de las tecnologías.
- La capacidad de la arquitectura tecnológica para adaptarse a cambios.
- Adaptación a las necesidades para optimizar la calidad y eficiencia en la gestión de datos.
- Criterio de aceptación sobre la metodología propuesta.

Para el procesamiento de los resultados se calculó la frecuencia de cada categoría de la escala de Likert definida en la encuesta realizada y se obtuvieron los porcentajes de concordancia de cada categoría de acuerdo a las formulaciones propuestas. Para calificar los resultados según la puntuación más alta y la que mayor importancia tiene se calculó en un índice porcentual (IP), para medir la opinión que tienen los especialistas y si están conformes con las acciones emprendidas hasta ahora para mejorar la gestión de datos.

Se define la interpretación de los resultados como:

- Muy Favorable si el índice es mayor que 80%
- Favorable si el índice está entre 80% y 50%
- Insatisfacción si el índice es de 49% y 25%
- Muy Desfavorable si el índice es menor que 25%

La Figura 4 muestra que el índice porcentual relacionado con la valoración de los expertos, sobre los aspectos planteados, es superior a 80% en la mayor parte de los casos. Este resultado evidencia el resultado muy favorable a la valoración que los expertos le otorgan a la metodología propuesta.

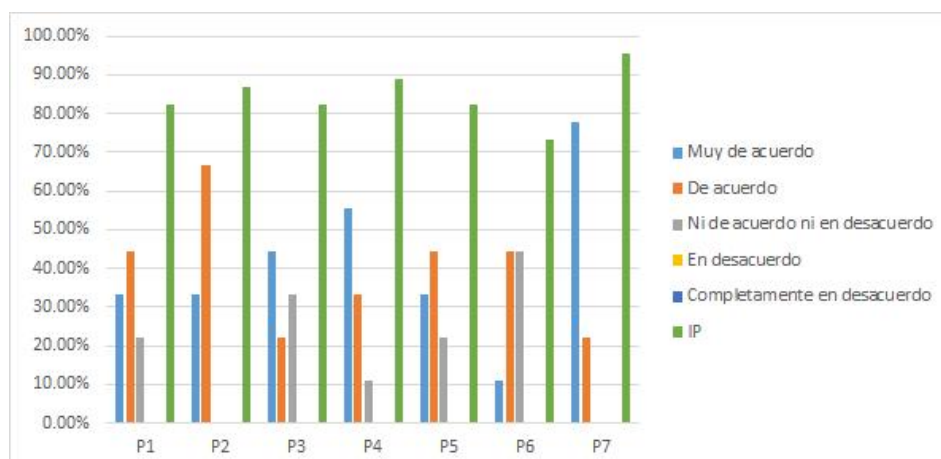


Fig.3 Índice porcentual relacionado con la valoración de los expertos

Los resultados obtenidos a través del escalamiento de Likert demuestran que las características y factibilidad de la metodología propuesta, recibieron una valoración alta por parte de los expertos, encontrándose de acuerdo con la efectividad de la integración de datos clínicos.

4. Conclusiones

A partir de la revisión del estado del arte con respecto a la integración de datos clínicos y teniendo en cuenta las características del CIM se desarrolló una metodología para la integración de datos clínicos del centro. Esta metodología abarca la caracterización de datos, el proceso de diseño, la implementación y la validación, donde se incluye la utilización de indicadores para medir la validez de los datos en la integración de cada estudio. Se validó esta propuesta metodológica mediante la aplicación de un caso de estudio consistente en la integración de 12 ensayos para cáncer de pulmón y de una encuesta utilizando el método de expertos. Como resultado de la aplicación del caso de

estudio se concluye que la metodología propuesta y herramientas seleccionadas jugaron un papel sumamente importante en la integración de los datos, facilitando tecnológicamente esta tarea. Por su parte la validación mediante el método de expertos arrojó una valoración alta por parte de los expertos, encontrándose de acuerdo con la efectividad de la metodología. La metodología diseñada permite obtener una integración efectiva de los datos, y un aumento de su calidad, redundando en mejoras en cuanto al proceso de análisis de los datos.

Referencias

1. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics*. 2018;114:57-65. doi:10.1016/j.ijmedinf.2018.03.013
2. Murphy EC, Ferris FL, O'Donnell WR. An Electronic Medical Records System for Clinical Research and the EMR–EDC Interface. *Invest Ophthalmol Vis Sci*. 2007;48(10):4383-4389. doi:10.1167/iovs.07-0345
3. Definition of clinical research - NCI Dictionary of Cancer Terms - NCI. February 2, 2011. Accessed May 23, 2022. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/clinical-research>
4. Gorrry C. The ABCs of Clinical Trials in Cuba. *MEDICC Rev*. 2016;18(3):9-14. doi:10.37757/MR2016.V18.N3.3
5. EpiData. EpiData General Flyer. http://www.epidata.dk/downloads/epidataflyer_general.pdf. May 17, 2022. Accessed May 17, 2022. http://www.epidata.dk/downloads/epidataflyer_general.pdf
6. Cardona MA, Hernández JLP, Valdés LS, et al. Uso del sistema XAVIA SIDEDEC en las investigaciones clínicas durante la pandemia de Covid19. *Revista de Información científica para la Dirección en Salud INFODIR*. 2022;(1). <https://revinfodir.sld.cu/index.php/infodir/article/view/1302>
7. Vega Izaguirre L, Quintana Díaz VM, Tamayo Peña R, et al. Sistema para el manejo de datos de Ensayos Clínicos XAVIA SIDEDEC. *Revista Cubana de Informática Médica*. 2021;13(1). Accessed May 17, 2022. http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S1684-18592021000100005&lng=es&nrm=iso&tlng=pt
8. Pentaho Data Integration: ¿Qué es y para qué sirve? Entel Comunidad Empresas. April 12, 2021. Accessed April 25, 2023. <https://ce.entel.cl/articulos/pentaho-data-integration/>
9. Ram K. Git can facilitate greater reproducibility and increased transparency in science. Published online 2013.
10. Referencia de datos | Metabase Learn. Accessed March 13, 2025. <https://www.metabase.com/learn/metabase-basics/querying-and-dashboards/data-browser>
11. Garrote PR, Rojas M del C. La validación por juicio de expertos: dos investigaciones cualitativas en Lingüística aplicada. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*. 2015;(18):124-139. doi:10.26378/rnlael918259
12. Taherdoost H. What Is the Best Response Scale for Survey and Questionnaire Design; Review of Different Lengths of Rating Scale / Attitude Scale / Likert Scale. *International*

Journal of Academic Research in Management (IJARM). 2019;8. Accessed March 13, 2025.
<https://hal.science/hal-02557308>

Conflicto de Intereses

Los autores declaran que no poseen conflicto de intereses.

Contribución de los autores

1. Conceptualización
2. Curación de datos
3. Adquisición de fondos: La investigación no requirió financiamiento externo
4. Investigación
5. Metodología
6. Administración del proyecto
7. Recursos
8. Supervisión
9. Validación
10. Redacción – revisión y edición

Arasay Montes De Santis <https://orcid.org/0000-0002-0206-410X>

Conceptualización y sistematización de las ideas centrales del estudio; elaboración del marco teórico, y la recopilación y análisis de datos; aplicación de técnicas de validación y análisis estadístico; interpretación crítica de los resultados y redacción del manuscrito original.

Jorge Luis Palomina Hernández <https://orcid.org/0000-0001-7527-8478>

Conceptualización y sistematización de las ideas fundamentales de la investigación; revisión crítica y perfeccionamiento de las versiones preliminares; redacción de secciones clave del manuscrito; la elaboración de las conclusiones y adaptación del contenido a las normas de publicación requeridas por la revista.