CLIAS

CENTRO DE INTELIGENCIA ARTIFICIAL Y SALUD PARA AMÉRICA LATINA Y EL CARIBE

Experiencias y aprendizajes sobre prácticas de producción, estandarización y uso de datos en salud en una comunidad de conocimiento

30 DE DICIEMBRE DE 2025

Comunidad de Conocimiento

Una iniciativa de CLIAS

IECS INSTITUTO DE EFECTIVIDAD CLINICA Y SANITARIA

CiPS IMPLEMENTACIÓN E INNOVACIÓN EN POLÍTICAS DE SALUD

IDRC · CRDI Canadá

UK International Development Partnership | Progress | Prosperity

Contenido

01. Introducción

La Comunidad de Conocimiento del Centro de Inteligencia Artificial y Salud para América Latina y el Caribe (CLIAS) constituye un espacio regional de intercambio, análisis y co-creación orientado a promover el desarrollo y uso responsable de la IA en el ámbito de la salud. El Grupo de Trabajo "Datos y Métodos", que es parte de esa Comunidad, funciona como un espacio específico dedicado a los desafíos técnicos, éticos y operativos vinculados al uso de datos. Desde la generación y gestión de datos hasta el despliegue de modelos en entornos reales, el grupo analiza prácticas, identifica obstáculos y promueve enfoques responsables, inclusivos y adaptados al contexto regional, con el objetivo de mejorar la calidad, confiabilidad y aplicabilidad de las soluciones de IA en salud, bajo principios de equidad, transparencia y protección de derechos.

La gestión y uso de datos en salud se han convertido en un componente central para el diseño de políticas públicas, la mejora de la atención y la producción de conocimiento. Sin embargo, este potencial convive con múltiples desafíos: la heterogeneidad de las fuentes de datos, la calidad y completitud de los registros, la presencia dominante de datos no estructurados y las tensiones entre marcos conceptuales clínicos, sociales y operativos.

Este documento sistematiza aprendizajes derivados de dos experiencias desarrolladas en el marco de la Comunidad, ambas presentadas y discutidas colectivamente.

El objetivo del documento es reflexionar sobre la calidad de los datos en salud, las estrategias metodológicas y conceptuales involucradas en su producción, estandarización y uso, así como sobre las oportunidades emergentes asociadas a la interoperabilidad y al empleo de datos sintéticos.

Las experiencias analizadas tienen orígenes de datos y abordajes metodológicos distintos, pero comparten problemáticas estructurales:

- Una experiencia enfocada en salud materna, basada en la transformación de datos clínicos y la generación de datos sintéticos bajo el modelo OMOP (*Observational Medical Outcomes Partnership*) ;

- Otra centrada en la detección automática de consultas de salud sexual y reproductiva en personas con discapacidad, apoyada en datos no estructurados de historias clínicas electrónicas y procesos de anotación manual.

A partir de estas experiencias, el documento sistematiza aprendizajes y conclusiones que aportan a la reflexión colectiva sobre los modos de producir, estructurar y utilizar datos en salud.

02. Enfoques y consideraciones sobre la calidad de los datos en salud

La utilización de datos en salud implica comprender que los registros disponibles son el resultado de procesos complejos de atención, documentación y gestión. Su calidad, en términos de consistencia, completitud y validez, depende tanto de las infraestructuras tecnológicas como de las prácticas institucionales y de los supuestos conceptuales que estructuran el registro clínico.

Desde la experiencia de la CDC, la calidad de los datos en salud puede analizarse a partir de al menos cuatro dimensiones interrelacionadas:

- **CALIDAD DE LOS DATOS:** entendida como la consistencia interna, la completitud de los registros, la validez clínica y la coherencia temporal. Deficiencias en cualquiera de estos aspectos limitan la interpretabilidad y el uso analítico de la información.

- **CANTIDAD DE DATOS:** particularmente relevante para el desarrollo de modelos analíticos y de inteligencia artificial, donde la escasez de registros puede restringir la capacidad de generalización y detección de patrones.

- **CONTEXTUALIZACIÓN DE LOS DATOS:** entender cómo, por quiénes y con qué propósito se generan los registros, evitando lecturas que no contemplen las prácticas concretas de atención.

- **ENFOQUE INTERDISCIPLINARIO:** articulación de saberes clínicos, sociales, técnicos y éticos desde las etapas iniciales de los proyectos, como condición para mejorar tanto la calidad como la relevancia de los datos.

Entre las estrategias utilizadas para abordar estas dimensiones se incluyen metodologías estructuradas de análisis (como CRISP-DM), procesos de estandarización semántica, técnicas de procesamiento de lenguaje natural aplicadas a datos no estructurados y, en determinados contextos, el uso de datos sintéticos como complemento a los datos reales.

03. Experiencia 1: Estandarización y datos sintéticos en salud materna

La primera experiencia se centró en el estudio de la morbilidad materna, abordando dificultades estructurales frecuentes en el uso de datos clínicos reales: dispersión de fuentes de información, utilización de vocabularios orientados a fines administrativos o de facturación, predominio de texto libre y restricciones de acceso derivadas de consideraciones éticas y de privacidad.

Frente a este escenario, el proyecto adoptó el modelo de datos OMOP como estrategia central. OMOP es un modelo de datos común diseñado para estructurar información clínica de manera estandarizada, apoyado en vocabularios controlados y en una arquitectura que separa los datos de las herramientas analíticas. Su implementación permite que datos provenientes de distintas instituciones y sistemas puedan analizarse con herramientas compartidas y compararse entre sí.

Desde la perspectiva de la comunidad, la adopción de OMOP no constituye únicamente una decisión técnica, sino una estrategia para mejorar la calidad de los datos y fomentar la interoperabilidad, reduciendo la dependencia de modelos propietarios y promoviendo prácticas reproducibles y colaborativas en investigación en salud.

El proceso de transformación de los datos al modelo OMOP implicó decisiones complejas, mapeos manuales y la articulación de conocimientos clínicos y técnicos, especialmente en contextos donde los vocabularios disponibles no fueron concebidos con fines de investigación. Esta experiencia evidenció que la estandarización es un proceso interpretativo y situado, que requiere acuerdos y trabajo interdisciplinario sostenido.

3.1 LA INCORPORACIÓN DE DATOS SINTÉTICOS COMO ESTRATEGIA METODOLÓGICA

En el marco de esta experiencia se identificó que la cantidad y la variabilidad de los datos reales disponibles no alcanzaban para sostener el desarrollo y la evaluación de modelos con mayor robustez, lo que llevó a explorar la generación de datos sintéticos mediante *Synthea*[1] como estrategia complementaria.

[1] ***Synthea**: herramienta de código abierto para la generación de datos clínicos sintéticos, utilizada para pruebas, desarrollo metodológico y formación sin comprometer datos reales.*

El uso de datos sintéticos permitió:

- experimentar con definiciones clínicas y fenotipos sin comprometer la privacidad de las personas,
- evaluar pipelines analíticos y procesos de transformación a OMOP,
- y compartir datasets, código y resultados de manera abierta.

No obstante, la experiencia también evidenció límites relevantes, en particular la dificultad de capturar la heterogeneidad clínica real y el riesgo de generar poblaciones sintéticas excesivamente homogéneas si el modelado no incorpora suficiente complejidad clínica y contextual.

Estos aprendizajes permitieron comprender los datos sintéticos no como un reemplazo de los datos reales, sino como una herramienta complementaria para el desarrollo metodológico, la puesta a prueba de enfoques y la formación, particularmente relevante en contextos donde el acceso a datos clínicos se encuentra restringido.

04. Experiencia 2: Datos no estructurados, discapacidad y salud sexual y reproductiva

La segunda experiencia abordó un problema distinto pero complementario: la detección automática de consultas de salud sexual y reproductiva en personas con discapacidad a partir de historias clínicas electrónicas con predominio de datos no estructurados.

En este caso, el desafío central consistió en traducir conceptos complejos del campo de la discapacidad y los derechos sexuales y reproductivos, en definiciones operativas consensuadas por perspectivas diversas sobre la problemática, reconocidas de forma consistente por diferentes personas expertas y susceptibles de ser identificadas y utilizadas por modelos computacionales.

4.1 DATOS NO ESTRUCTURADOS Y ABORDAJE METODOLÓGICO

Los datos no estructurados presentan múltiples dificultades, entre las que se incluyen la ambigüedad semántica, la variabilidad en el lenguaje clínico, la ausencia de campos específicos

para registrar determinados fenómenos y las tensiones entre distintos modelos de interpretación, en particular entre el modelo médico y el modelo social de la discapacidad.

El abordaje metodológico incluyó mesas de trabajo participativas, definiciones consensuadas de conceptos, selección dirigida de registros y un proceso iterativo de anotación manual para consolidar criterios consistentes entre diferentes personas expertas. Este proceso permitió visibilizar que la representación de la discapacidad en la historia clínica electrónica suele ser fragmentaria, implícita y atravesada por supuestos, lo que impacta directamente en la calidad de los datos disponibles.

4.2 ANOTACIÓN DE DATOS Y CONSTRUCCIÓN DE ETIQUETAS

En esta experiencia, el proceso de anotación de datos y construcción de etiquetas constituyó el núcleo metodológico del abordaje. Una etiqueta se definió como una categoría asignada a un fragmento de texto o a un registro completo de la historia clínica electrónica para señalar la presencia de un fenómeno de interés, como una consulta de salud sexual y reproductiva, una mención de discapacidad o la coexistencia de ambos.

A medida que avanzó el proceso de anotación, se hizo necesario explicitar y sistematizar los criterios utilizados. De este modo, se elaboró un manual de anotación, construido de manera colectiva e iterativa, que permitió traducir los acuerdos conceptuales alcanzados en criterios operativos concretos. En este manual se definieron:

- las definiciones operativas de cada etiqueta,
- los criterios de inclusión y exclusión,
- las reglas para resolver ambigüedades,
- y los mecanismos para documentar desacuerdos y revisiones.

Este trabajo no sólo habilitó el entrenamiento de modelos automáticos, sino que generó un recurso metodológico reutilizable, que transparenta las decisiones tomadas y contribuye a mejorar la calidad y trazabilidad del análisis de datos en salud.

4.3. ANÁLISIS DE SESGOS

Con los datos etiquetados automáticamente se entrenaron diferentes modelos predictivos. Se analizó el comportamiento de estos modelos, en particular, la distribución de errores desde un punto de vista de equidad. Se observó que se producían más errores en clases minoritarias: en el caso de hombres, no se identificaron registros sobre salud sexual y reproductiva en una proporción mayor que en el resto de clases. En cambio, se observó que se identificaban más registros de

personas con discapacidad erróneamente como de consultas sobre salud sexual y reproductiva en el caso de adultos, más que en otras franjas de edad. Y la casuística más grave: en niños se produjo una mayor cantidad de errores al no identificar como registros de atención sobre salud sexual casos que sí lo eran. Estos errores pueden ser especialmente graves, al no identificar población especialmente vulnerable en situaciones que pueden ser potencialmente complejas (violencia, trastornos graves).

Este análisis caracteriza áreas problemáticas en este tipo de aplicación de inteligencia artificial, y permite identificar oportunidades de mejora para futuros desarrollos del proyecto, por ejemplo, obteniendo más datos específicamente para estas clases minoritarias.

05. Conclusiones y aprendizajes compartidos con la CDC

La A partir del trabajo colectivo y del intercambio en la comunidad, se identifican los siguientes aprendizajes clave:

- La **estandarización de datos mediante modelos comunes como OMOP** constituye una decisión metodológica y estratégica para mejorar la calidad de los datos, favorecer la interoperabilidad y habilitar la producción colaborativa de conocimiento.
- Los **datos sintéticos** se consolidan como una herramienta complementaria relevante, especialmente para el desarrollo metodológico, la validación de enfoques analíticos y la formación, sin sustituir el valor de los datos reales.
- El abordaje de **datos no estructurados** requiere metodologías específicas que reconozcan la complejidad del lenguaje clínico y las prácticas de registro.
- La **construcción de etiquetas y manuales de anotación** emerge como un aprendizaje metodológico central, al permitir operacionalizar conceptos complejos, mejorar la calidad de los datos y generar recursos reutilizables.
- La **calidad y la cantidad de datos** deben abordarse de manera integrada, ya que grandes volúmenes de datos de baja calidad pueden reproducir sesgos e invisibilizar poblaciones.
- El principal valor de las experiencias sistematizadas reside tanto en los resultados técnicos como en los **procesos, recursos y aprendizajes metodológicos** generados.

AUTORÍA:

Analía Pastrana y **Mario Rossi** - Coordinadores del Grupo de Trabajo de Datos y Métodos de la Comunidad de Conocimiento del CLIAS.

COLABORARON:

Cintia Cejas - Coordinadora General del CLIAS. **Martín Saban** - Investigador del CLIAS. **Cender Quispe** - Asistente de investigación del CLIAS. **Victoria Bruschini** - Gestora de Comunidades del CLIAS.

AGRADECIMIENTOS:

Agradecemos a los integrantes de la Comunidad que colaboraron con la propuesta de temas y la escritura del presente documento, aportando ideas, consensos y desarrollos sustantivos. Especialmente a: ***Ever Augusto Torres Silva*** *- Netux SAS, Colombia;* ***Laura Alonso Alemany*** *- Facultad de Matemática, Astronomía, Física y Computación, Universidad Nacional de Córdoba, Argentina;* ***María Cecilia Palermo*** *- CONICET-IIGG/UBA;* ***Verónica Xhardez****- ARPHAI/CIECTI - Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación;* ***Sabrina Laura López*** *- ARPHAI - Argentinian Public Health Research on Data Science and Artificial Intelligence for Epidemic Prevention y* ***María Victoria Tiseyra*** *- CONICET.*

CLIAS

CENTRO DE INTELIGENCIA
ARTIFICIAL Y SALUD
PARA AMÉRICA LATINA
Y EL CARIBE

Comunidad de Conocimiento

Una iniciativa de CLIAS

IECS INSTITUTO DE EFECTIVIDAD CLINICA Y SANITARIA

CIPS IMPLEMENTACIÓN E INNOVACIÓN EN POLÍTICAS DE SALUD

IDRC · CRDI Canada

UK International Development
Partnership | Progress | Prosperity