

## **Información clínica en español: cómo aprovecharla sin poner en riesgo la privacidad**

Daniel Mauricio Ospina Leguizamón<sup>1</sup>

Director: Jhon Alexis Piracoca Arcos<sup>2</sup>

### **Resumen**

El crecimiento de la información clínica en formato no estructurado, como las historias clínicas, las notas médicas y los reportes narrativos, deja una gran tarea al tener como principal reto su análisis y aprovechamiento. En este caso, el procesamiento de lenguaje natural (NLP) y los modelos de lenguaje pueden comprenderse como herramientas clave y de gran repercusión para poder extraer y tener una comprensión más directa y exacta de su contenido sin dejar de lado la protección de información relevante expuesta en los mismos. Se desarrolló este artículo de divulgación mediante una revisión de literatura descriptiva y cualitativa de cincuenta fuentes académicas, técnicas y normativas, con el objetivo de exponer la importancia de los datos no estructurados en salud, describir las principales aplicaciones del NLP en este campo y analizar sus desafíos más relevantes, como la precisión de los modelos, la privacidad de los pacientes y la desidentificación de información sensible. Los hallazgos muestran que gran parte de la información clínica más valiosa reside en el texto libre, que las técnicas basadas en transformadores y modelos de lenguaje permiten extraerla con precisión creciente también en español y que la desidentificación y la interoperabilidad son condiciones indispensables para su uso ético. Se concluye que los datos no estructurados, manejados de forma segura, representan una oportunidad real para fortalecer la investigación y con ellas dar un aporte significativo en

---

<sup>1</sup> Estudiante, Ingeniería de Sistemas, Universidad Santo Tomás, Tunja, Colombia. [daniel.ospinal@usantoto.edu.co](mailto:daniel.ospinal@usantoto.edu.co)

<sup>2</sup> Docente, Facultad de Ingeniería de Sistemas, Universidad Santo Tomás, Tunja, Colombia.  
[jhon.piracoca@usantoto.edu.co](mailto:jhon.piracoca@usantoto.edu.co)

decisiones clínicas, como también dejar una base clara de cómo avanzar en la gestión de la información de diferentes formatos.

**Palabras clave:** procesamiento de lenguaje natural, datos no estructurados, historia clínica electrónica, desidentificación, interoperabilidad.

### **Abstract**

The growth of unstructured clinical information, such as medical records, clinical notes, and narrative reports, has created new challenges for its organization, analysis, and use in the healthcare sector. In this context, natural language processing (NLP) and language models have become key tools for extracting, structuring, and protecting the relevant information contained in medical texts. This popular science article was developed through a descriptive, qualitative literature review of fifty academic, technical, and regulatory sources, with the aim of explaining the importance of unstructured data in healthcare, describing the main applications of NLP in this field, and analyzing its most significant challenges, such as model accuracy, patient privacy, and the de-identification of sensitive information. The findings show that much of the most valuable clinical information resides in free text, that transformer-based techniques and language models can extract it with growing accuracy also in Spanish and that de-identification and interoperability are indispensable conditions for its ethical use. It is concluded that unstructured data, when handled securely, represents a real opportunity to strengthen research, support clinical decision-making, and modernize health information systems.

**Keywords:** natural language processing, unstructured data, electronic health record, de-identification, interoperability.

## Introducción

Supóngase que el personal de la salud, en específico los médicos o encargados a primera mano de pacientes, debe tomar decisiones rápidas sobre un paciente que ya ha sido atendido varias veces en las cuales en cada situación se ha dejado evidencia de cada síntoma, característica o diagnóstico y por ende, en el sistema se registran algunos datos organizados, como la edad, los resultados de laboratorio o los diagnósticos codificados. Sin embargo, gran parte de la información más valiosa no está en formatos estructurados, sino en notas médicas, descripciones de síntomas, antecedentes y observaciones escritas en texto libre. Allí pueden surgir detalles clave para comprender mejor la evolución del paciente. Pero ese mismo texto libre es también uno de los mayores retos para la investigación en salud: aprovechar esa información sin perderla en el desorden y sin poner en riesgo la privacidad de quien la generó.

Desde hace varios años , los datos se reconocen como los recursos más importantes para producir conocimiento y apoyar la toma de decisiones (Dash et al., 2019), y la literatura reciente reafirma su relevancia en el ámbito de la salud (Jerfy et al., 2024; Kim et al., 2024). Sin embargo, no todos tienen el mismo formato: algunos son estructurados, organizados en tablas, filas y columnas donde se puede almacenar aquella información simple y de fácil manejo, ya que se distinguen por tener un grupo o lugar al que pertenecen y pueden almacenarse en gran medida, mientras que otros son no estructurados y se presentan en formatos más abiertos, como texto, imágenes, audio o video (Ministerio de Ciencia, Tecnología e Innovación, 2022). Aunque estos

últimos no se ajustan con facilidad a categorías predefinidas, su fortaleza radica en que describen con mayor detalle situaciones concretas ocurridas en un momento determinado, información que suele almacenarse tal como se genera, sin una clasificación previa. Esta diferencia es relevante, pues los datos estructurados resultan más fáciles de almacenar y analizar, pero los no estructurados suelen contener una riqueza de información que no se puede registrar en una base de datos tradicional (Li et al., 2022). Es decir, lo más valioso no siempre está en la estructura, sino en lo que alguien escribió para explicar un caso específico y único, por lo tanto, merecen mayor atención como objeto de estudio. Esos detalles valiosos, muchas veces, solo pueden encontrarse en los registros directos del personal que está presente en cada momento de la atención.

Esta situación se vuelve especialmente relevante en el sector salud. Las historias clínicas electrónicas no solo reúnen datos organizados, como signos vitales, fórmulas o resultados diagnósticos, sino también notas narrativas que médicos y otros profesionales elaboran al momento de su turno o jornada laboral. Dónde su único objetivo es explicar a detalle los acontecimientos o circunstancias que pueden llegar a suceder con cada paciente. Ese texto libre puede contener contexto, matices y detalles que los registros estructurados no siempre reflejan. Por eso mismo el gran interés por estudiar los datos no estructurados. Diversos estudios han mostrado que esta información complementa de manera importante a los registros estructurados. Además, puede aportar un panorama más completo del estado de salud de los pacientes (Kim et al., 2024; Seinen et al., 2025).

Sin embargo, aprovechar estos datos no es sencillo, ¿Por qué? A diferencia de una tabla bien organizada, el texto clínico varía mucho de un profesional a otro. Puede incluir abreviaturas, expresiones ambiguas, errores de digitación o distintas formas de describir una misma situación. Esto hace que su análisis sea más complejo y que extraer información útil requiera herramientas capaces de interpretar lenguaje natural y en ciertas ocasiones preparado para ciertas áreas como lo sería en este caso el sector médico. En este punto cobra relevancia el procesamiento de lenguaje natural, una rama de la inteligencia artificial que permite trabajar con textos de forma automática. Con estas herramientas es posible identificar patrones, extraer entidades, clasificar información y convertir documentos narrativos en datos útiles para la investigación (Jerfy et al., 2024). Estas herramientas han empezado a mostrar un gran potencial, especialmente cuando los investigadores las aplican a historias clínicas y otros registros médicos en texto libre (Koleck et al., 2019).

Pero el problema no es solo técnico. En salud, trabajar con datos también significa trabajar con información íntima y sensible. Una historia clínica puede contener nombres, fechas, direcciones, diagnósticos, antecedentes personales y otros elementos que no deben circular libremente. En Colombia, la normativa protege de manera especial este tipo de información (Congreso de Colombia, 1981, 2012). Además, exige aplicar medidas que eviten la identificación del paciente al tratar la información (Ministerio de Salud, 1999). La anonimización y la desidentificación deben entenderse no solo como algo obligatorio, sino como la manera correcta y ética en que los profesionales usan cualquier información. Si los datos no se protegen adecuadamente, el riesgo no solo es legal, sino también humano: se compromete la confianza del paciente y se debilita su uso.

Cuando los datos se manejan de forma segura, sus beneficios pueden ser muy grandes y sus aportes de gran magnitud. El análisis de información clínica no estructurada puede ayudar a encontrar patrones de enfermedades y mejorar el seguimiento de pacientes o de aquellas situaciones donde algo ciertamente pequeño es por lo general una reacción o síntoma fuera de lo usual. También apoya en las decisiones clínicas que se presentan comúnmente y abre nuevas posibilidades para la investigación médica (Hossain et al., 2023). En ese caso, el desafío actual no consiste en decidir si se deben usar estos datos o no, sino en aprender a utilizarlos bien. Esto implica contar con herramientas adecuadas, aplicar criterios de protección de datos y adoptar una visión que permita transformar información dispersa en conocimiento útil. Desde esta perspectiva, los datos no estructurados dejan de ser un problema desordenado para convertirse en una oportunidad real de innovación y avance.

A pesar de todo, la literatura evidencia un conocimiento poco visto. Los investigadores han desarrollado y validado la mayoría de las herramientas de procesamiento de lenguaje natural para el ámbito clínico en inglés, que su adaptación al español clínico y, en particular, a un contexto como el colombiano sigue siendo limitada y dispersa. Este artículo busca contribuir a cerrar esa brecha.

Por lo cual el objetivo principal de este artículo de divulgación es explicar, de forma accesible, cómo aprovechar la información clínica no estructurada con énfasis en el idioma español sin comprometer la privacidad de los pacientes. Para desarrollar esta idea, el documento se organizó de la siguiente manera. La sección de metodología describe la revisión de literatura realizada. La

sección de resultados presenta los hallazgos sobre el valor del texto clínico, las bases técnicas de los modelos de lenguaje, las aplicaciones del NLP para extraer y estructurar información, el procesamiento del español clínico, la desidentificación, la interoperabilidad y el marco normativo colombiano. La discusión contrasta estos hallazgos con la literatura existente. Finalmente, las conclusiones sintetizan los aportes, las limitaciones y las líneas de trabajo futuras.

### **Metodología**

Este artículo tuvo revisión de literatura, informes de grandes colaboradores del área, investigaciones y más estudios complementarios en enfoque descriptivo y cualitativo, orientada a analizar el papel de los datos no estructurados, su aprovechamiento en investigación y la importancia fundamental de su anonimización o desidentificación cuando se trata de información sensible médica. Para ello, se revisaron estudios previos, documentos institucionales y normativa relacionada con la gestión de datos clínicos, especialmente aquellos vinculados con historias clínicas electrónicas, notas médicas y otros registros narrativos. En el plano normativo se consideraron, entre otros, la ley que crea la historia clínica electrónica interoperable (Congreso de Colombia, 2020), la reglamentación de los datos clínicos para la interoperabilidad (Ministerio de Salud y Protección Social, 2021), la adopción del Resumen Digital de Atención en Salud (Ministerio de Salud y Protección Social, 2025) y las guías técnicas sobre anonimización de datos (DANE, 2024).

A partir de esta revisión, se identificaron las principales diferencias entre datos estructurados y no estructurados, así como los retos que surgen en su procesamiento, protección entre otros y

también de igual manera, se examinaron las implicaciones éticas y legales del uso de información en salud, prestando especial atención a la privacidad del paciente y a la necesidad de aplicar procesos de anonimización o desidentificación que permitan aprovechar los datos sin comprometer la identidad de las personas.

La selección de fuentes se centró en publicaciones académicas, documentos técnicos y normas relevantes, principalmente en español e inglés, con prioridad en estudios actuales y directamente relacionados con el ámbito de la salud. La búsqueda se organizó mediante cadenas orientadas a los ejes temáticos del trabajo, tal como se resume en la Tabla 1.

**Tabla 1**

*Cadenas de búsqueda empleadas en la revisión de literatura*

| Cadena de búsqueda                                                               | Total  | Seleccionadas |
|----------------------------------------------------------------------------------|--------|---------------|
| "unstructured data" AND healthcare AND "electronic health records"               | 4.563  | 9             |
| "clinical notes" AND anonymization AND privacy                                   | 5.467  | 11            |
| "Spanish electronic health records" / NLP clínico en español                     | 256    | 6             |
| Revisión complementaria de referencias, documentos normativos y depuración final | —      | 24            |
| Total                                                                            | 10.286 | 50            |

*Fuente: elaboración propia.*

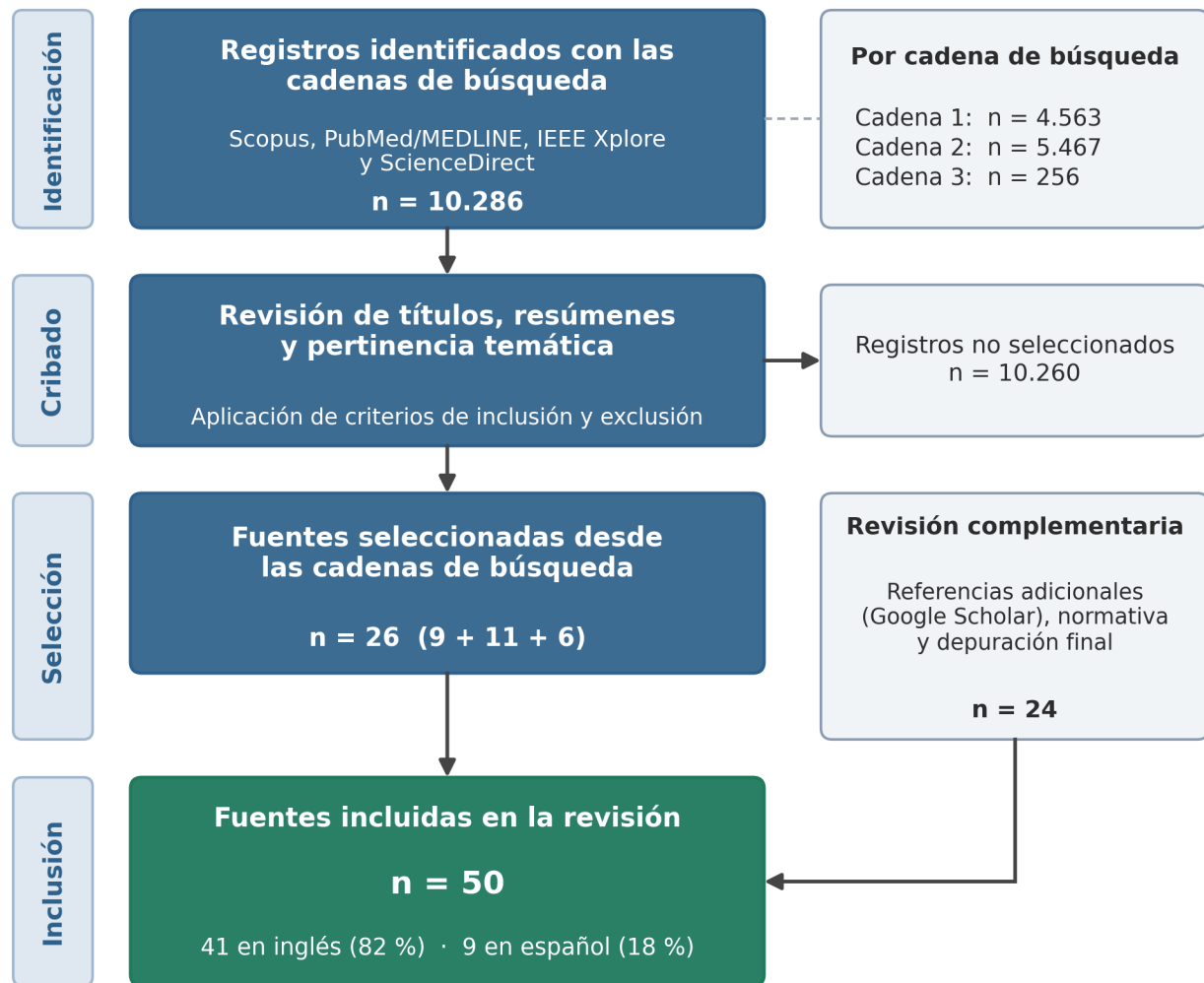
La búsqueda se realizó en bases de datos académicas indexadas, principalmente Scopus, PubMed/MEDLINE, IEEE Xplore y ScienceDirect, complementadas con Google Scholar para la revisión de referencias adicionales. Los criterios de inclusión fueron: publicaciones en español o inglés, relación directa con el procesamiento de lenguaje natural o los modelos de lenguaje aplicados a texto clínico, la desidentificación de datos sensibles o la interoperabilidad de la historia clínica, y prioridad a estudios de los últimos cinco años, sin descartar trabajos fundacionales más antiguos cuando eran pertinentes. Se excluyeron blogs, páginas web sin autor identificable o de gran relevancia y fuentes sin revisión por pares o respaldo institucional. El

proceso de búsqueda y selección se documentó siguiendo principios generales de reporte transparente afines a los de PRISMA, aunque sin constituir una revisión sistemática formal, dado el carácter descriptivo y cualitativo de este artículo de divulgación.

Con base en estas cadenas de búsqueda y contribuciones se identificó un conjunto amplio de publicaciones relacionadas con el uso de datos no estructurados en salud, el procesamiento de lenguaje natural aplicado a historias clínicas electrónicas y la anonimización de información sensible. Tras una revisión inicial de títulos, resúmenes y pertinencia temática, se seleccionaron 50 referencias para el desarrollo del artículo, priorizando aquellas con mayor relación con el objetivo del estudio, actualidad, aplicabilidad al contexto de la salud y aportes sobre texto clínico, desidentificación, normalización e investigación basada en registros médicos. De estas, 26 provienen directamente de las cadenas de búsqueda y 24 de la revisión complementaria de referencias y documentos normativos (Tabla 1). Estas 50 fuentes conforman la totalidad del corpus analizado y corresponden a la lista de referencias del artículo; es decir, todas las fuentes citadas, incluidas las empleadas para contextualizar la introducción, hacen parte de la revisión.

La Figura 1 resume este proceso, desde los registros identificados con las cadenas de búsqueda hasta las fuentes finalmente incluidas en la revisión.

**Figura 1.** Diagrama de flujo del proceso de identificación y selección de fuentes.



*Fuente: elaboración propia a partir de la Tabla 1.*

En cuanto al idioma, de las 50 fuentes seleccionadas, 41 (82 %) están publicadas en inglés y 9 (18 %) en español; estas últimas corresponden a los ocho documentos normativos y técnicos colombianos y al estudio de Báez et al. (2022). Además, cinco estudios (10 %) se centran explícitamente en el procesamiento del español clínico (Báez et al., 2022; García Subies et al., 2024; Mata et al., 2025; Moreno-Barea et al., 2025; Zotova et al., 2026), lo que evidencia el predominio del inglés en la producción científica sobre NLP clínico.

Posteriormente, las fuentes seleccionadas se analizaron de manera cualitativa con el propósito de identificar temas recurrentes en su manejo. En este proceso se organizaron los aportes en varios ejes: la diferencia entre datos estructurados y los que no son estructurados, el valor del texto clínico como fuente de información única, los desafíos del procesamiento en español, los riesgos éticos y legales derivados del manejo de datos sensibles, las estrategias de anonimización o desidentificación y la interoperabilidad de la información clínica.

Este análisis permitió reconocer que tan importante fue la información para comprender la trayectoria clínica de los pacientes. Y que no siempre se encuentra en campos estructurados, sino en notas médicas, reportes narrativos, resúmenes de egreso y otros documentos que en gran medida los toman como simple papeleo. Al mismo tiempo, evidenció que el aprovechamiento de estos datos requiere herramientas especializadas, criterios de calidad y manejo ético, como medidas de protección que reduzcan el riesgo de identificación de las personas.

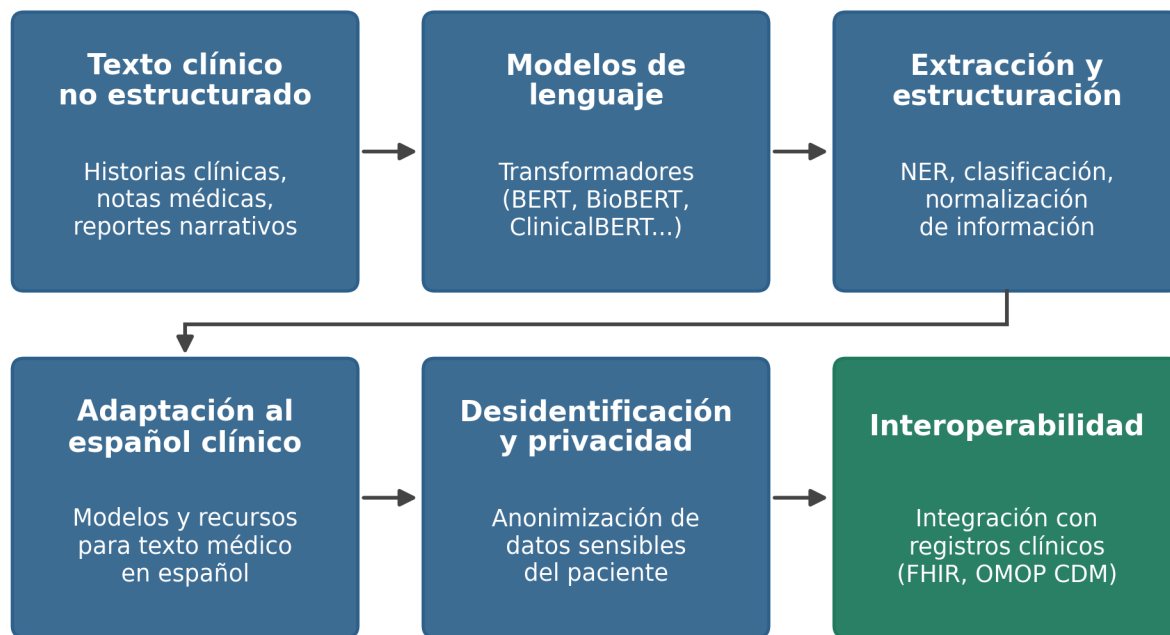
### **Resultados**

En el análisis general de las cincuenta fuentes seleccionadas se organizaron los hallazgos en seis ejes: el valor del texto clínico como fuente de información, las bases técnicas de los modelos de lenguaje, las aplicaciones del NLP y los modelos de lenguaje para extraer información, el procesamiento del español clínico, la desidentificación y la privacidad y construcción de registros. En una visión más fuera de lo específico, varias revisiones sistemáticas confirman el crecimiento sostenido del NLP en salud y su impacto en la toma de decisiones (Jerfy et al., 2024; Hossain et al., 2023), su despliegue a lo largo de toda la trayectoria del paciente, desde el ingreso

hasta el egreso (Klug et al., 2024), y la reciente medida de uso de los grandes modelos de lenguaje en la documentación clínica y en dominios especializados como el pediátrico y quirúrgico (Woo et al., 2026; Daskalo et al., 2026). Para reflejar el conteo de las 50 fuentes seleccionadas, las tablas incluyen una columna de numeración (N.º) consecutiva del 1 al 50.

La Figura 2 resume de manera visual este recorrido: parte del texto clínico no estructurado, pasa por los modelos de lenguaje y las tareas de extracción y estructuración de información sin comprometer su privacidad, continúa con su adaptación al español clínico y la desidentificación de datos sensibles, y culmina en la interoperabilidad de los registros, condición necesaria para su aprovechamiento en investigación y en la toma de decisiones clínicas.

**Figura 2.** Pipeline de procesamiento de lenguaje natural (NLP) en el ámbito clínico.



*Investigación en salud, seguimiento de pacientes y apoyo a decisiones clínicas*

*Fuente: elaboración propia.*

### El texto clínico como fuente de información

Un primer hallazgo consistente es que gran parte de la información clínica útil reside en el texto libre y no en los campos estructurados. Las revisiones sobre NLP de síntomas documentados en narrativas de historias clínicas muestran que estos registros contienen señales relevantes que los campos codificados no capturan (Koleck et al., 2019), y los trabajos sobre estandarización de información clínica no estructurada evidencian la necesidad de sistemas capaces de capturar y normalizar ese contenido (Kreimeyer et al., 2017). Estudios recientes cuantifican esta complementariedad entre códigos estructurados y notas de texto libre, mostrando que ambos aportan información que el otro no contiene (Seinen et al., 2025), y otros advierten sobre los retos y oportunidades del análisis de datos basado en historias clínicas electrónicas (Kim et al.,

2024). Este panorama se sintetiza en las revisiones de NLP neuronal para datos no estructurados en EHR (historia clínica electrónica) (Li et al., 2022) y en trabajos que logran reconstruir la historia natural del paciente integrando información dispersa (Najafabadipour et al., 2020). La Tabla 2 reúne las revisiones y estudios sobre el valor del texto clínico citados en este apartado, junto con las revisiones generales sobre datos y NLP en salud mencionadas en la introducción y al inicio de esta sección; el trabajo de Najafabadipour et al. (2020) se presenta en la Tabla 4, con las demás aplicaciones de extracción.

**Tabla 2**

*Revisiones generales y estudios sobre el valor del texto clínico*

| N.º | Referencia              | Enfoque                                                              | Aporte                                                                                          |
|-----|-------------------------|----------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| 1   | Koleck et al. (2019)    | Revisión sistemática de NLP de síntomas en narrativas de EHR         | Muestra que el texto libre registra síntomas que los campos codificados no capturan.            |
| 2   | Kreimeyer et al. (2017) | Revisión sistemática de sistemas de NLP clínico                      | Evidencia la necesidad de capturar y estandarizar la información clínica no estructurada.       |
| 3   | Seinen et al. (2025)    | Complementariedad entre códigos estructurados y notas de texto libre | Cuantifica la información que aportan las notas frente a los códigos estructurados.             |
| 4   | Kim et al. (2024)       | Análisis de datos basados en EHR                                     | Expone los retos y oportunidades del análisis e interpretación de estos datos.                  |
| 5   | Li et al. (2022)        | Revisión de NLP neuronal para datos no estructurados en EHR          | Sintetiza los métodos neuronales aplicados al texto de la historia clínica.                     |
| 6   | Dash et al. (2019)      | Revisión sobre big data en salud                                     | Describe la gestión, el análisis y las perspectivas de los grandes volúmenes de datos en salud. |
| 7   | Jerfy et al. (2024)     | Revisión del impacto del NLP en salud                                | Documenta el impacto creciente del NLP en la atención en salud y la salud pública.              |
| 8   | Hossain et al. (2023)   | Revisión sistemática de NLP en EHR                                   | Relaciona el uso del NLP en la EHR con la toma de decisiones en salud.                          |
| 9   | Klug et al. (2024)      | Revisión sistemática del NLP clínico                                 | Analiza su aplicación a lo largo de la trayectoria del paciente, del ingreso al egreso.         |
| 10  | Woo et al. (2026)       | Revisión de alcance sobre LLM en la documentación clínica            | Sintetiza el uso de grandes modelos de lenguaje para documentar la atención.                    |

| N.º | Referencia            | Enfoque                            | Aporte                                                 |
|-----|-----------------------|------------------------------------|--------------------------------------------------------|
| 11  | Daskalo et al. (2026) | Revisión sistemática de LLM en EHR | Examina su uso en la atención pediátrica y quirúrgica. |

*Fuente: elaboración propia a partir de las referencias citadas.*

### Bases técnicas: de los modelos de lenguaje a los transformadores

El avance del NLP clínico se apoya en una familia de modelos basados en la arquitectura de transformadores. En términos sencillos, un transformador funciona como un lector muy atento que, en lugar de leer palabra por palabra de forma aislada, analiza toda la oración a la vez y pondera cómo se relaciona cada palabra con las demás para entender su significado según el contexto.

La Tabla 3 resume los modelos de referencia que hicieron posible el procesamiento moderno del texto médico, desde las arquitecturas generales hasta sus adaptaciones al dominio biomédico y clínico.

**Tabla 3**

*Modelos de lenguaje y arquitecturas de referencia para el procesamiento de texto clínico*

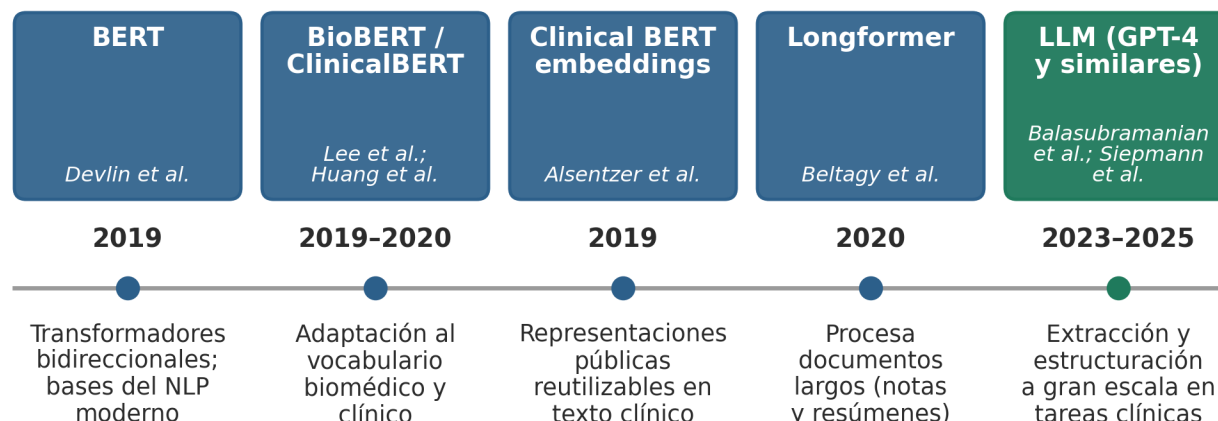
| N.º | Modelo / arquitectura                                                   | Referencia              | Aporte principal                                                                             |
|-----|-------------------------------------------------------------------------|-------------------------|----------------------------------------------------------------------------------------------|
| 12  | BERT ( <i>Bidirectional Encoder Representations from Transformers</i> ) | Devlin et al. (2019)    | Transformadores bidireccionales que sentaron las bases del NLP moderno.                      |
| 13  | BioBERT                                                                 | Lee et al. (2020)       | Adaptación de BERT al dominio biomédico mediante preentrenamiento con literatura científica. |
| 14  | ClinicalBERT                                                            | Huang et al. (2019)     | Modelo entrenado con notas clínicas para predecir el reingreso hospitalario.                 |
| 15  | Clinical BERT embeddings                                                | Alsentzer et al. (2019) | Representaciones públicas especializadas en texto clínico.                                   |
| 16  | Longformer                                                              | Beltagy et al. (2020)   | Transformador para documentos largos, útil ante notas clínicas extensas.                     |

*Fuente: elaboración propia a partir de las referencias citadas.*

La progresión que refleja la Tabla 3 no es únicamente cronológica, sino también conceptual. Los primeros modelos de propósito general, como BERT (Devlin et al., 2019), demostraron que era posible capturar el significado de una palabra a partir de su contexto completo, algo especialmente valioso en medicina, donde un mismo término puede cambiar de sentido según la frase. Sobre esa base, BioBERT (Lee et al., 2020) y ClinicalBERT (Huang et al., 2019) incorporaron el vocabulario y las particularidades del lenguaje biomédico y clínico, mientras que los embeddings clínicos de acceso público (Alsentzer et al., 2019) son representaciones numéricas que ubican cada palabra en una especie de mapa de significados, donde los términos con sentidos parecidos quedan cerca unos de otros, y que los investigadores reutilizan sin partir de cero. Longformer (Beltagy et al., 2020), por su parte, respondió a un problema práctico frecuente: las notas clínicas y los resúmenes de egreso suelen ser largos y superan la capacidad de los modelos tradicionales, por lo que se requieren arquitecturas capaces de procesar documentos extensos sin perder información.

La Figura 3, elaborada a partir de los modelos reseñados en la Tabla 3, sintetiza esta evolución de forma visual: muestra cómo los modelos de propósito general, como BERT, dieron paso a variantes especializadas en el dominio biomédico y clínico (BioBERT, ClinicalBERT), a representaciones reutilizables (Clinical BERT embeddings) y a arquitecturas capaces de procesar documentos extensos (Longformer), hasta llegar a los grandes modelos de lenguaje que hoy se aplican directamente a tareas clínicas de extracción y estructuración de información.

**Figura 3.** Evolución y comparación de modelos de lenguaje para texto clínico.



Fuente: elaboración propia.

### Aplicaciones del NLP y los modelos de lenguaje para extraer y estructurar información

Un conjunto amplio de estudios aplica NLP y modelos de lenguaje a tareas concretas de extracción, clasificación y estructuración de información clínica. La Tabla 4 sintetiza doce trabajos representativos, que abarcan desde la lectura automática de informes de patología y cartas de alta hasta la detección de síntomas y el fenotipado, y que en conjunto muestran una precisión creciente en la conversión de texto narrativo en datos utilizables.

**Tabla 4**

*Estudios sobre extracción y estructuración de información clínica mediante NLP y modelos de lenguaje*

| N.º | Referencia                    | Aplicación / tarea                                                                             | Hallazgo principal                                                             |
|-----|-------------------------------|------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| 17  | Balasubramanian et al. (2025) | Extracción estructurada de informes de patología con LLM (modelos de lenguaje de gran escala). | Los LLM extraen datos estructurados de reportes narrativos con alta fidelidad. |
| 18  | Siepmann et al. (2025)        | Extracción de información de cartas de alta con GPT-4                                          | Automatiza la lectura de resúmenes de egreso no estructurados.                 |
| 19  | Garcia-Carmona et al. (2025)  | Recuperación de información de pacientes con LLM                                               | Evalúa sistemáticamente la precisión de la recuperación en informes médicos.   |

| N.º | Referencia                      | Aplicación / tarea                                                           | Hallazgo principal                                                    |
|-----|---------------------------------|------------------------------------------------------------------------------|-----------------------------------------------------------------------|
| 20  | Vithanage et al. (2025)         | Extracción en registros de cuidado de adultos mayores                        | Evaluación comparativa de varios LLM sobre EHR no estructuradas.      |
| 21  | Bazoge et al. (2025)            | Clasificación de insuficiencia cardíaca en notas francesas                   | Los LLM clasifican y extraen información de notas clínicas.           |
| 22  | Zeinali et al. (2024)           | Detección de síntomas de cáncer (Symptom-BERT)                               | Mejora la detección de síntomas en notas clínicas de EHR.             |
| 23  | Guan et al. (2025)              | Detección temprana de deterioro cognitivo (CD-Tron)                          | Modelo clínico de gran escala aplicado sobre EHR.                     |
| 24  | Mata et al. (2025)              | Detección de síntomas de insuficiencia cardíaca en español                   | Optimización multicriterio de modelos sobre EHR en español.           |
| 25  | Najafabadipour et al. (2020)    | Reconstrucción de la historia natural del paciente                           | Integra datos dispersos para trazar la trayectoria clínica.           |
| 26  | Pinheiro da Silva et al. (2023) | Reconocimiento de entidades y extracción de relaciones                       | Integra ontologías con registros médicos.                             |
| 27  | Hurtado et al. (2025)           | Fenotipado profundo en radiología                                            | Combina transformadores con datos enlazados.                          |
| 28  | López-Úbeda et al. (2026)       | Normalización a SNOMED CT (nomenclatura clínica estandarizada internacional) | Modelos pequeños con razonamiento en cadena para normalizar términos. |

*Fuente: elaboración propia a partir de las referencias citadas.*

Como muestra la Tabla 4, la tendencia más reciente combina modelos de lenguaje de gran tamaño con tareas clínicas específicas. Estos trabajos coinciden en que los modelos de lenguaje ya no se limitan a clasificar texto, sino que permiten extraer variables estructuradas, normalizar terminología y reconstruir la trayectoria del paciente a partir de documentos dispersos, lo que amplía de manera notable el aprovechamiento de la información no estructurada y su enfoque en brindar más contexto en cada situación.

Una parte importante que conviene mencionar es la diversidad de tareas que agrupan estos trabajos. Algunos se concentran en la extracción de variables concretas o comunes como los informes de patología (Balasubramanian et al., 2025), las cartas de alta (Siepmann et al., 2025) o la recuperación de información de pacientes a partir de informes médicos (García-Carmona et al., 2025); otros en la detección de condiciones específicas, como los síntomas de cáncer (Zeinali et al., 2024), el deterioro cognitivo (Guan et al., 2025) o la insuficiencia cardíaca (Bazoge et al.,

2025; Mata et al., 2025); y otros en tareas más estructurales, como el reconocimiento de entidades y relaciones (Pinheiro da Silva et al., 2023) o la normalización terminológica hacia vocabularios estándar (López-Úbeda et al., 2026). Así, el NLP clínico puede considerarse una herramienta aplicable a problemas reales y muy distintos. No obstante, cada aplicación exige adaptar y validar los modelos al contexto en el que se usarán y cuál va a ser su objetivo o su enfoque, pues un sistema que funciona bien con un tipo de nota o especialidad no necesariamente rinde igual en otra.

### El reto del español clínico

Buena parte del desarrollo del NLP clínico se ha concentrado en estudios en idioma inglés, por lo que el procesamiento del texto médico en español constituye un reto específico y una línea de investigación en expansión. La Tabla 5 reúne los trabajos revisados que abordan directamente el español clínico.

**Tabla 5**

*Procesamiento del lenguaje natural clínico en español con enfoque experimental*

| N.º | Referencia                  | Enfoque                                                                         | Aporte                                                              |
|-----|-----------------------------|---------------------------------------------------------------------------------|---------------------------------------------------------------------|
| 29  | Báez et al. (2022)          | NLP de texto clínico en español (listas de espera, Chile)                       | Demuestra la viabilidad del NLP clínico en español en un caso real. |
| 30  | García Subies et al. (2024) | Modelos codificadores clínicos en español (NER y clasificación)                 | Análisis comparativo de modelos disponibles para el español.        |
| 31  | Moreno-Barea et al. (2025)  | NER (reconocimiento de entidades nombradas) para des identificar EHR en español | Reconocimiento de entidades orientado a proteger datos personales.  |
| 32  | Zotova et al. (2026)        | Modelos generativos para enlace de entidades clínicas en español                | Vincula entidades clínicas mediante modelos generativos.            |

*Fuente: elaboración propia a partir de las referencias citadas.*

Estos trabajos comparten un mensaje de fondo: el español clínico ya no depende exclusivamente de herramientas pensadas para el inglés. La experiencia chilena con listas de espera (Báez et al., 2022) muestra que es posible aplicar NLP a texto clínico real en español y obtener información útil para la gestión sanitaria. Los análisis comparativos de modelos en gran medida en español (García Subies et al., 2024) ofrecen una guía sobre qué herramientas funcionan mejor para tareas específicas como reconocimiento de las entidades, y los enfoques generativos para el enlace de entidades (Zotova et al., 2026) apuntan hacia la siguiente generación. Parte de este trabajo se orienta directamente a la desidentificación y al manejo adecuado de la información, de modo que pueda utilizarse sin comprometer principios éticos (Moreno-Barea et al., 2025), pues conecta el reto lingüístico con el de la privacidad y la ética virtual. Para un país como Colombia, donde la totalidad de las historias clínicas se redacta en español, disponer de estos recursos deja de ser un asunto académico o investigativo y se convierte en una condición práctica para aprovechar la información no estructurada sin dejar de lado su origen.

### **Desidentificación y protección de la privacidad**

La información sensible es otro componente central de esta revisión. Ya que las notas clínicas contienen datos personales y confidenciales, su uso en investigación exige procesos de anonimización o desidentificación robustos que permitan claridad y transparencia al momento de su estudio ya que no solo se está manejando un simple dato si no que se debería tomar como fuente de la vida y privacidad de un individuo. Sin embargo, la desidentificación no debe entenderse únicamente como un requisito legal: es una condición habilitante que permite el uso secundario de los datos con fines de investigación y gestión clínica.

La Tabla 6 sintetiza los enfoques revisados, que van desde revisiones sistemáticas de métodos hasta marcos concretos basados en modelos de lenguaje.

**Tabla 6**

*Enfoques para la desidentificación y de la privacidad en la salud clínica*

| N.º | Referencia             | Enfoque                                                                     | Aporte                                                            |
|-----|------------------------|-----------------------------------------------------------------------------|-------------------------------------------------------------------|
| 33  | Faustini et al. (2026) | Revisión sistemática de desidentificación (texto, imagen y datos tabulares) | Ofrece un panorama comparado de los métodos de desidentificación. |
| 34  | Negash et al. (2023)   | Revisión de alcance de desidentificación de texto libre                     | Sintetiza enfoques sobre información personal de salud.           |
| 35  | Dorémus et al. (2025)  | LLM de tamaño moderado para des identificar registros de urgencias          | Desidentificación fiable, validada e implementada.                |
| 36  | Jia y Nishi (2025)     | Marco de anonimización en dos etapas para registros narrativos              | Flexible y adaptable a distintos modelos de lenguaje.             |
| 37  | Zhong et al. (2025)    | Revisión de alcance sobre privacidad y LLM en salud                         | Reúne consideraciones de privacidad al usar LLM.                  |

*Fuente: elaboración propia a partir de las referencias citadas.*

Estos estudios dibujan una evolución en la protección de datos personales en el sector médico.

Las revisiones sistemáticas (Faustini et al., 2026; Negash et al., 2023) muestran que durante años la desidentificación se apoyó en reglas y diccionarios mas no en una fuerte aplicación de formas y maneras claras de evitar su filtración, con resultados desiguales según el tipo de dato. Los enfoques más recientes muestran que se está priorizando recurrir a modelos de lenguaje para reconocer y ocultar la información identificable de forma más robusta y directa, incluso en escenarios exigentes como los registros de urgencias médicas (Dorémus et al., 2025), y proponen marcos flexibles que se adaptan a distintos modelos según los recursos disponibles (Jia y Nishi, 2025). La literatura también advierte que la misma tecnología basada en modelos de lenguaje que ayuda a proteger los datos puede, si se emplea sin cuidado, generar nuevos riesgos de filtración (Zhong et al., 2025). Por ello, la desidentificación no debe verse como un paso único, sino como un proceso que requiere validación y constante revisión durante su uso y manejo.

### Interoperabilidad y construcción de registros clínicos

La interoperabilidad y la construcción de registros a partir de datos no estructurados conforman otro componente relevante de esta revisión. Para que la información extraída sea reutilizable, debe ajustarse a modelos y estándares comunes. La Tabla 7 resume los trabajos revisados sobre este tema.

**Tabla 7**

*Interoperabilidad y construcción de registros a partir de datos no estructurados*

| N.º | Referencia                    | Tema                                                                                                                           | Aporte                                                            |
|-----|-------------------------------|--------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------|
| 38  | González-Castro et al. (2021) | Modelo de datos interoperable (CASIDE, basado en FHIR)                                                                         | Estructura información de supervivencia oncológica interoperable. |
| 39  | Tabari et al. (2024)          | Implementaciones de modelos de datos basados en FHIR                                                                           | Revisión del estado del arte de los modelos FHIR.                 |
| 40  | Liu et al. (2020)             | Sistema de recuperación de cohortes basado en OMOP CDM ( <i>Observational Medical Outcomes Partnership Common Data Model</i> ) | Recupera cohortes desde repositorios de datos clínicos.           |
| 41  | Tavabi et al. (2024)          | Construcción de registros desde notas no estructuradas                                                                         | Pipeline de NLP de bajos recursos para grandes registros.         |
| 42  | Taseh et al. (2025)           | NLP frente a códigos CIE (Clasificación Internacional de Enfermedades) para construir registros                                | Compara el NLP con la codificación diagnóstica tradicional.       |

*Fuente: elaboración propia a partir de las referencias citadas.*

La lectura conjunta de estos trabajos indica que extraer información del texto clínico es solo una parte del problema; la otra es lograr que esa información pueda compartirse y reutilizarse. Los modelos basados en el estándar FHIR (*Fast Healthcare Interoperability Resources*, estándar de interoperabilidad en salud) (González-Castro et al., 2021; Tabari et al., 2024) buscan que los datos tengan una estructura común, de modo que distintos sistemas puedan entenderse entre sí. En paralelo, los modelos comunes de datos, como el empleado en los sistemas de recuperación de cohortes (Liu et al., 2020), permiten formular preguntas de investigación sobre grandes repositorios clínicos. Los estudios que construyen registros directamente desde notas no

estructuradas (Tavabi et al., 2024) o que comparan el NLP con la codificación diagnóstica tradicional (Taseh et al., 2025) muestran que el texto libre, procesado adecuadamente, puede alimentar registros de calidad comparable o incluso superior a la de los basados únicamente en códigos. La interoperabilidad conecta el dato extraído con su uso en investigación y gestión.

### Marco normativo colombiano

Finalmente, el aprovechamiento de la información clínica en Colombia se enmarca en un conjunto de normas y guías que regulan tanto el manejo de la historia clínica como la protección de los datos personales. La Tabla 8 sintetiza este marco.

**Tabla 8**

*Marco normativo y de gestión de datos aplicable en Colombia*

| N.º | Norma / documento                 | Referencia                                            | Contenido                                                  |
|-----|-----------------------------------|-------------------------------------------------------|------------------------------------------------------------|
| 43  | Ley 23 de 1981                    | Congreso de Colombia (1981)                           | Normas en materia de ética médica.                         |
| 44  | Resolución 1995 de 1999           | Ministerio de Salud (1999)                            | Normas para el manejo de la historia clínica.              |
| 45  | Ley 1581 de 2012                  | Congreso de Colombia (2012)                           | Protección de datos personales.                            |
| 46  | Ley 2015 de 2020                  | Congreso de Colombia (2020)                           | Historia clínica electrónica interoperable.                |
| 47  | Resolución 866 de 2021            | Ministerio de Salud y Protección Social (2021)        | Datos clínicos relevantes para la interoperabilidad.       |
| 48  | Resolución 1888 de 2025           | Ministerio de Salud y Protección Social (2025)        | Adopción del Resumen Digital de Atención en Salud (RDA) .  |
| 49  | Guía de anonimización de datos    | DANE (2024)                                           | Lineamientos para la anonimización de datos estructurados. |
| 50  | Gestión de datos de investigación | Ministerio de Ciencia, Tecnología e Innovación (2022) | Definiciones y conceptos básicos sobre datos.              |

*Fuente: elaboración propia a partir de la normativa citada.*

Este marco normativo evidencia una progresión coherente en Colombia. Las primeras normas fijaron los principios de la ética médica (Congreso de Colombia, 1981) y las reglas para el manejo de la historia clínica (Ministerio de Salud, 1999); posteriormente, la legislación de protección de datos personales (Congreso de Colombia, 2012) estableció garantías generales sobre el tratamiento de la información sensible. Más recientemente, el país avanzó hacia la historia clínica electrónica interoperable (Congreso de Colombia, 2020) y hacia su reglamentación técnica (Ministerio de Salud y Protección Social, 2021), un proceso que continúa con la adopción del Resumen Digital de Atención en Salud (RDA) (Ministerio de Salud y Protección Social, 2025). A este marco también se le suman las guías técnicas sobre anonimización de datos publicadas por el DANE (2024) y los lineamientos de gestión de datos de investigación del Ministerio de Ciencia, Tecnología e Innovación (2022). Con esto se puede ver que usar la información clínica no es solo cuestión de cumplir con requisitos técnicos, sino que también implica un marco jurídico y ético que el país ha ido construyendo poco a poco.

### **Discusión**

Los resultados obtenidos permiten interpretar el uso de la información clínica no estructurada como un campo en el que tres componentes, el valor del texto libre, la madurez creciente de los modelos de lenguaje y la exigencia irrenunciable de proteger la privacidad, no operan de forma aislada, sino que se refuerzan y condicionan entre sí, por lo que conviene leerlos de manera conjunta en lugar de retomar cada hallazgo por separado.

La evolución técnica que se muestra en las tablas 3 y 4 solo tiene sentido si se entiende junto con ese primer hallazgo: si el texto libre no fuera realmente un activo informativo relevante, como lo confirman las revisiones sobre síntomas documentados en notas clínicas y sobre la complementariedad entre código estructurado y texto libre (Koleck et al., 2019; Li et al., 2022; Seinen et al., 2025), entonces no tendría mucho sentido seguir desarrollando modelos cada vez más sofisticados para procesarlo. Por eso, el paso de arquitecturas especializadas basadas en BERT (Devlin et al., 2019; Lee et al., 2020; Huang et al., 2019) hacia los grandes modelos de lenguaje aplicados a la extracción de información (Balasubramanian et al., 2025; Siepmann et al., 2025; Vithanage et al., 2025) no es simplemente un avance tecnológico que ocurre por sí solo, sino una respuesta a la necesidad de aprovechar mejor ese texto libre. Esto coincide con el cambio que ha tenido el NLP clínico, que pasó de enfocarse en tareas aisladas a aplicarse a lo largo de toda la atención en salud (Klug et al., 2024), y con su llegada a dominios más especializados (Woo et al., 2026; Daskalo et al., 2026). Aun así, esta madurez técnica no es pareja en todos los idiomas: sigue estando concentrada mayormente en inglés, lo cual se convierte en la primera limitación que atraviesa todo este campo.

Esa limitación se relaciona con dos problemas que en la literatura casi siempre se tratan por separado, aunque en la práctica van de la mano: la adaptación al español y la protección de los datos. Por un lado, ya existen recursos y modelos para el español clínico (Báez et al., 2022; García Subies et al., 2024; Mata et al., 2025; Zotova et al., 2026), pero siguen siendo pocos comparados con todo lo que hay en inglés, lo que confirma que hay una brecha de idioma que afecta directamente el aprovechamiento de la información clínica en países hispanohablantes como Colombia. Por otro lado, la desidentificación no debería verse como un paso final después de extraer la información, sino como algo que hay que resolver antes de poder reutilizarla

(Faustini et al., 2026; Negash et al., 2023). Los trabajos más recientes, que usan modelos de lenguaje para esto, ya muestran resultados confiables (Dorémus et al., 2025; Jia & Nishi, 2025), incluso en español (Moreno-Barea et al., 2025). Sin embargo, usar modelos de lenguaje para desidentificar también trae un riesgo que no se puede ignorar: la misma tecnología que protege los datos, si se usa mal, puede terminar filtrándolos (Zhong et al., 2025). Y no es casualidad que muchos de los recursos que están ayudando a cerrar la brecha del español sean, al mismo tiempo, herramientas de protección de datos; esto da a entender que estos dos problemas se podrían trabajar juntos en vez de uno después del otro. A todo esto se suma el tema de la interoperabilidad: estándares como FHIR y los modelos comunes de datos (González-Castro et al., 2021; Tabari et al., 2024; Liu et al., 2020) son los que finalmente definen si esa información, ya extraída, traducida al español y sin identificadores, se puede realmente usar a mayor escala. Si no se resuelve esta parte, todo el avance técnico y lingüístico se queda en estudios sueltos que no llegan a tener un impacto real en la práctica clínica.

Por último, vale la pena pensar todo esto en el caso de un país como Colombia. Ya existe un marco normativo que contempla la historia clínica electrónica interoperable, cada vez hay más modelos de NLP disponibles para el español, y las técnicas de desidentificación han madurado bastante, así que las condiciones están dadas para que hospitales y grupos de investigación empiecen a usar la información no estructurada de forma responsable. El problema principal ya no es tanto si la tecnología existe, porque en gran parte ya existe, sino cómo adaptarla al contexto local, validarla con datos propios y armar flujos de trabajo que tengan la protección de datos incorporada desde el inicio y no como un parche después. Por eso, más que un problema de conocimiento, lo que hay entre el potencial de estas herramientas y su uso real en la práctica clínica e investigativa es, sobre todo, un problema de implementación.

## **Conclusiones**

A continuación se presentan los aportes del trabajo en relación con los tres objetivos específicos que se plantearon al inicio: mostrar la importancia de los datos no estructurados en el sector salud, describir las principales aplicaciones del NLP en este campo y, por último, analizar los desafíos más relevantes. Para cada uno de estos se explican los resultados principales, si el objetivo se cumplió o no, las limitaciones que tuvo la revisión y, finalmente, qué se podría hacer en el futuro.

En cuanto al primer objetivo, mostrar la importancia de los datos no estructurados en el sector salud, este se considera cumplido, puesto que los resultados evidencian que buena parte de la información clínica más valiosa se encuentra en el texto libre y no en los campos estructurados. Además, ese texto no reemplaza a los datos codificados, sino que más bien los complementa, ya que aporta contexto, matices y detalles sobre cómo evoluciona el paciente que de otra forma se perderían. Así lo muestran, por ejemplo, varios estudios sobre síntomas registrados en notas clínicas y sobre cómo se complementan el código estructurado y el texto libre (Tabla 2), lo cual confirma que este hallazgo es, en últimas, el punto de partida que justifica por qué se siguen desarrollando modelos de lenguaje cada vez más avanzados. Ahora bien, como limitación, hay que tener en cuenta que esta es una revisión de literatura descriptiva y cualitativa, y no un estudio experimental, así que esta conclusión no se contrastó con datos clínicos reales, y aunque las cincuenta fuentes elegidas son representativas, tampoco cubren toda la producción científica del área. Por eso, como trabajo futuro, sería útil evaluar pipelines de NLP sobre corpus clínicos en español, idealmente en el contexto colombiano, para así poder comprobar de forma empírica el valor del texto libre que se identificó en esta revisión.

Con respecto al segundo objetivo, describir las principales aplicaciones del NLP y los modelos de lenguaje en este campo, este también se considera cumplido, ya que los resultados dejan ver una evolución técnica bastante clara: se pasó de modelos especializados basados en BERT hacia los grandes modelos de lenguaje aplicados a la extracción de información, abarcando tareas que van desde extraer variables concretas hasta detectar condiciones específicas y reconocer entidades. Y, de hecho, ya hay avances propios para el español clínico. Todo esto se apoya en las Tablas 3, 4 y 5, donde se organizan, respectivamente, los modelos de referencia divididos y bien referenciados, doce aplicaciones concretas y los recursos disponibles para el español. Sin embargo, como limitación, el desarrollo de estas herramientas se ha concentrado sobre todo en inglés, lo que puede sesgar lo que se sabe de su desempeño real en español y, por lo tanto, sigue siendo una brecha que limita su uso en contextos hispanohablantes como el colombiano. En ese sentido, como trabajo futuro, se recomienda desarrollar y evaluar estos pipelines directamente sobre corpus clínicos en español, además de comparar modelos pequeños y grandes en tareas de extracción para saber qué arquitecturas funcionan mejor en este idioma.

Finalmente, el tercer objetivo, analizar los desafíos más relevantes del aprovechamiento de estos datos es decir, la precisión de los modelos, la privacidad de los pacientes y la desidentificación de información sensible, también se cumplió. Los resultados muestran que la desidentificación no debería verse solamente como un trámite legal que se hace al final, sino más bien como algo necesario para poder usar los datos de forma secundaria. Y en esa misma línea, la interoperabilidad, a través de estándares como FHIR y modelos comunes de datos, es lo que termina definiendo si esa información, ya extraída y protegida, se puede reutilizar a mayor escala en investigación y gestión clínica. Esto se apoya en las Tablas 6, 7 y 8, que organizan los enfoques de desidentificación, los mecanismos de interoperabilidad y el marco normativo

colombiano, el cual, cabe decir, avanza poco a poco hacia la historia clínica electrónica interoperable. Aun así, como limitación, al no ser un estudio experimental, la revisión no evaluó qué tan precisos son estos modelos ni las estrategias de desidentificación sobre datos clínicos reales, sino lo que reporta la literatura. Por eso, como trabajo futuro, se recomienda avanzar hacia flujos interoperables basados en FHIR que incorporen la protección de datos desde el diseño, de modo que la desidentificación y la interoperabilidad dejen de manejarse como pasos separados y así se pueda investigar con registros médicos de forma responsable.

### Referencias

- Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (pp. 72–78). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-1909>
- Balasubramanian, J. B., Adams, D., Roxanis, I., Berrington de Gonzalez, A., Coulson, P., Almeida, J. S., & García-Closas, M. (2025). Leveraging large language models for structured information extraction from pathology reports. *Journal of Pathology Informatics*, 19, 100521. <https://doi.org/10.1016/j.jpi.2025.100521>
- Báez, P., Arancibia, A. P., Chaparro, M. I., Bucarey, T., Núñez, F., & Dunstan, J. (2022). Procesamiento de lenguaje natural para texto clínico en español: el caso de las listas de espera en Chile. *Revista Médica Clínica Las Condes*, 33(6), 576–582. <https://doi.org/10.1016/j.rmclc.2022.10.002>
- Bazoge, A., Wargny, M., Constant dit Beaufile, P., Morin, E., Daille, B., Gourraud, P.-A., & Hadjadj, S. (2025). Assessing large language models for acute heart failure classification and information extraction from French clinical notes. *Computers in Biology and Medicine*, 195, 110609. <https://doi.org/10.1016/j.combiomed.2025.110609>
- Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2004.05150>
- Congreso de Colombia. (1981, 18 de febrero). Ley 23 de 1981. Por la cual se dictan normas en materia de ética médica. <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=68760>

Congreso de Colombia. (2012, 17 de octubre). Ley 1581 de 2012. Por la cual se dictan

disposiciones generales para la protección de datos personales.

<https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981>

Congreso de Colombia. (2020, 31 de enero). Ley 2015 de 2020. Por medio de la cual se crea la

historia clínica electrónica interoperable y se dictan otras disposiciones.

[https://www.minsalud.gov.co/Normatividad\\_Nuevo/Ley%202015%202020.pdf](https://www.minsalud.gov.co/Normatividad_Nuevo/Ley%202015%202020.pdf)

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare:

Management, analysis and future prospects. *Journal of Big Data*, 6, Article 54.

<https://doi.org/10.1186/s40537-019-0217-0>

Daskalo, C., Abu-Ashour, W., Tshimula, J. M., Amoei, M., Guadagno, E., & Poenaru, D. (2026).

Large language models for electronic health records in pediatric and surgical care: A systematic review. *Journal of Pediatric Surgery*, 162956.

<https://doi.org/10.1016/j.jpedsurg.2026.162956>

Departamento Administrativo Nacional de Estadística [DANE]. (2024). Guía para la

anonimización de datos estructurados. <https://www.dane.gov.co/files/sen/registros-administrativos/guia-anonimizacion-datos2024.pdf>

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep

bidirectional transformers for language understanding. In *Proceedings of the 2019*

*Conference of the North American Chapter of the Association for Computational*

*Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp.

4171–4186). Association for Computational Linguistics.

<https://doi.org/10.18653/v1/N19-1423>

- Dorémus, O., Russon, D., Contrand, B., Guerra-Adames, A., Avalos-Fernandez, M., Gil-Jardiné, C., & Lagarde, E. (2025). Harnessing moderate-sized language models for reliable patient data deidentification in emergency department records: Algorithm development, validation, and implementation study. *JMIR AI*, 4. <https://doi.org/10.2196/57828>
- Faustini, P., McIver, A., Sullivan, R., & Dras, M. (2026). De-identification of clinical data: A systematic review of free text, image and tabular data approaches. *International Journal of Medical Informatics*, 208, 106225. <https://doi.org/10.1016/j.ijmedinf.2025.106225>
- Garcia-Carmona, A. M., Prieto, M.-L., Puertas, E., & Beunza, J.-J. (2025). Leveraging large language models for accurate retrieval of patient information from medical reports: Systematic evaluation study. *JMIR AI*, 4. <https://doi.org/10.2196/68776>
- García Subies, G., Barbero Jiménez, Á., & Martínez Fernández, P. (2024). A comparative analysis of Spanish clinical encoder-based models on NER and classification tasks. *Journal of the American Medical Informatics Association*, 31(9), 2137–2146. <https://doi.org/10.1093/jamia/ocae054>
- González-Castro, L., Cal-González, V. M., Del Fiol, G., & López-Nores, M. (2021). CASIDE: A data model for interoperable cancer survivorship information based on FHIR. *Journal of Biomedical Informatics*, 124, 103953. <https://doi.org/10.1016/j.jbi.2021.103953>
- Guan, H., Novoa-Laurentiev, J., & Zhou, L. (2025). CD-Tron: Leveraging large clinical language model for early detection of cognitive decline from electronic health records. *Journal of Biomedical Informatics*, 166, 104830. <https://doi.org/10.1016/j.jbi.2025.104830>
- Hossain, E., Rana, R. K., Higgins, N. S., Soar, J., Barua, P. D., Pisani, A. R., & Turner, K. (2023). Natural language processing in electronic health records in relation to healthcare

decision-making: A systematic review. *Computers in Biology and Medicine*, 155,

106649. <https://doi.org/10.1016/j.combiomed.2023.106649>

Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission [Preprint]. arXiv.

<https://doi.org/10.48550/arXiv.1904.05342>

Hurtado, L.-F., Marco-Ruiz, L., Segarra, E., Castro-Bleda, M. J., Bustos-Moreno, A., de la

Iglesia-Vayá, M., & Vallalta-Rueda, J. F. (2025). Leveraging transformers-based models

and linked data for deep phenotyping in radiology. *Computer Methods and Programs in*

*Biomedicine*, 260, 108567. <https://doi.org/10.1016/j.cmpb.2024.108567>

Jerfy, A., Selden, O., & Balkrishnan, R. (2024). The growing impact of natural language

processing in healthcare and public health. *Inquiry: The Journal of Health Care*

*Organization, Provision, and Financing*, 61, 469580241290095.

<https://doi.org/10.1177/00469580241290095>

Jia, J., & Nishi, H. (2025). A flexible two-stage anonymization framework for narrative medical

records adapting to various language models. *Computers in Biology and Medicine*, 195,

110624. <https://doi.org/10.1016/j.combiomed.2025.110624>

Kim, M. K., Rouphael, C., McMichael, J., Welch, N., & Dasarathy, S. (2024). Challenges in and

opportunities for electronic health record-based data analysis and interpretation. *Gut and*

*Liver*, 18(2), 201–208. <https://doi.org/10.5009/gnl230272>

Klug, K., Beckh, K., Antweiler, D., Chakraborty, N., Baldini, G., Laue, K., Hosch, R., Nensa, F.,

Schuler, M., & Giesselbach, S. (2024). From admission to discharge: A systematic

review of clinical natural language processing along the patient journey. *BMC Medical*

Informatics and Decision Making, 24, Article 238. <https://doi.org/10.1186/s12911-024-02641-w>

Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. *Journal of the American Medical Informatics Association*, 26(4), 364–379. <https://doi.org/10.1093/jamia/ocy173>

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M., & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, 73, 14–29. <https://doi.org/10.1016/j.jbi.2017.07.012>

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>

Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W. P., Nuzumlalı, M. Y., Rosand, B., Li, Y., Zhang, M., Chang, D., Taylor, R. A., Krumholz, H. M., & Radev, D. (2022). Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46, 100511. <https://doi.org/10.1016/j.cosrev.2022.100511>

Liu, S., Wang, Y., Wen, A., Wang, L., Hong, N., Shen, F., Bedrick, S., Hersh, W., & Liu, H. (2020). Implementation of a cohort retrieval system for clinical data repositories using the Observational Medical Outcomes Partnership Common Data Model: Proof-of-concept system validation. *JMIR Medical Informatics*, 8(10). <https://doi.org/10.2196/17376>

López-Úbeda, P., Martín-Noguerol, T., & Luna, A. (2026). Integrating semantic retrieval and chain-of-thought reasoning in small language models for SNOMED CT normalization.

International Journal of Medical Informatics, 211, 106340.

<https://doi.org/10.1016/j.ijmedinf.2026.106340>

Mata, J., Pachón, V., Manovel, A., Maña, M. J., & de la Villa, M. (2025). Multicriteria optimization of language models for heart failure with preserved ejection fraction symptom detection in Spanish electronic health records: Comparative modeling study.

Journal of Medical Internet Research, 27. <https://doi.org/10.2196/76433>

Ministerio de Ciencia, Tecnología e Innovación. (2022, 9 de julio). Definiciones y conceptos básicos. Gestión de datos de investigación. [https://red-](https://red-documentacion.minciencias.gov.co/Gestion_Datos_Investigacion/gestion-datos)

[documentacion.minciencias.gov.co/Gestion\\_Datos\\_Investigacion/gestion-datos](https://red-documentacion.minciencias.gov.co/Gestion_Datos_Investigacion/gestion-datos)

Ministerio de Salud. (1999, 8 de julio). Resolución 1995 de 1999. Por la cual se establecen normas para el manejo de la historia clínica.

[https://www.minsalud.gov.co/normatividad\\_nuevo/resoluci%C3%93n%201995%20de%201999.pdf](https://www.minsalud.gov.co/normatividad_nuevo/resoluci%C3%93n%201995%20de%201999.pdf)

Ministerio de Salud y Protección Social. (2021, 25 de junio). Resolución 866 de 2021. Por la cual se reglamenta el conjunto de elementos de datos clínicos relevantes para la interoperabilidad de la historia clínica en el país y se dictan otras disposiciones.

<https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/resolucion-866-de-2021.pdf>

Ministerio de Salud y Protección Social. (2025, 15 de septiembre). Resolución 1888 de 2025. Por medio de la cual se adopta el Resumen Digital de Atención en Salud (RDA) en el marco de la Interoperabilidad de la Historia Clínica Electrónica (IHCE), se establece el

mecanismo para su implementación a nivel nacional y se dictan otras disposiciones.

[https://www.minsalud.gov.co/Normatividad\\_Nuevo/Resolucion%20No%201888%20de%202025.pdf](https://www.minsalud.gov.co/Normatividad_Nuevo/Resolucion%20No%201888%20de%202025.pdf)

Moreno-Barea, F. J., López-García, G., Mesa, H., Ribelles, N., Alba, E., Jerez, J. M., & Veredas,

F. J. (2025). Named entity recognition for de-identifying Spanish electronic health records. *Computers in Biology and Medicine*, 185, 109576.

<https://doi.org/10.1016/j.combiomed.2024.109576>

Najafabadipour, M., Zanin, M., Rodríguez-González, A., Torrente, M., Nuñez García, B., Cruz

Bermudez, J. L., Provencio, M., & Menasalvas, E. (2020). Reconstructing the patient's natural history from electronic health records. *Artificial Intelligence in Medicine*, 105, 101860. <https://doi.org/10.1016/j.artmed.2020.101860>

Negash, B., Katz, A., Neilson, C. J., Moni, M., Nesca, M., Singer, A., & Enns, J. E. (2023). De-identification of free text data containing personal health information: A scoping review of reviews. *International Journal of Population Data Science*, 8(1), Article 2153.

<https://doi.org/10.23889/ijpds.v8i1.2153>

Pinheiro da Silva, D., da Rosa Fröhlich, W., de Mello, B. H., Vieira, R., & Rigo, S. J. (2023).

Exploring named entity recognition and relation extraction for ontology and medical records integration. *Informatics in Medicine Unlocked*, 43, 101381.

<https://doi.org/10.1016/j.imu.2023.101381>

Seinen, T. M., Kors, J. A., van Mulligen, E. M., & Rijnbeek, P. R. (2025). Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study. *Journal of Medical Internet Research*, 27, e66910. <https://doi.org/10.2196/66910>

- Siepmann, R. M., Baldini, G., Schmidt, C. S., Truhn, D., Müller-Franzes, G. A., Dada, A., Kleesiek, J., Nensa, F., & Hosch, R. (2025). An automated information extraction model for unstructured discharge letters using large language models and GPT-4. *Healthcare Analytics*, 7, 100378. <https://doi.org/10.1016/j.health.2024.100378>
- Tabari, P., Costagliola, G., De Rosa, M., & Boeker, M. (2024). State-of-the-art Fast Healthcare Interoperability Resources (FHIR)–based data model and structure implementations: Systematic scoping review. *JMIR Medical Informatics*, 12. <https://doi.org/10.2196/58445>
- Taseh, A., Moradian, A. D., Chan, M., Sirls, E., Nazarian, A., Batmanghelich, K., & Bean, J. F. (2025). Performance of natural language processing versus International Classification of Diseases codes in building registries for patients with fall injury: Retrospective analysis. *JMIR Medical Informatics*, 13. <https://doi.org/10.2196/66973>
- Tavabi, N., Pruneski, J., Golchin, S., Singh, M., Sanborn, R., Heyworth, B., Landschaft, A., Kimia, A., & Kiapour, A. (2024). Building large-scale registries from unstructured clinical notes using a low-resource natural language processing pipeline. *Artificial Intelligence in Medicine*, 151, 102847. <https://doi.org/10.1016/j.artmed.2024.102847>
- Vithanage, D., Yu, P., Xie, Q., Xu, H., Wang, L., & Deng, C. (2025). A comprehensive evaluation of large language models for information extraction from unstructured electronic health records in residential aged care. *Computers in Biology and Medicine*, 197, 111013. <https://doi.org/10.1016/j.combiomed.2025.111013>
- Woo, B. F. Y., Cato, K., Cho, H., You, S. B., & Song, J. (2026). The use of large language models in clinical documentation: A scoping review. *International Journal of Nursing Studies*, 176, 105322. <https://doi.org/10.1016/j.ijnurstu.2025.105322>

Zeinali, N., Albashayreh, A., Fan, W., & Gilbertson White, S. (2024). Symptom-BERT:

Enhancing cancer symptom detection in EHR clinical notes. *Journal of Pain and Symptom Management*, 68(2), 190–198.e1.

<https://doi.org/10.1016/j.jpainsymman.2024.05.015>

Zhong, X., Li, S., Chen, Z., Ge, L., Yu, D., Wang, S., You, L., & Shang, H. (2025).

Considerations for patient privacy of large language models in health care: Scoping review. *Journal of Medical Internet Research*, 27. <https://doi.org/10.2196/76571>

Zotova, E., Cuadros, M., & Rigau, G. (2026). Generative models for clinical entity linking in Spanish. *Array*, 30, 100805. <https://doi.org/10.1016/j.array.2026.100805>