PANDORA AI

Un modelo de inteligencia artificial para la extracción automática de datos clínicos no estructurados e implementación de puntajes de riesgo clínico

AUTORES
Daniel Jimenez, Julian Martinez, Natalia Castano-Villegas, Isabella Llano, Jose Zea

AFILIACIONES
Arkangel AI

INTRODUCCIÓN

Una de las principales fuentes de sesgo en la investigación observacional es la calidad de las fuentes de información secundaria, a menudo registros clínicos. Datos valiosos suelen estar ocultos en formatos de texto, en las notas de los médicos o perdidos en bases de datos de baja calidad (1). Sin embargo, los avances en tecnología y computación han permitido recopilar, organizar, operar, analizar e interpretar grandes cantidades de datos y utilizar inteligencia artificial para realizar tareas repetitivas y que consumen mucho tiempo, impulsando soluciones innovadoras en salud. Los modelos de lenguaje (LLMs) pueden proporcionar una alternativa para la extracción de información clínica (2).

OBJETIVO

Queríamos desarrollar un modelo de inteligencia artificial que no solo pudiera recuperar con precisión cualquier tipo de dato a partir de texto y hacerlo accesible para médicos e investigadores, sino que también implementara automáticamente cualquier guía clínica para propósitos de diagnóstico, utilizando la información extraída. Este estudio presenta la validación inicial de nuestro modelo de inteligencia artificial generativa, PANDORA.

METODOLOGÍA

Pandora tiene dos algoritmos: uno recupera información de los registros electrónicos de salud (EHRs) y el otro realiza recomendaciones diagnósticas basadas en la puntuación de la escala clínica seleccionada. En este caso, utilizamos la escala PUMA, validada en varios países de América Latina y China para la detección de EPOC. Estas dos fases se integran mediante las capacidades del modelo de lenguaje GPT-4o.

Se utilizaron dos fuentes de información clínica (EHRs): La base de datos MIMIC-IV-Note, un conjunto de notas de alta clínica de pacientes reales y anonimizados del Beth Israel Deaconess Medical Center en Boston, EE. UU. Y una guía diseñada por nuestro equipo médico para fabricar registros de consulta externa basados en la estructura de una historia clínica estándar en Colombia.

Las capacidades de extracción del primer algoritmo fueron evaluadas utilizando la base de datos EHR-DS-QA como referencia, un conjunto de preguntas creadas a partir de casos en MIMIC. Usamos métricas semánticas como BERTScore, Semantic Score y Relevance Score. La interacción entre los algoritmos se evaluó aplicando los criterios de la escala PUMA, que guiaron la extracción de características específicas. Basado en esto, PANDORA asignó una puntuación de 0 a 9, con un umbral de 5 para recomendar diagnósticos de EPOC.

Las recomendaciones del segundo algoritmo fueron evaluadas utilizando dos enfoques: La primera usó la base de datos MIMIC como fuente de EHRs, donde se evaluó manualmente la extracción, puntuación y recomendación. La segunda usó la base de datos con casos sintéticos, donde además de las tres capacidades anteriores, evaluamos la plausibilidad y estructura de los casos clínicos.

RESULTADOS

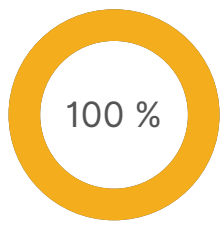
Todos los puntajes semánticos estuvieron por encima del 90% (Fig. 1), lo que indica una buena extracción en general, con una comprensión adecuada de preguntas, respuestas con significados coherentes y un contexto apropiado. Las recomendaciones de Pandora para el diagnóstico de EPOC utilizando la calculadora PUMA tuvieron una precisión del 86% en comparación con el estándar en la base de datos MIMIC-IV y del 100% en casos sintéticos evaluados por humanos. La capacidad del sistema para aplicar correctamente el puntaje PUMA fue del 98% para MIMIC-IV y del 95% en casos evaluados por humanos. Las capacidades de extracción del modelo, tanto según la evaluación humana como el estándar, fueron del 100% para los EHRs de MIMIC-IV y del 99% para los casos sintéticos (Figs. 2-3).

FIG 3. EVALUACIÓN HUMANA DE PANDORA USANDO MIMIC

Recomendación

Métrica	Valor
Sensibilidad	**0.885**
Especificidad	0.700
Precisión	0.815
Exactitud	**0.794**
Puntaje F1	0.835

Extracción
Proporción de positivos



Puntuación
Proporción de positivos

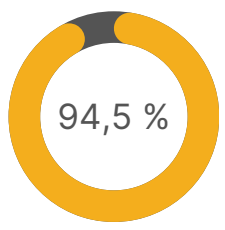


Fig 1. EVALUACIÓN AUTOMÁTICA DEL ALGORITMO DE EXTRACCIÓN

Métricas Semánticas

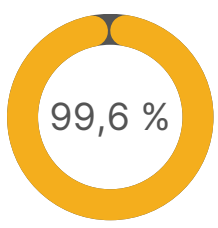
Métrica	Puntaje
BERTScore	0.911
SemanticScore	0.925
RelevanceScore	0.901

FIG 2. EVALUACIÓN HUMANA DE PANDORA CON CASOS SINTÉTICOS

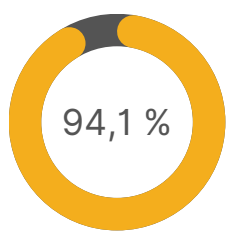
Recomendación

Metric	Value
Sensibilidad	**1.0**
Especificidad	0.200
Precisión	0.238
Exactitud	**0.360**
Puntaje F1	0.385

Extracción
Proporción de positivos



Puntuación
Proporción de positivos



DISCUSSION

En esta evaluación inicial, PANDORA demostró que puede extraer adecuadamente datos estructurados a partir de fuentes no estructuradas, como registros médicos y notas de alta, con buena concordancia cuando se compara con evaluaciones humanas. Además, exploramos su capacidad para interactuar con una calculadora de riesgo validada (capacidad de puntuación), la escala PUMA para la evaluación del riesgo de EPOC, lo que reveló que nuestro modelo comprende con precisión las condiciones necesarias para ciertas puntuaciones. En cuanto a la capacidad de recomendación, pudo identificar el riesgo de EPOC en todos los casos sintéticos y en el 89% de las notas de alta de MIMIC-IV.
Los valores bajos de especificidad y precisión se explican por la naturaleza de la escala clínica utilizada. La escala PUMA está validada para el tamizaje en población de mayor edad y con antecedentes de tabaquismo intenso (4-6), y su umbral de riesgo de EPOC se estableció en consecuencia. Investigaciones futuras que utilicen esta u otras escalas para EPOC podrían emplear un umbral más alto para mejor adaptación a la población general.

CONCLUSIONS

Esta evaluación inicial es el primer paso hacia la validación y el lanzamiento de una herramienta clínica y de investigación que permitirá la aplicación de puntuaciones diagnósticas de diversas enfermedades a información previamente atrapada en formatos que la hacían inaccesible. Incluso instituciones sin bases de datos estructuradas podrán utilizarla y aprovechar todo el conocimiento actualmente escrito en texto libre.

BIBILIOGRAPHY

(1) Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. J Med Ethics. 2020 Jun;46(6):367-377. doi: 10.1136/medethics-2019-105472. Epub 2020 May 26. PMID: 32457202; PMCID: PMC7279205. **(2)** Gu B, Shao V, Liao Z, Carducci V, Brufau SR, Yang J, et al. Scalable information extraction from free text electronic health records using large language models [Internet]. medRxiv; 2024 [cited 2024 Aug 29]. p. 2024.08.08.24311237. Available from: https://www.medrxiv.org/content/10.1101/2024.08.08.24311237v1 **(3)** Bastidas G. AR, Estupiñán B. MF, Arias B. JS, Estrada H. M, López O. J, Mateus M. SL, et al. Validación externa y reproducibilidad del cuestionario PUMA para el diagnóstico de EPOC en una población latinoamericana: Validación externa del cuestionario PUMA. Rev Chil Enfermedades Respir. 2022 Mar;38(1):11–9. **(4)** Herrera AC, Oca MM de, Varela MVL, Aguirre C, Schiavi E, Jardim JR, et al. COPD Underdiagnosis and Misdiagnosis in a High-Risk Primary Care Population in Four Latin American Countries. A Key to Enhance Disease Diagnosis: The PUMA Study. PLOS ONE. 2016 Apr 13;11(4):e0152266. **(5)** Schiavi E, Stirbulov R, Hernández Vecino R, Mercurio S, Di Boscio V. COPD Screening in Primary Care in Four Latin American Countries: Methodology of the PUMA Study. Arch Bronconeumol Engl Ed. 2014 Nov 1;50(11):469–74. **(6)** PUMA screening tool to detect COPD in high-risk patients in Chinese primary care-A validation study - PubMed [Internet]. [cited 2024 Aug 26]. Available from: https://pubmed.ncbi.nlm.nih.gov/36084011/