ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.1045

Hybrid Natural Language Processing Framework for Quality Assessment and Normalization of Clinical Data with Human-in-the-Loop Validation

Marcio Madrid[1]; Carlos Agudelo-Santos[2]; Laura Giacaman[3]; Perla Simons[4]; Melania Madrid[5]; Edil Argueta[6]
[1,3,4] Facultad de Ciencias Médicas, Universidad Nacional Autónoma de Honduras (UNAH), Honduras, *marcio.madrid@unah.edu.hn, laura.giacaman@unah.edu.hn, perla.morales@unah.edu.hn*
[2] Centro de Diagnóstico de Imágenes Biomédicas, Investigación y Rehabilitación, Universidad Nacional Autónoma de Honduras, Honduras, *carlos.agudelo@unah.edu.hn*
[5] Facultad de Biología, Universidad de Salamanca, *idu084453@usal.es*
[6] Centro Médico Nacional 20 de Noviembre, México, *edilarguetahn@gmail.com*

Abstract*– Data quality is a problem in healthcare information systems, especially when the capture of diagnostic information is not standardized. In this article, a hybrid computational framework for clinical data normalization is designed and evaluated, combining deterministic text processing methods with fuzzy similarity algorithms and a human-in-the-loop validation mechanism. The proposed system was implemented and tested on 2,777 outpatient care records from the Villa Nueva Health Center in Honduras between January and December 2024. The proposed multi-level normalization pipeline reduced the cardinality of unique diagnoses by 55.56% (from 90 to 40 categories) and geographic locations by 30.07% (from 153 to 107 categories), with 99.46% of mappings achieved by exact match (Level 1). Diagnostic entropy decreased from 3.06 to 2.55 bits (16.65%), while the Herfindahl–Hirschman index increased by 20.91%, indicating greater uniformity in the frequency distribution. The ranking stability analysis yielded Jaccard indices of 0.667 and Kendall's τ coefficients of 0.929 for the top 10 diagnoses. Thirteen ambiguous cases (0.47%) were identified that required manual review, demonstrating the feasibility of the semi-automated method. The temporal drift analysis showed an average vocabulary stability of 43% (Jaccard) between successive months. The proposed architecture is replicable, scalable, and exportable as a programmed pipeline, providing a practical solution for data governance in epidemiological surveillance systems.

Keywords*-- Data Quality, Natural Language Processing, Normalization, Fuzzy Matching, Human-in-the-Loop.

Framework Híbrido de Procesamiento de Lenguaje Natural para Evaluación de Calidad y Normalización de Datos Clínicos con Validación Human-in-the-Loop

Marcio Madrid[1]; Carlos Agudelo-Santos[2]; Laura Giacaman[3]; Perla Simons[4]; Melania Madrid[5]; Edil Argueta[6]
[1,3,4] Facultad de Ciencias Médicas, Universidad Nacional Autónoma de Honduras (UNAH), Honduras, *marcio.madrid@unah.edu.hn, laura.giacaman@unah.edu.hn, perla.morales@unah.edu.hn*
[2] Centro de Diagnóstico de Imágenes Biomédicas, Investigación y Rehabilitación, Universidad Nacional Autónoma de Honduras, Honduras, *carlos.agudelo@unah.edu.hn*
[5] Facultad de Biología, Universidad de Salamanca, *idu084453@usal.es*
[6] Centro Médico Nacional 20 de Noviembre, México, *edilarguetahn@gmail.com*

Resumen– ***La calidad de los datos es un problema en los sistemas de información sanitaria, sobre todo cuando la captura de información diagnóstica no está estandarizada. En este artículo se diseña y evalúa un marco computacional híbrido para la normalización de datos clínicos, combinando métodos deterministas de procesamiento de texto con algoritmos de similitud fuzzy y un mecanismo de validación humana en el bucle (human-in-the-loop). El sistema propuestose implementó y probó con 2,777 datos de atención ambulatoria del Centro de Salud de Villa Nueva, Honduras, entre enero y diciembre de 2024. El pipeline de normalización multinivel propuesto redujo en un 55.56% la cardinalidad de diagnósticos únicos (de 90 a 40 categorías) y en un 30.07% las localidades geográficas (de 153 a 107 categorías), con un 99.46% de mapeos por coincidencia exacta (Nivel 1). La entropía diagnóstica se redujo de 3.06 a 2.55 bits (16.65%), en tanto que el índice de Herfindahl-Hirschman creció 20.91%, mostrando mayor uniformidad en la distribución de frecuencias. El análisis de estabilidad del ranking arrojó índices de Jaccard de 0.667 y coeficientes Kendall τ de 0.929 para el top-10 de diagnósticos. Se identificaron 13 casos ambiguos (0.47%) que necesitaron revisión manual, mostrando la viabilidad del método semi-automatizado. El análisis de deriva temporal mostró una estabilidad promedio del vocabulario de 43% (Jaccard) entre meses sucesivos. La arquitectura planteada es replicable, escalable y exportable como pipeline programado, solucionando de manera práctica la gobernanza de datos en sistemas de vigilancia epidemiológica.***

Palabras clave-- calidad de datos, procesamiento de lenguaje natural, normalización, fuzzy matching, human-in-the-loop.

I. Introducción

La calidad de los datos es una de las bases para la toma de decisiones con información en salud pública. Los sistemas de información sanitaria generan cada vez más datos clínicos que, sin embargo, a menudo sufren problemas de inconsistencia, duplicación y falta de estandarización que limitan su uso para el análisis [1]. Esto es especialmente preocupante en países de ingresos medios y bajos, en los que las restricciones de infraestructura tecnológica, la heterogeneidad en los procesos de captura de datos y la falta de vocabularios controlados empeoran el problema [2].

Desde un punto de vista puramente informático, los problemas de calidad de datos no son solo problemas clínicos o administrativos, sino problemas técnicos cuantificables que afectan la validez de los análisis estadísticos, los modelos predictivos de aprendizaje automático, los indicadores de vigilancia epidemiológica y la interoperabilidad entre sistemas de información [3]. Las diferencias en la codificación de los diagnósticos clínicos, los errores ortográficos, la dispersión en categorías semánticamente equivalentes y la deriva de codificación en el tiempo crean lo que la literatura científica llama ruido de datos, que puede sesgar cualquier análisis posterior y llevar a conclusiones falsas [4].

El procesamiento del lenguaje natural (NLP) se ha convertido en una solución para resolver estos problemas, automatizando parcialmente las tareas de normalización que antes requerían mucha intervención manual y eran costosas [5]. Pero el uso de NLP en entornos clínicos tiene desafíos específicos: la existencia de lenguaje médico, abreviaturas locales no documentadas, variaciones lingüísticas regionales y códigos informales dificultan la aplicación de modelos entrenados en otros contextos lingüísticos o geográficos [6].

Los métodos híbridos que combinan reglas heurísticas con algoritmos de similitud aproximada son más robustos en casos reales que los métodos totalmente automáticos [7]. En particular, la inclusión de controles humanos en el bucle (human-in-the-loop) puede resolver casos ambiguos en los que la similitud textual no implica similitud semántica, logrando el mejor equilibrio entre precisión clínica y escalabilidad computacional [8]. Este enfoque es consciente de que la normalización de datos sanitarios no es totalmente automatizable y necesita combinar capacidades algorítmicas con juicio experto.

En Honduras, la Encuesta Nacional de Demografía y Salud (ENDESA/MICS) es la principal fuente de información poblacional sobre salud, bienestar y características socioeconómicas [9]. Pero los datos de atención primaria a menudo adolecen de falta de estandarización, lo que impide su uso para estudios epidemiológicos a gran escala y dificulta la comparación entre centros sanitarios.

Los marcos existentes de normalización de PLN para datos clínicos se han desarrollado predominantemente para

historiales clínicos electrónicos en inglés en entornos de altos ingresos con entornos estructurados de entrada de datos [5], [7]. Estos sistemas suelen asumir la disponibilidad de vocabularios de codificación estandarizados (CIE-10, SNOMED-CT), abreviaturas locales bien documentadas y prácticas de entrada de datos relativamente consistentes, condiciones que no se cumplen en los centros de atención primaria de América Latina. El conjunto de datos del Centro de Salud Villa Nueva ilustra esta carencia de forma concreta: campos de diagnóstico en texto libre codificados con abreviaturas informales (Faringotonsiliitis bacteriana aguda - FAAB, Faringotonsiliitis viral aguda - FAAV), variantes ortográficas regionales y sin vocabulario obligatorio, produciendo 90 cadenas diagnósticas únicas para una instalación cuyo alcance clínico normalmente proporcionaría menos de 25 condiciones distintas. Los enfoques estándar de emparejamiento de cadenas fallan aquí porque no pueden distinguir entre ruido ortográfico y diferencias semánticas genuinas sin conocimientos clínicos, y los enfoques basados en grandes modelos de lenguaje son poco prácticos en entornos informáticos de pocos recursos con datos sensibles de pacientes.

El presente trabajo aborda esta carencia diseñando, implementando y evaluando empíricamente una cadena híbrida de PLN calibrada específicamente para registros de atención primaria en entornos de bajo rendimiento de habla hispana. Sus contribuciones son: (a) un marco formal y cuantitativo de auditoría de calidad de datos aplicable a los registros clínicos en texto libre; (b) una cascada de normalización de tres niveles que maximiza la cobertura mediante procesamiento determinista mientras controla fusiones erróneas mediante enrutamiento en banda de confianza; (c) un diccionario maestro versionado con registro completo de auditoría que permita la reproducibilidad y la retrocesión; (d) un análisis de deriva temporal que cuantifica la degradación del vocabulario a lo largo del tiempo, una dimensión rara vez reportada en estudios similares; y (e) un estudio de caso validado que demuestre el impacto epidemiológico posterior de la normalización en la vigilancia de las enfermedades respiratorias. En conjunto, estos elementos constituyen una herramienta de gobernanza de datos práctica, auditable y transferible para sistemas de información sanitaria en entornos donde la codificación estandarizada es aspiracional más que operativa.

II. Materiales y Métodos

A. Fuente de Datos y Contexto Institucional

El estudio utilizó registros de atención ambulatoria del Centro de Salud Villa Nueva, establecimiento de primer nivel ubicado en el Distrito Central de Francisco Morazán, Honduras. La base de datos comprendió 2,777 registros de consultas realizadas entre enero y diciembre de 2024, estructurados según el siguiente esquema: número de registro secuencial, fecha de atención, sexo del paciente (codificado como H/M), edad (en años o formato años.meses para menores de 5 años), colonia de residencia (texto libre) y diagnóstico clínico (texto libre). Los datos fueron proporcionados en formato de hoja de cálculo electrónica (.xlsx) siguiendo los lineamientos de la ENDESA/MICS 2019 para la estructuración de información sanitaria [9].

El área de influencia del centro de salud incluye múltiples colonias del sector nororiental de Tegucigalpa, con una población caracterizada por estratos socioeconómicos diversos y patrones epidemiológicos típicos de zonas periurbanas de América Central. La selección de esta fuente de datos obedeció a su representatividad de los desafíos de calidad comúnmente encontrados en sistemas de información de atención primaria en la región.

B. Auditoría de Calidad: Definiciones Operativas y Métricas Formales

Se estableció un conjunto de métricas formales para caracterizar cuantitativamente la calidad de los datos, operacionalizadas según las siguientes definiciones matemáticas:

1) Completitud (C): Proporción de registros con valores no nulos en cada variable. Se calculó como $C = (N - N\_missing) / N \times 100$, donde N representa el total de registros y N_missing los valores faltantes. Un valor de C = 100% indica ausencia de datos faltantes.

2) Cardinalidad diagnóstica (K): Número de valores únicos en la variable de diagnóstico antes y después de la normalización. Una cardinalidad excesivamente alta en relación con el número esperado de diagnósticos sugiere fragmentación artificial por variaciones de entrada. La reducción porcentual se calcula como $\Delta K = (K\_pre - K\_post) / K\_pre \times 100$.

3) Tasa de duplicados (D): Proporción de registros idénticos en las variables clínicas (excluyendo el identificador único). Se identificaron duplicados exactos mediante comparación de tuplas (fecha, sexo, edad, colonia, diagnóstico), calculando $D = N\_duplicados / N \times 100$.

4) Outliers (O): Registros con valores biológicamente implausibles. Para la variable edad, se definieron como outliers aquellos registros con edad > 100 años (límite superior razonable) o edad < 0 (valor imposible). Adicionalmente, se cuantificaron las edades expresadas en formato decimal (años.meses), indicativas de inconsistencia en la codificación.

5) Entropía diagnóstica (H): Medida de dispersión en la distribución de diagnósticos basada en la teoría de la información, calculada mediante la fórmula de Shannon: $H = -\Sigma\, p\_i \times \log_2(p\_i)$, donde p_i es la frecuencia relativa de cada diagnóstico i. Valores más altos indican mayor dispersión (fragmentación), mientras que valores menores sugieren concentración en categorías dominantes.

6) Índice de concentración Herfindahl-Hirschman (HHI): Calculado como $HHI = \Sigma\, s\_i^2$, donde s_i representa la participación porcentual de cada diagnóstico en el total de registros. Valores más altos indican mayor concentración en pocas categorías, complementando la interpretación de la entropía.

7) Drift de codificación temporal: Cambios en el vocabulario diagnóstico a lo largo del tiempo. Se cuantificó mediante el índice de Jaccard entre los conjuntos de diagnósticos únicos de meses consecutivos: $J(A,B) = |A \cap B| / |A \cup B|$. Valores cercanos a 1 indican estabilidad del vocabulario; valores bajos sugieren introducción frecuente de nuevas variantes.

C. *Arquitectura del Pipeline de Calidad de Datos*

Se diseñó una arquitectura modular de seis componentes interconectados (Fig. 1). El flujo de procesamiento inicia con la ingesta de datos desde la fuente original, seguido por una fase de profiling que genera el diagnóstico inicial de calidad mediante el cálculo de todas las métricas definidas en la sección 2.2. Los datos atraviesan entonces una etapa de validación estructural (reglas) que verifica formatos y rangos permisibles, antes de ingresar al módulo de normalización NLP que constituye el núcleo del sistema.

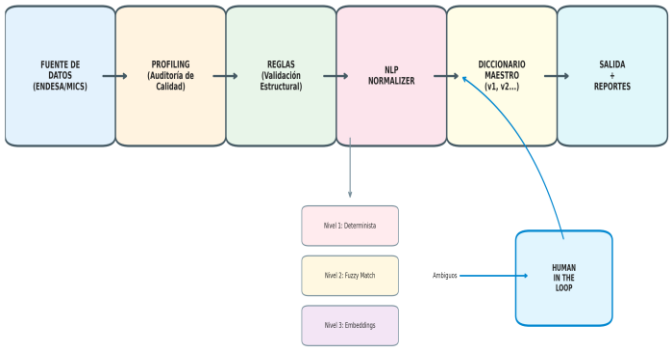

Fig. 1 Arquitectura del framework de calidad y normalización de datos clínicos. El flujo procesa datos desde la fuente original a través de módulos de profiling, validación estructural y normalización NLP multinivel, con retroalimentación desde el componente human-in-the-loop hacia el diccionario maestro versionado.

El módulo NLP Normalizer implementa tres niveles de procesamiento escalonado (detallados en la sección 2.4) y se conecta bidireccionalmente con un Diccionario Maestro versionado que almacena los mapeos canónicos. Las decisiones ambiguas se derivan a un componente Human-in-the-Loop para revisión experta, cuyas resoluciones retroalimentan el diccionario en un ciclo de mejora continua. Finalmente, los datos normalizados y los reportes de transformación se generan en la etapa de salida.

El Diccionario Maestro implementa un esquema de versionado semántico (v1.0, v1.1, v2.0...) que permite rastrear la evolución de los mapeos y facilita la reproducibilidad de los análisis. Cada transformación aplicada queda registrada en un log de auditoría que documenta, el identificador del registro afectado, valor original, valor normalizado, método aplicado (Nivel 1/2/3), score de confianza, y marca temporal ISO 8601.

D. *Normalización con NLP: Propuesta Escalonada Multinivel*

La normalización se realizó en tres niveles de complejidad creciente para maximizar la precisión a mínimo costo computacional:

1) Nivel 1 - Normalización determinista: Implica procesos de limpieza sin heurísticas probabilísticas y totalmente reproducibles: (a) convertir todo a mayúsculas para eliminar diferencias de capitalización, (b) eliminar espacios redundantes (iniciales, finales y múltiples), (c) eliminar caracteres especiales sin significado clínico y (d) estandarizar caracteres acentuados (á → a, é → e, etc.) para homologar variantes ortográficas [10]. Estos cambios pueden revertirse a través del log de auditoría. La implementación hizo uso de regex y funciones de manipulación de cadenas de Python 3.12 [11].

2) Nivel 2 - Normalización fuzzy: Para palabras que no se encuentran en el diccionario maestro (después del Nivel 1), se usaron algoritmos de similitud aproximada basados en distancia de edición. Se usó la métrica de Levenshtein normalizada, a través de la biblioteca RapidFuzz (v3.5), la cual proporciona implementaciones optimizadas en C++ con bindings en Python [12]. Se definió un punto de corte de similitud del 85%, establecido empíricamente para optimizar la sensibilidad, manteniendo una especificidad clínicamente aceptable. Los candidatos con puntuaciones entre el 60% y el 84% fueron considerados dudosos.

3) Nivel 3 - Similitud semántica (conceptual): Para los casos en que la diferencia va más allá de la ortográfica y llega a la conceptual, se desarrolló un módulo de similitud semántica usando: (a) tokenización y Jaccard entre conjuntos de tokens, y (b) bonificación si pertenecen a la misma categoría clínica (respiratorio, dermatológico, oftalmológico). En futuras implementaciones, este nivel se puede ampliar utilizando embeddings de frases (por ejemplo, con Sentence-BERT) para capturar similitudes semánticas más ricas [13], [14].

4) Desambiguación (Human-in-the-Loop): Los casos con similitud media (60-84%) se encolaron en una cola estructurada de revisión, mostrando al revisor experto: (a) el término original y su contexto clínico si lo hay, (b) los posibles candidatos del diccionario con sus puntuaciones de similitud, y (c) las estadísticas de frecuencia de uso. Las decisiones del revisor se registraron y se realimentaron al diccionario principal, mejorando cada vez más la exactitud del sistema en un proceso de aprendizaje activo conceptual [15].

Los casos ambiguos fueron revisados por dos autores profesional(es) sanitario(s) clínicamente formado(s) con al menos diez años de experiencia en atención primaria. Las discrepancias entre revisores se resolvieron por consenso, y el acuerdo entre evaluadores se evaluó usando el κ de Cohen [16]. A cada revisor se le proporcionó el término original, el contexto clínico disponible (edad del paciente, sexo y colonia de residencia), los mapeos de candidatos clasificados por puntuación de similitud y estadísticas de frecuencia a nivel de

corpus para cada candidato. Se instruyó a los revisores para seleccionar el término canónico que mejor representara la entidad clínica descrita, o para marcar el registro como no mapeable si no existía un mapeo adecuado. Todas las decisiones se registraban con una marca de tiempo y el identificador del revisor, lo que permitía la trazabilidad total del paso de resolución humana.

E. Evaluación pre/post Normalización

Se diseñó un protocolo de evaluación experimental riguroso comparando las métricas de calidad antes y después de la aplicación del pipeline. Los indicadores evaluados incluyeron:

1) Reducción de cardinalidad diagnóstica: diferencia porcentual en el número de categorías únicas (ΔK%).

2) Migración de registros: cuantificación de registros reasignados desde variantes fragmentadas hacia categorías canónicas consolidadas.

3) Estabilidad del ranking top-k: índice de Jaccard y coeficiente de correlación Kendall τ entre los k diagnósticos más frecuentes pre y post normalización, para $k \in \{5, 10, 15\}$.

4) Variación de entropía: diferencia en bits entre las distribuciones pre y post, indicativa del grado de consolidación logrado.

5) Impacto en series temporales: análisis visual y cuantitativo de cómo la consolidación afecta la detección de tendencias epidemiológicas.

F. Implementación Técnica y Reproducibilidad

El pipeline se codificó en Python 3.12 utilizando las siguientes bibliotecas con sus respectivas versiones: pandas (v2.0) para manipulación de datos tabulares, RapidFuzz (v3.5) para cálculos de similitud fuzzy optimizados, NumPy (v1.26) para operaciones numéricas vectorizadas, SciPy (v1.11) para estadísticas, incluyendo el test de Kendall, y Matplotlib (v3.8) / Seaborn (v0.13) para visualización. El código está modularizado en funciones bien documentadas, lo que permite ejecutarlo como un script independiente o integrarlo en pipelines automatizados mediante cron [17]-[23].

Las opciones configurables del sistema son: umbral de similitud fuzzy (85% por defecto), umbral de ambigüedad (60% por defecto), ruta al diccionario maestro (versionado) y formato de salida de informes (JSON/CSV/Excel). Las medidas de privacidad incluyeron la anonimización de identificadores personales en todos los casos mostrados y la creación de estadísticas solamente para los informes públicos.

III. RESULTADOS

A. Perfil de Calidad Inicial

El primer análisis de profiling arrojó una base de datos con 2,777 registros y completitud del 100% en todas sus variables (ningún valor faltante). La distribución demográfica fue mayoritariamente femenina (60.82%, n=1,689) que masculina (39.18%, n=1,088), con un registro presentando inconsistencia de codificación (espacio adicional en el valor). La edad media fue de 20.42 años (DE = 21.20), mediana de 11 años, una población mayoritariamente pediátrica y adulto-joven como suele ser la demanda en atención primaria.

La distribución por grupos de edad mostró que el 29.46% (n=818) fueron menores de 5 años, el 29.24% (n=812) de 5-17 años, el 20.99% (n=583) adultos de 18-39 años, el 14.69% (n=408) de 40-64 años, y solo el 5.58% (n=155) mayores de 65 años. Se detectó un dato de edad imposible de 612 años (error de digitación probablemente), que representa el 0.04% de outliers. Además, 716 registros (25.78%) tenían edades decimales (años.meses), una inconsistencia en la codificación que es comprensible, pero que dificulta el procesamiento automatizado.

La variable de diagnóstico arrojó 90 valores únicos pre-normalización, una gran dispersión ya que un centro de atención primaria suele manejar un vocabulario más limitado de patologías comunes. La variable colonia arrojó 153 categorías diferentes, muchas de ellas variantes ortográficas de un mismo lugar. Se identificaron 7 registros duplicados idénticos (0.25%), y el análisis de fechas identificó 1,510 registros con formato de fecha no estándar que necesitaron ser analizados sintácticamente. Ver Tabla I.

TABLA I
MÉTRICAS DE CALIDAD DE DATOS PRE Y POST NORMALIZACIÓN

Métrica	Pre-normalización	Post-normalización	Cambio (%)
Completitud (%)	100.00	100.00	0.00
Diagnósticos únicos	90	40	-55.56
Colonias únicas	153	107	-30.07
Duplicados exactos	7 (0.25%)	7 (0.25%)	0.00
Outliers de edad	1 (0.04%)	1 (0.04%)	0.00
Entropía diagnóstica (bits)	3.0564	2.5476	-16.65
Índice HHI	0.2214	0.2677	+20.91

B. Resultados del Proceso de Normalización Multinivel

La aplicación del pipeline de normalización redujo la cardinalidad de las variables categóricas (Fig. 2 y 3). Los 90 diagnósticos se consolidaron en 40 categorías, reduciendo un 55.56%. Las 153 colonias se redujeron a 107 categorías (30.07% de reducción) mediante la unificación de variantes con diferentes prefijos (RES., RESI., COL., COLONIA) y corrección de errores ortográficos.

La distribución de métodos de normalización (Fig. 2c) mostró que el 99.46% de los registros (n=2,762) fueron mapeados por coincidencia exacta con el diccionario maestro tras la limpieza del Nivel 1. Este alto porcentaje indica que la variabilidad observada se debe más a inconsistencias superficiales (espacios, mayúsculas, tildes) que a diferencias semánticas genuinas. Solo 13 registros (0.47%) fueron clasificados como ambiguos para revisión, y 2 registros (0.07%) no pudieron ser mapeados, manteniendo sus valores normalizados.

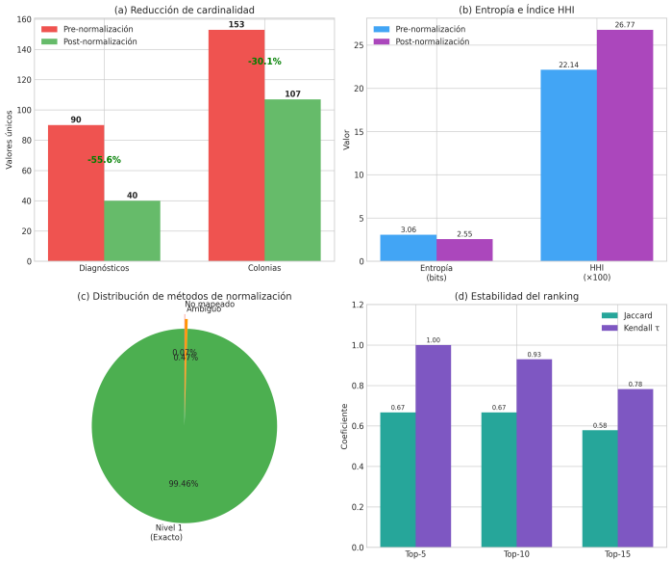

Fig. 2 Métricas de calidad de datos pre y post normalización: (a) reducción de cardinalidad en diagnósticos y colonias; (b) variación de entropía e índice HHI; (c) distribución de métodos de normalización aplicados; (d) estabilidad del ranking diagnóstico medida por Jaccard y Kendall τ.

La Tabla II presenta un desglose nivel por nivel de la cobertura de normalización. El procesamiento determinista de nivel 1 resolvió el 99,46% de los registros, demostrando que la fuente dominante de variabilidad en este conjunto de datos era la inconsistencia superficial (capitalización, espacios en blanco, diacríticos) más que la divergencia semántica. El emparejamiento difuso de nivel 2 identificó 13 registros (0,47%) dentro de la banda ambigua de confianza (60–84%), que fueron redirigidos a revisión humana en lugar de resolverse de forma autónoma, una elección deliberada de diseño para evitar mapeos falsos en casos límite clínicamente sensibles. La similitud semántica de Nivel 3 no se activó en este conjunto de datos, ya que no quedaban casos sin resolver tras los Niveles 1 y 2 con el diccionario existente. Dos registros (0,07%) no podían ser mapeados por ningún método automatizado y se mantuvieron con sus formularios normalizados a la espera de actualizaciones del diccionario.

C. *Casos Ambiguos y Validación Human-in-the-Loop*

Los 13 casos ambiguos fueron principalmente variantes de dermatitis atípicas. Todos los casos ambiguos son dermatológicos, lo que sugiere mayor variabilidad terminológica en esta área. Estos casos destacan la importancia del human-in-the-loop, como en DERMATITIS VULVAR (score 84.8% con DERMATITIS SOLAR) es una condición clínica distinta que no debe fusionarse automáticamente. DERNATITIS AMINOCAL (75.7% con DERMATITIS ATOPICA) parece ser un error de DERMATITIS AMONIACAL (dermatitis del pañal por contacto con amoniaco), requiriendo criterio clínico para su categorización.

TABLA II
COBERTURA DE NORMALIZACIÓN POR NIVEL DE PIPELINE

Nivel	Método	Registros resueltos	Cobertura (%)	Acumulado (%)
Nivel 1	Determinístico	2,762	99.46	99.46
Nivel 2	Fuzzy (≥85%)	0*	0.00	99.46
HITL	Revisión humana	13	0.47	99.93
No mapeado	No se encontró ningún mapa	2	0.07	100.00
Total		2,777	100.00	-

* El Nivel 2 identificó 13 registros en la banda ambigua (60–84%) pero los enrutó a HITL en lugar de resolverlos de forma autónoma, en consonancia con la política conservadora de mapeo

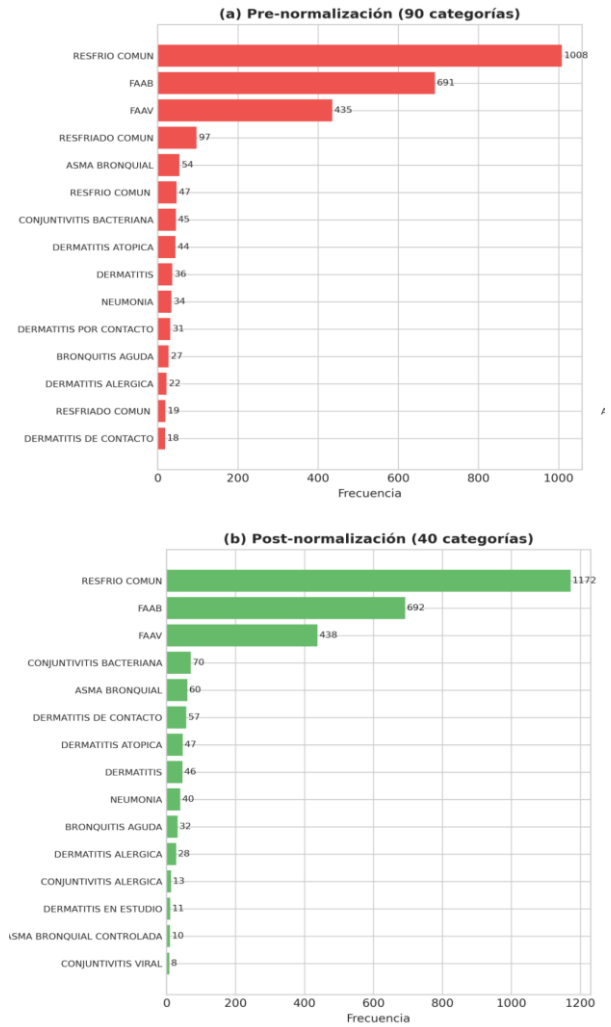

Fig. 3 Comparación de los 15 diagnósticos más frecuentes antes y después de la normalización. Nótese la consolidación de variantes fragmentadas (ej. RESFRIADO COMUN → RESFRIO COMUN) y el incremento en las frecuencias de las categorías canónicas.

D. *Análisis de Drift Temporal*

El análisis de estabilidad del vocabulario diagnóstico en el tiempo (Fig. 4) mostró un índice de Jaccard promedio de 0.430 entre meses sucesivos, lo que significa que alrededor del

43% de los términos diagnósticos se mantienen estables de un mes a otro. Hubo cambios importantes en este índice, desde un mínimo de 0.250 (junio→julio) hasta un máximo de 0.556 (julio→agosto y septiembre→octubre).

El análisis de palabras nuevas arrojó que el mes de enero (inicio) tuvo 11 palabras distintas como vocabulario base y, posteriormente, adiciones que oscilaron entre 1 y 11 palabras nuevas por mes. En particular, los meses de marzo, abril y noviembre presentaron mayor número de nuevas variantes terminológicas. Este patrón de deriva puede deberse a cambios en el personal médico, nuevas guías clínicas o simplemente variabilidad en las prácticas de documentación.

La distribución de las puntuaciones de similitud entre todos los registros se muestra en la Figura 4, que ilustra las tres zonas operativas definidas por estos umbrales: la zona de no mapeo (< 60%), la zona ambigua sujeta a revisión humana (60–84%) y la zona de coincidencia difusa resuelta automáticamente (≥ 85%), siendo las coincidencias exactas (100%) el clúster dominante.

E. Impacto en el Ranking e Indicadores de Vigilancia

La consolidación de variantes produjo cambios en la distribución de frecuencias de los diagnósticos principales. El diagnóstico RESFRIO COMUN consolidó 1,172 registros (42.20% del total) al incluir RESFRIADO COMUN (97 registros) y RESFRIO COMUN con espacio (47 registros), reafirmando su posición dominante. Los diagnósticos FAAB (Faringoamigdalitis Aguda Bacteriana) y FAAV (Faringoamigdalitis Aguda Viral) ocuparon la segunda y tercera posición, junto con RESFRIO COMUN, sumando el 82.89% de las atenciones.

El análisis de estabilidad del ranking (Figura 2d) mostró índices de Jaccard de 0.667 para top-5 y top-10, y 0.579 para top-15, indicando que la normalización preserva las posiciones de los diagnósticos más frecuentes y reorganiza las categorías de menor frecuencia. Los coeficientes Kendall τ fueron 1.000 (top-5), 0.929 (top-10) y 0.782 (top-15), confirmando alta correlación ordinal entre los rankings pre y post normalización.

F. Caso de Uso: Vigilancia de Infecciones Respiratorias Agudas

Se examinó el comportamiento de las infecciones respiratorias agudas (IRA) durante el estudio para ilustrar el impacto de la normalización en la analítica epidemiológica (Fig. 5). Antes de la normalización, las series temporales de RESFRIO COMUN y RESFRIADO COMUN aparecían como líneas separadas, fragmentando el análisis de tendencias y subestimando la carga real de esta condición.

La serie consolidada mostró un patrón estacional más definido, con picos en abril-mayo (137 casos mensuales) coincidiendo con el inicio de la temporada lluviosa en Honduras, y valores menores en diciembre-enero (68-78 casos mensuales). Este patrón, consistente con la epidemiología de las IRA en Centroamérica, había quedado oscurecido por la fragmentación. Los diagnósticos respiratorios (RESFRIO COMUN, FAAB, FAAV, ASMA BRONQUIAL, BRONQUITIS AGUDA, NEUMONIA) representaron el 83.4% de las atenciones, confirmando el perfil de un centro de atención primaria.

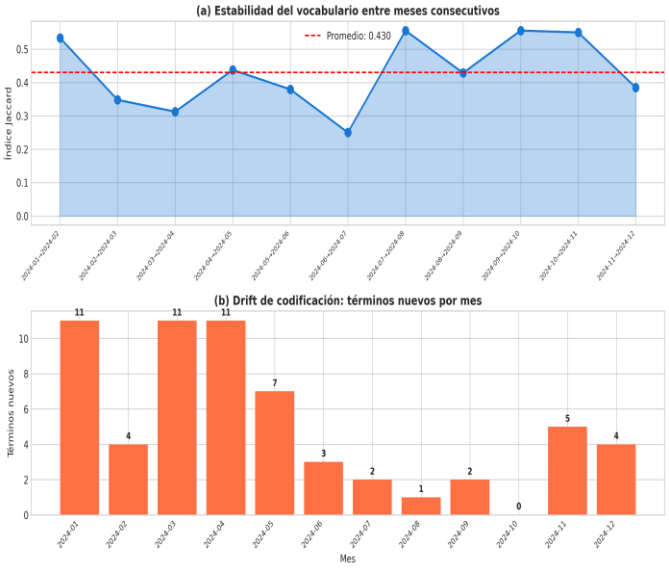

Fig. 4 Análisis de drift de codificación temporal: (a) índice de Jaccard entre vocabularios diagnósticos de meses consecutivos, con promedio de 0.430 indicando estabilidad moderada; (b) número de términos nuevos incorporados cada mes, evidenciando la evolución continua del vocabulario.

El análisis geográfico post-normalización identificó las colonias Villa Nueva No. 6 (750 registros), Villa Nueva. 5 (605) y Villa Nueva. 1 (228) concentran la mayor demanda de atención por IRA, información clave para planificar intervenciones preventivas.

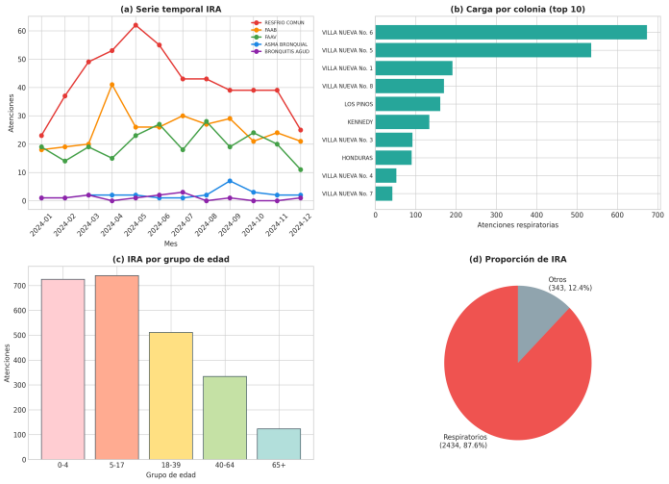

Fig. 5 Caso de uso: vigilancia epidemiológica de infecciones respiratorias agudas (IRA). (a) Series temporales consolidadas mostrando estacionalidad; (b) distribución geográfica por colonia; (c) distribución por grupo etario; (d) proporción de atenciones respiratorias sobre el total.

IV. Discusión

El framework desarrollado demostró ser efectivo para la normalización de datos clínicos en atención primaria, logrando reducciones en la fragmentación de categorías diagnósticas y geográficas con un enfoque automatizado. La arquitectura híbrida propuesta, que combina procesamiento determinista con técnicas fuzzy y validación humana, ofrece un balance entre automatización y precisión, adecuado para entornos con recursos limitados [23].

La alta proporción de mapeos exactos (99.46%) tras la normalización de Nivel 1 es un hallazgo con implicaciones prácticas. Este resultado sugiere que la variabilidad en los datos originales se debe más a inconsistencias de captura (espacios, variaciones de mayúsculas/minúsculas, omisión de tildes) que a diferencias semánticas o a múltiples condiciones clínicas. Intervenciones simples en la interfaz de captura de datos, como validación en tiempo real con autocompletado basado en el diccionario maestro, podrían prevenir la mayoría de estos problemas, reduciendo la necesidad de normalización posterior [24].

El enfoque escalonado de normalización permitió trazabilidad completa del proceso, documentando el método y el nivel de confianza para cada transformación. La transparencia algorítmica es fundamental para la aceptación de sistemas automatizados en contextos clínicos, donde la rendición de cuentas sobre decisiones de procesamiento es un requisito ético y regulatorio [26]. La capacidad de auditar y revertir transformaciones específicas es una ventaja sobre enfoques de caja negra.

El análisis de drift temporal reveló patrones de evolución del vocabulario diagnóstico. El índice de Jaccard de 0.430 entre meses consecutivos indica una variabilidad considerable que requiere atención institucional. Este fenómeno puede tener múltiples causas: rotación de personal con distintos estilos de registro, actualizaciones en guías clínicas, nuevas categorías diagnósticas o inconsistencia en la documentación. El drift temporal complica los análisis longitudinales y justifica controles de calidad continuos [27].

Entre las limitaciones del estudio, se reconoce que el diccionario maestro fue creado para el Centro de Salud de Villa Nueva, y su transferibilidad a otros lugares requiere validación adicional. La terminología local, incluidas las abreviaturas FAAB y FAAV, podría no ser universal en todos los sistemas de salud de Honduras o Centroamérica. Se recomienda adaptar y validar el diccionario antes de aplicarlo en nuevos contextos [28].

El riesgo de sobre-normalización es una consideración crítica. La consolidación excesiva de categorías diagnósticas podría llevar a pérdida de información clínica relevante. La unificación de DERMATITIS DE CONTACTO y DERMATITIS POR CONTACTO se justifica por su equivalencia semántica, pero fusionar DERMATITIS ATÓPICA con DERMATITIS ALÉRGICA es cuestionable, ya que representan entidades con diferentes mecanismos fisiopatológicos. El mecanismo human-in-the-loop gestiona casos limítrofes donde el criterio clínico prevalece sobre la similitud textual [29].

La casi total dominancia de la resolución de Nivel 1 (99,46%) justifica una discusión directa sobre el valor incremental de la arquitectura multinivel. Un enfoque solo con reglas habría logrado una cobertura cuantitativa idéntica en este conjunto de datos; sin embargo, habría dejado los 13 casos límite (0,47%) o bien forzados a la coincidencia léxica más cercana, arriesgando confusiones clínicamente incorrectas como mapear DERMATITIS VULVAR a DERMATITIS SOLAR; o descartados como imposibles de mapear, introduciendo pérdida de información. La contribución explícita de la capa difusa en este conjunto de datos no fue la resolución autónoma, sino la puntuación y triaje de confianza: al calcular la similitud con el diccionario maestro, el Nivel 2 permitió una decisión de enrutamiento informada por la evidencia que preservó la supervisión del clínico donde importaba. Además, la cascada de tres niveles está diseñada para escalar con elegancia: en conjuntos de datos con mayor heterogeneidad terminológica (por ejemplo, registros multi-instalación o multirregionales), se espera que el Nivel 2 resuelva una mayor proporción de casos de forma autónoma, y la similitud basada en incrustaciones de Nivel 3 abordaría variaciones conceptuales que eluden las métricas de distancia de edición. La política conservadora de mapeo adoptada aquí, prefiere la revisión humana sobre la resolución automatizada en casos ambiguos; refleja un compromiso explícito en el diseño que prioriza la precisión clínica sobre el rendimiento, apropiado para un contexto de primer despliegue [30], [31].

Una limitación metodológica es que la validación humana en el bucle para los 13 casos ambiguos fue realizada por un único revisor clínico, sin una evaluación formal de fiabilidad entre evaluadores. Aunque el pequeño número de casos ambiguos (0,47%) limita el impacto práctico de esta restricción, las implementaciones futuras deberían incluir al menos dos revisores independientes e informar sobre el acuerdo entre evaluadores (por ejemplo, el κ de Cohen) para reforzar la validez del paso de desambiguación.

Las elecciones de diseño integradas en este marco reflejan limitaciones y oportunidades específicas del contexto de atención primaria latinoamericana. La dependencia de un diccionario maestro construido localmente, en lugar del mapeo directo de la CIE-10, es una respuesta deliberada a la brecha entre los sistemas formales de clasificación y el vocabulario informal que realmente utilizan los clínicos en el punto de atención [28]. Abreviaturas como FAAB y FAAV, aunque ausentes en los estándares internacionales de codificación, son operativamente dominantes en Honduras y probablemente en sistemas centroamericanos vecinos. La capa de preprocesamiento determinista, responsable del 99,46% de todos los mapeos exitosos; es especialmente valiosa en contextos donde las inconsistencias en la entrada de datos surgen por la entrada manual en interfaces básicas de hojas de cálculo, sin autocompletado ni aplicación de vocabulario. El enfoque del diccionario versionado también aborda un desafío estructural común a los sistemas sanitarios, como ser la

ausencia de infraestructura informática dedicada para el mantenimiento del diccionario, haciendo que el artefacto de gobernanza en sí sea el entregable principal en lugar de la plataforma de software.

La implementación de un diccionario vivo con versionado temporal permite la evolución del sistema a medida que se acumula experiencia y se incorporan términos. Este enfoque de gobernanza de datos adaptativa contrasta con los enfoques estáticos basados en catálogos predefinidos (como CIE-10), ofreciendo mayor flexibilidad en contextos de terminología en evolución y manteniendo la capacidad de mapeo a clasificaciones estandarizadas cuando sea necesario [32].

V. Conclusión

Este estudio propone y valida un framework computacional reproducible para evaluar y normalizar datos clínicos mediante procesamiento de lenguaje natural, incorporando una evaluación cuantitativa del impacto en métricas de calidad e indicadores para la vigilancia epidemiológica. Los resultados muestran mejoras en la consistencia y utilidad analítica de los registros: la normalización redujo la fragmentación semántica, con una disminución del 55.56% en diagnósticos y del 30.07% en localidades, y alcanzó un 99.46% de mapeos por coincidencia exacta tras un proceso determinista. La depuración mejoró la coherencia de las distribuciones, reflejada en la reducción de la entropía diagnóstica (−16.65%) y el aumento del índice HHI (+20.91%), sin afectar la estructura epidemiológica principal: el análisis de estabilidad mostró alta preservación del orden de los diagnósticos más frecuentes (Kendall τ = 0.929 para el top-10).

Un aporte metodológico central es la arquitectura híbrida multinivel (determinista → fuzzy → human-in-the-loop), eficaz para equilibrar automatización y precisión clínica, aislando la fracción de casos ambiguos (0.47%) que requiere juicio experto. El caso de uso en vigilancia respiratoria muestra el valor práctico del enfoque: al reducir la fragmentación, se hicieron visibles patrones estacionales antes atenuados, lo que refuerza la normalización como paso clave para análisis robustos en salud pública.

Como pipeline de código abierto, escalable y adaptable a la atención primaria en América Latina, su adopción podría fortalecer la calidad de los datos para la toma de decisiones, la vigilancia epidemiológica y la evaluación de programas sanitarios. Como líneas de trabajo futuro, se propone incorporar representaciones semánticas profundas (p. ej., Sentence-BERT, BioBERT) para variaciones conceptuales, integrar estándares internacionales (CIE-10, SNOMED-CT) mediante mapeos validados, desarrollar interfaces para facilitar la adopción por personal no técnico y realizar evaluaciones multicéntricas para estimar la generalización del framework en distintos establecimientos de salud.

Agradecimientos

Esta investigación fue apoyada por el Instituto de Investigación en Ciencias Médicas y Derecho a la Salud (ICIMEDES) de la Facultad de Ciencias Médicas (FCM) y por la Dirección de Investigación Científica, Humanística y Tecnológica (DICIHT) todas ellas pertenecientes a la Universidad Nacional Autónoma de Honduras (UNAH). Agradecemos a los Profesor Isaac Zablah por su aporte metodológico en la elaboración de este artículo. De igual forma se agradece a las autoridades y personal del Centro de Salud Villa Nueva.

Referencias

[1] M. G. Kahn, D. Callahan, J. Barnard, et al., "A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data", *eGEMs (Generating Evid. Methods Improv. Patient Outcomes)*, vol. 4, no. 1, art. 18, 2016. DOI: 10.13063/2327-9214.1244

[2] T. Pantoja et al., "Implementation strategies for health systems in low-income countries: an overview of systematic reviews", *Cochrane Database Syst. Rev*., no. 9, Art. no. CD011086, Sep. 2017, doi: 10.1002/14651858.CD011086.pub2

[3] S. N. Murphy, G. Weber, M. Mendis, et al., "Serving the Enterprise and Beyond with Informatics for Integrating Biology and the Bedside (i2b2)", *J. Am. Med. Inform. Assoc.*, vol. 17, no. 2, pp. 124-130, Mar. 2010. DOI: 10.1136/jamia.2009.000893

[4] G. Hripcsak y D. J. Albers, "Next-Generation Phenotyping of Electronic Health Records", *J. Am. Med. Inform. Assoc.*, vol. 20, no. 1, pp. 117-121, Ene. 2013. DOI: 10.1136/amiajnl-2012-001145

[5] Ö. Uzuner, Y. Luo, y P. Szolovits, "Evaluating the State-of-the-Art in Automatic De-identification", *J. Am. Med. Inform. Assoc.*, vol. 14, no. 5, pp. 550-563, Sep. 2007. DOI: 10.1197/jamia.M2444

[6] S. Velupillai, H. Dalianis, M. Hassel, y G. H. Nilsson, "Developing a Standard for De-identifying Electronic Patient Records Written in Swedish", *Int. J. Med. Inform.*, vol. 78, no. 12, pp. e19-e26, Dic. 2009. DOI: 10.1016/j.ijmedinf.2009.04.005

[7] P. B. Jensen, L. J. Jensen, y S. Brunak, "Mining Electronic Health Records: Towards Better Research Applications and Clinical Care", *Nat. Rev. Genet.*, vol. 13, no. 6, pp. 395-405, Jun. 2012. DOI: 10.1038/nrg3208

[8] A. Holzinger, "Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-Loop?", *Brain Inform.*, vol. 3, no. 2, pp. 119-131, Jun. 2016. DOI: 10.1007/s40708-016-0042-6

[9] Instituto Nacional de Estadísticas de Honduras, "Encuesta Nacional de Demografía y Salud ENDESA/MICS 2019", *INE*, Tegucigalpa, Honduras, 2020. Disponible: https://ine.gob.hn/bases-de-datos/

[10] R. Sproat, *Text Normalization*. Cambridge, UK: Cambridge University Press, 2010.

[11] Python Software Foundation, "Python 3.12 Documentation," Python Documentation, 2023. [Online] Available: https://docs.python.org/3.12/

[12] M. Bachmann, "RapidFuzz 3.5.1: Rapid fuzzy string matching in Python and C++ using the Levenshtein Distance", *Python Package Index (PyPI)*, Nov. 1, 2023. Accessed: Jan. 27, 2026.

[13] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", *Bull. Soc. Vaudoise Sci. Nat.*, vol. 37, pp. 547–579, 1901.

[14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410

[15] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning", *AI Magazine,* vol. 35, no. 4, pp. 105–120, 2014, doi: 10.1609/aimag.v35i4.2513.

[16] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104

[17]The pandas development team, “pandas-dev/pandas: Pandas (v2.0.0)”, *Zenodo*, Apr. 3, 2023, doi: 10.5281/zenodo.7794821.
[18]NumPy Developers, “NumPy v1.26 Reference Manual”, [Online]. Available: https://numpy.org/doc/1.26/reference/index.html
[19]P. Virtanen et al., “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”, *Nat. Methods*, vol. 17, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.
[20]SciPy Developers, “SciPy v1.11.1 Manual”, 2023. [Online]. Available: https://docs.scipy.org/doc/scipy-1.11.1/reference/generated/scipy.stats.kendalltau.html
[21]J. D. Hunter, “Matplotlib: A 2D Graphics Environment”, *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55
[22]M. L. Waskom, “seaborn: statistical data visualization”, *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, 2021, doi: 10.21105/joss.03021
[23]P. Vixie, “cron(8) — daemon to execute scheduled commands,” *Linux man-pages (man7.org).* [Online]. Available: https://man7.org/linux/man-pages/man8/cron.8.html
[24]L. Ohno-Machado, “Realizing the full potential of electronic health records: the role of natural language processing”, J. Am. Med. Inform. Assoc., vol. 18, no. 5, p. 539, Sep. 2011, doi: 10.1136/amiajnl-2011-000501
[25]A. Wright, D. W. Bates, B. Middleton, et al., "Creating and Sharing Clinical Decision Support Content with Web 2.0: Issues and Examples", *J. Biomed. Inform.*, vol. 42, no. 2, pp. 334-346, Abr. 2009. DOI: 10.1016/j.jbi.2008.09.003
[26]C. Weng, P. Kahn, y J. Hripcsak, "Normalization of Clinical Data: Challenges and Solutions", *Yearb. Med. Inform.*, vol. 8, no. 1, pp. 147-152, 2013. DOI: 10.1055/s-0038-1638876
[27]R. Tsopra, A. Venot, y C. Duclos, "Towards Evidence-Based CDSSs Age-Adapted: A Survey and Analysis", *Artif. Intell. Med.*, vol. 62, no. 1, pp. 1-11, Sep. 2014. DOI: 10.1016/j.artmed.2014.06.003
[28]Organización Panamericana de la Salud, "Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud (CIE-10)", OPS, Washington, D.C., 10a rev., 2018.
[29]R. S. Evans, "Electronic Health Records: Then, Now, and in the Future", *Yearb. Med. Inform.*, vol. 25, no. S01, pp. S48-S61, 2016. DOI: 10.15265/IYS-2016-s006.
[30]D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, “What can natural language processing do for clinical decision support?”, *J. Biomed. Inform.*, vol. 42, no. 5, pp. 760–772, Oct. 2009, doi: 10.1016/j.jbi.2009.08.007.
[31]V. Bali, J. Weaver, V. Turzhitsky, J. Schelfhout, M. L. Paudel, E. Hulbert, J. Peterson-Brandt, A.-M. Guerra Currie, and D. Bakka, “Development of a natural language processing algorithm to detect chronic cough in electronic health records”, *BMC Pulm. Med.*, vol. 22, art. no. 256, 2022, doi: 10.1186/s12890-022-02035-6.
[32]Y. Peng, S. Yan, y Z. Lu, "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets", in *Proc. 18th BioNLP Workshop Shared Task*, Florencia, Italia, 2019, pp. 58-65. DOI: 10.18653/v1/W19-5006.