Design and Feasibility Evaluation of an R Pipeline for Automated Clinical Data Extraction from Informal Records

Beccar Varela, Luisa[/0009-0001-9939-6660]
Facultad de Medicina, Universidad de Buenos Aires, Argentina
luisabeccarvarela@gmail.com

Abstract. Retrospective clinical research in Argentine hospitals faces a critical challenge: the fragmentation of primary clinical records. In many healthcare institutions, inpatient information is documented not only in official paper-based records but also in Google Drive files. Although these documents facilitate physicians' daily workflow, they do not comply with security and confidentiality standards and present significant limitations for secondary data use. This study presents the design, feasibility assessment, and validation results of an R-based pipeline for the automated extraction and structuring of clinical data from unstructured text in a local computing environment. The tool supports patient identification for a retrospective clinical study on hospital-acquired anemia. Evaluated on 446 clinical records from three test files, the system achieved extraction rates ranging from 71.7% to 100%, depending on the variable, processed the entire corpus in 2 minutes and 17 seconds, and reached a Cohen's kappa coefficient of 0.71 (substantial agreement, $p < 0.001$) compared with manual annotation. The findings highlight implications for primary data quality, information security, and the limitations of a rule-based approach.

Keywords: Clinical Data Extraction, Natural Language Processing, Informal Medical Records, R (Programming Language), Feasibility Studies

Diseño y Evaluación de Factibilidad de un Pipeline en R para la Extracción Automatizada de Datos Clínicos desde Registros Informales

Resumen. La investigación clínica retrospectiva en hospitales argentinos enfrenta una barrera crítica: la fragmentación de los registros primarios asistenciales. En muchas instituciones de salud, la información de los pacientes internados se documenta, además de en los registros oficiales de papel, en archivos en la nube. Si bien facilitan el trabajo diario de los médicos, estos registros no cumplen con normas de seguridad y confidencialidad, y tienen severas limitaciones para una explotación secundaria. Este trabajo presenta el diseño, la evaluación de factibilidad y los resultados de la validación de un pipeline en R para la extracción y estructuración automatizada de datos clínicos desde texto no estructurado en un entorno local, para facilitar la selección de pacientes en un estudio clínico retrospectivo de anemia adquirida en internación. Evaluado sobre 446 registros, el sistema logró tasas de extracción de entre 71,7% y 100% según la variable, procesó el corpus en 2 minutos 17 segundos y alcanzó un coeficiente Kappa de 0,71 (concordancia considerable, $p<0,001$) respecto de la anotación manual. Se discuten las implicancias respecto a la calidad del dato

primario, la seguridad de la información y las limitaciones del enfoque basado en reglas.

Palabras clave: Extracción de Datos Clínicos, Procesamiento de Lenguaje Natural, Registros Médicos Informales, R (Lenguaje de Programación), Estudios de Factibilidad

1. Introducción

Como en muchas regiones latinoamericanas, en los hospitales la Ciudad de Buenos Aires existen sistemas oficiales de gestión de la información clínica; sin embargo, en la práctica asistencial persiste el uso de documentación paralela informal en la nube, de acceso fácil desde el smartphone para consulta rápida en la cama del paciente y seguimiento fuera del hospital (Luna et al., 2018). Esta dualidad refleja una brecha entre los sistemas institucionales y las necesidades operativas reales de los equipos de salud (Farías et al., 2023).

Desde una perspectiva de informática en salud, el almacenamiento de datos clínicos y sensibles en plataformas de propósito general, sin controles formales de acceso ni cifrado específico, no cumple con los estándares establecidos por la Ley 25.326 de Protección de Datos Personales de Argentina ni con las recomendaciones internacionales en la materia (Comité Jurídico Interamericano, 2021).

En este contexto, ante un pedido de colaboración para agilizar la selección de casos para una investigación clínica retrospectiva sobre anemia adquirida durante la internación, se optó por desarrollar una solución que no comprometiera la seguridad de los datos y que fuera implementable con recursos computacionales y técnicos limitados. Aunque ya se han desarrollado enfoques más complejos basados en técnicas avanzadas de NLP en español para la extracción automatizada de datos desde historias clínicas electrónicas (Petri, 2025), en el presente trabajo, dadas las limitaciones técnicas y de infraestructura disponibles, se optó por métodos simples de extracción basados en reglas y expresiones regulares, ejecutado en un entorno local, sin conexión a servicios externos.

El objetivo de este trabajo es presentar el diseño de un pipeline automatizado en R que busca en un corpus de pacientes aquellos elegibles para el estudio clínico: que sean mayores de 18 años de edad, sin antecedentes de embarazo, traumatismo ni hemorragia y que no hayan presentado anemia al comienzo de su internación. Además se evalúa su factibilidad técnica mediante la comparación contra una anotación manual como estándar de referencia.

2. Metodología

El pipeline fue desarrollado mediante el software estadístico R.

2.1 Fuente de datos

Los documentos fuente son archivos .docx que contienen tablas completadas por el equipo médico, cada una correspondiendo a un paciente distinto. Cada tabla incluye en texto libre datos identificatorios y clínicos, algunas veces precedidos por etiquetas

como “DOC/PAS/DNI:”, “HC:”, “FI:”, “FICM, “Laboratorio:”, y otros sin: el número de cama, nombre, edad. La heterogeneidad en la forma de registro de estas variables constituye el principal desafío de procesamiento.

2.2 Conversión y segmentación

El pipeline comienza con la transformación de los archivos .docx a archivos de texto plano (.txt), preservando la estructura línea a línea. Luego, cada .txt se segmenta en registros individuales mediante detección del inicio de cada tabla de paciente: una expresión regular identifica una secuencia de 4 o 5 dígitos seguidos de texto (nombre y edad), con al menos una etiqueta reconocida (DNI / DOC / PAS, HC, FI, FICM) en las tres líneas siguientes. Se excluyen falsos inicios que contengan las palabras "epicrisis" o ".doc". A cada registro identificado se le asigna un identificador único incremental como nombre de archivo.

2.3 Extracción de variables

Las variables de identificación (cama, nombre, edad, DNI/DOC/PAS, número de historia clínica, fechas de internación e ingreso a Clínica Médica) se extraen de las primeras cinco líneas de cada registro mediante patrones alternativos que cubren la variabilidad de formato observada en los documentos fuente. El valor de hemoglobina inicial se localiza en la sección "Laboratorios:" tomando la segunda posición del formato DD/MM: valor1/valor2/..., validada en rango clínico (3–24 g/dL). Los criterios de exclusión clínica (embarazo, hemorragia, traumatismo) se detectan mediante patrones con lookbehinds negativos para evitar falsos positivos en menciones negadas. Todas las fechas se conservan sin parseo para preservar la integridad del dato original.

2.4 Criterios de clasificación y manejo del sexo ausente

Cada paciente recibe una clasificación en la columna “comentario” según una lógica de prioridad ordenada: edad menor a 18 años, criterios clínicos (embarazo, hemorragia, traumatismo), anemia basal (Hb < 12 g/dL), ausencia del valor de hemoglobina, y finalmente evaluación por sexo cuando Hb se encuentra entre 12 y 13 g/dL (OMS, 2011). Los pacientes que superan todos los filtros reciben la clasificación "sigue".

Dado que el sexo no figura en los documentos fuente, los casos con "evaluar sexo" se resuelven mediante un módulo interactivo que presenta el nombre del paciente y solicita su determinación manual.

El producto final es un archivo Excel con los datos identificatorios, el comentario y una columna “decisión” donde se asigna CONTINUAR a quienes presentaron comentario "sigue" (candidatos para el estudio clínico) y EXCLUIR a todos los demás casos.

2.5 Diseño del estudio de evaluación

El pipeline se evaluó sobre tres archivos .docx de prueba (~9,31 MB en total), proporcionados por la investigadora responsable del estudio clínico en un pendrive, sin transferencia a redes externas. La anotación manual fue realizada de forma ciega

por la desarrolladora del pipeline sobre los mismos tres archivos, registrando las mismas variables que extrae el sistema.

Se emparejaron los registros deduplicados (434 del pipeline y 435 del manual) mediante left_join exacto y stringdist, logrando aparear 431 pares: 299 por coincidencia exacta y 132 por similitud de nombre (≤2 caracteres de diferencia).

Las métricas de evaluación fueron: tasa de extracción por variable (porcentaje de registros con valor extraído sobre el total), porcentaje de concordancia por variable entre pipeline y anotación manual (registros iguales / total de registros en el gold standard); sensibilidad (S), especificidad (E), valor predictivo positivo (VPP) y valor predictivo negativo (VPN) de la columna decisión, tomando CONTINUAR como condición positiva y la anotación manual como gold standard, y el coeficiente Kappa de Cohen no ponderado para la variable decisión, preferido sobre la concordancia simple dado el desbalance entre categorías. La eficiencia se evaluó comparando el ratio de pacientes elegibles identificados por minuto entre el pipeline y la anotación manual.

3. Resultados

3.1 Procesamiento del corpus de prueba

El pipeline procesó los tres archivos y detectó 445 registros (312 del archivo A, 10 del archivo B, 123 del archivo C). Se identificaron 13 registros duplicados y 4 triplicados. Tras la deduplicación se obtuvieron 434 registros únicos. El módulo interactivo de determinación de sexo fue activado para 45 registros. La corrida completa tomó 2 minutos y 17 segundos.
La anotación manual de los mismos tres archivos arrojó 446 registros (313 del A, 10 del B, 123 del C), con 13 duplicados y 4 triplicados, resultando en 435 registros únicos. La anotación completa tomó aproximadamente 8 horas.

3.2 Tasas de extracción por variable

El pipeline alcanzó el 100% en las variables de segmentación (cama, nombre, edad) y mostró tasas superiores al 96% en DNI, hemoglobina inicial y fecha de ingreso a Clínica Médica. El número de historia clínica registró la tasa más baja (71,7%) en ambas modalidades, lo que refleja inconsistencias en el llenado manual de los documentos fuente.

3.3 Decisión de elegibilidad

Sobre los 445 registros, el pipeline clasificó 131 (29,4%) como CONTINUAR y 314 (70,6%) como EXCLUIR. La anotación manual clasificó 182 (40,8%) como CONTINUAR y 264 (59,2%) como EXCLUIR. La diferencia en la distribución de causas de exclusión se analiza en la sección 3.5.

3.4 Desempeño diagnóstico de la decisión de elegibilidad

Al realizar la tabla de contingencia para la variable “decisión” del pipeline vs la anotación manual (Gold Standard) se obtuvieron las métricas : sensibilidad = 69,4%; especificidad = 98,8%; VPP = 97,7%; VPN = 81,9%. Los falsos positivos fueron escasos (n=3; 0,7%), lo que indica que el pipeline raramente incluye pacientes que

deberían ser excluidos. La principal limitación identificada son los falsos negativos (n=55), que representan pacientes elegibles no capturados en la primera pasada automatizada.

3.5 Concordancia por variable y Kappa de Cohen

Las variables de segmentación (cama, nombre, edad) alcanzaron el 99,1%; DNI y número de historia clínica, el 98,6%; las variables de fecha, entre 96,6% y 96,8%. La hemoglobina inicial mostró 74,0% de concordancia y la variable *comentario* el menor valor (69,7%), debido a las diferencias en la detección de criterios clínicos de exclusión descritas en la sección 2.3.

Dado el desbalance entre categorías —la clase EXCLUIR es mayoritaria tanto en el pipeline (70,6%) como en la anotación manual (59,2%)—, se calculó el coeficiente Kappa de Cohen no ponderado para la variable decisión (Newman & Kohn, 2020). Con una concordancia esperada por azar (Pe) de 0,54 y una concordancia observada de 0,87. El Coeficiente Kappa de Cohen no ponderado resultó en 0,71, con una $p < 0,001$. Según la escala de Landis & Koch (1977), este valor corresponde a una concordancia considerable, e indica que el pipeline y el anotador manual coinciden en la decisión de elegibilidad un 71% más de lo que ocurriría por clasificación aleatoria. La discordancia observada es asimétrica: la mayor parte corresponde a falsos negativos (FN=55).

3.6 Eficiencia de detección

El pipeline identificó 128 candidatos elegibles (CONTINUAR, post-deduplicación) en 2 minutos y 17 segundos, equivalente a un ratio de aproximadamente 56 registros por minuto. La anotación manual identificó 180 pacientes elegibles en 8 horas (480 minutos), con un ratio de 0,375 pacientes elegibles por minuto (1 paciente cada 2 minutos y medio). Aplicando el Kappa (κ=0,71) como estimación de la fracción de verdaderos positivos capturados, el pipeline recuperaría aproximadamente 91 pacientes elegibles reales en sus 2 minutos y 17 segundos, con un ratio ajustado de ~36 pacientes elegibles correctamente identificados por minuto, frente a 0,375 de la anotación manual: una diferencia de aproximadamente 97 veces.

4. Discusión

Los resultados demuestran que el pipeline es capaz de procesar documentos clínicos informales y producir una tabla estructurada con alta tasa de extracción en las variables centrales del estudio, con un nivel de concordancia considerable respecto de la revisión clínica experta (κ=0,71). La reducción del tiempo de procesamiento —de 8 horas a poco más de 2 minutos sobre el mismo corpus— valida la factibilidad técnica del enfoque para el problema de preselección de cohorte planteado.

La principal desventaja observada en los resultados es la sensibilidad moderada del pipeline (69,4%), que derivada de 55 falsos negativos.

El presente trabajo no propone continuar ni validar el uso de plataformas en la nube de propósito general para el almacenamiento de datos clínicos, sino documentar esa práctica como una realidad existente que aún no ha sido resuelta en el sistema de salud local. Los datos de salud son datos sensibles: su almacenamiento en sistemas sin

controles formales de acceso ni cifrado específico vulnera la normativa vigente y compromete los derechos de los pacientes.

La implementación de sistemas formales de registro clínico con resguardos adecuados es una obligación institucional cuyos beneficios trascienden la atención diaria: bajo condiciones éticas y técnicas apropiadas, esos mismos datos pueden convertirse en insumo para la investigación clínica retrospectiva y el avance del conocimiento médico

Limitaciones técnicas. La dependencia de expresiones regulares genera fragilidad ante cambios de formato en los documentos fuente; la ausencia de un esquema fijo obliga a ajustar los patrones cada vez que se modifica la estructura de las tablas. El enfoque basado en reglas también tiene dificultades para discernir contextos semánticos complejos (por ejemplo, distinguir antecedentes médicos de condiciones activas). El estudio se realizó sobre tres archivos de prueba proporcionados por la investigadora; la evaluación sobre el corpus completo (>8.000 registros) constituye el paso de validación siguiente.

5. Conclusiones

El pipeline permite convertir, segmentar, extraer variables críticas y clasificar pacientes de forma estructurada y reproducible. Los resultados sobre tres archivos de prueba (446 registros) demuestran la viabilidad técnica del enfoque: alta especificidad (98,8%) en la identificación de candidatos elegibles, concordancia considerable con el gold standard clínico (κ=0,71) y una reducción de la carga de trabajo de aproximadamente 97 veces respecto de la anotación manual. La baja sensibilidad (69,4%) implica que el pipeline descarta pacientes elegibles por lo que podrían revisarse casos excluidos y obtener más candidatos para el estudio clínico. El trabajo demuestra que metodologías de bajo costo computacional, centradas en la soberanía local del dato, son viables en contextos de infraestructura limitada.

Declaraciones

Origen del proyecto y uso de datos clínicos. El pipeline fue desarrollado a solicitud de la investigadora responsable del estudio clínico, quien proporcionó los tres archivos de muestra en un pendrive, evitando el uso de redes externas o servicios de nube para el traslado de los datos primarios. Se describe exclusivamente el diseño e implementación del pipeline de extracción; no se publican datos de pacientes. En ningún momento se transfirieron archivos con datos de pacientes a servidores externos ni a herramientas de inteligencia artificial de acceso público.

Uso de inteligencia artificial. Durante la preparación de este trabajo, la autora utilizó herramientas de asistencia basadas en inteligencia artificial (Claude, Anthropic, 2026; Gemini, Google, 2026; Comet, Perplexity AI, 2026) como apoyo en el desarrollo iterativo del pipeline en R y como asistente de redacción. En todos los casos, la autora definió los objetivos, tomó todas las decisiones técnicas y clínicas, revisó y validó cada resultado, y asume plena responsabilidad por el contenido de esta publicación. Los archivos con datos de pacientes no fueron compartidos con ninguna herramienta de IA en ninguna etapa del proceso.

Referencias

Farías, M. A., Badino, M., Báscolo, E., y García Saisó, S. (2023). La transformación digital como estrategia para fortalecer las funciones esenciales de salud pública en las Américas. Revista Panamericana de Salud Pública, 47, e150. https://doi.org/10.26633/RPSP.2023.150

Ministerio de Salud GCBA. (2010). Resolución 123/GCABA/SSASS/10: Sistema de Gestión Hospitalario SIGEHOS. Boletín Oficial de la Ciudad de Buenos Aires, 3514.

Comité Jurídico Interamericano. (2021). Principios actualizados sobre la privacidad y la protección de datos personales, con anotaciones. Organización de los Estados Americanos. https://www.oas.org/es/sla/cji/docs/Publicacion_Proteccion_Datos_Personales_Principios_Actualizados_2021.pdf

Petri, Javier. (2025). Extracción de información de historias clínicas electrónicas escritas en español para realizar inteligencia epidémica. (Tesis de Grado. Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales.). Recuperado de https://hdl.handle.net/20.500.12110/seminario_nCOM000838_Petri

World Health Organization. (2011). Haemoglobin concentrations for the diagnosis of anemia and assessment of severity. WHO/NMH/NHD/MNM/11.1. https://www.who.int/publications/i/item/WHO-NMH-NHD-MNM-11.1

Luna, D - Otero, C - Plazotta, F - Campos, F (2018). Sistemas de información para la Salud - 1a ed. - Buenos Aires: Sociedad Italiana de Beneficencia en Buenos Aires. https://www.hospitalitaliano.org.ar/hiba/es/news/sistemas-de-informacion-para-la-salud

Newman, T. B., & Kohn, M. A. (2020). Evidence-based diagnosis: An introduction to clinical epidemiology (2nd ed.). Cambridge University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://pubmed.ncbi.nlm.nih.gov/843571/