Generación de un dataset de problemas relacionados con hipertensión arterial en el sistema electrónico de información en salud de la Ciudad Autónoma de Buenos Aires

Camila Agostina Ebensrtejin, 0009-0005-0780-2545, Ministerio de Salud de la Ciudad Autónoma de Buenos Aires, GOGIES, cebensrtejin@buenosaires.gob.ar.
Carolina Mengoni Goñalons, 0000-0001-7052-8642, Ministerio de Salud de la Ciudad Autónoma de Buenos Aires, GOGIES, cmengoni@buenosaires.gob.ar.
Florencia Faretta, Ministerio de Salud de la Ciudad Autónoma de Buenos Aires, GOGIES, ffaretta@buenosaires.gob.ar.

Resumen. La hipertensión arterial (HTA) es una de las enfermedades crónicas no transmisibles y el factor de riesgo cardiovascular más prevalente en la población mundial. En el Ministerio de Salud del Gobierno de la Ciudad Autónoma de Buenos Aires se está llevando adelante un proceso de mejora continua para la identificación, el seguimiento y el control de pacientes con HTA en los efectores públicos, con el fin de optimizar la atención y gestionar de manera más eficiente los recursos. Dentro del Ministerio, la Gerencia Operativa de Gestión de Información y Estadística es la encargada de identificar pacientes a partir de los registros sanitarios en las historias clínicas electrónicas del sistema de información utilizado en los efectores públicos, y de construir indicadores de seguimiento y control. El proceso de identificación comenzó con un conjunto de reglas sencillas, con prioridad en la sensibilidad por sobre la especificidad. Con el tiempo se aplicaron mejoras para aumentar el valor predictivo positivo y la especificidad. El objetivo de este trabajo es describir los avances en la generación de nuevas reglas para la identificación de personas con HTA, y el establecimiento de un circuito de monitoreo y evaluación periódica.

La nueva versión del proceso incorpora motivos de consulta relacionados con HTA que eran inicialmente excluidos, y además incluye una actualización periódica para validar nuevas expresiones ingresadas en el sistema de información. Se usaron herramientas de procesamiento de texto y técnicas cuantitativas para evaluar resultados. Se eliminaron falsos positivos e incorporaron términos asociados a la enfermedad.

Palabras clave: hipertensión arterial, sistema de información en salud, procesamiento de texto, motivos de consulta.

Generation of a Dataset of Problems Related to Hypertension (HTN) in the Electronic Health Information System of Buenos Aires City

Abstract. Arterial hypertension (HTN) is one of the most prevalent non-communicable chronic diseases and the most common cardiovascular risk factor in the global population. The Ministry of Health of the Government of the City of Buenos Aires is carrying out a continuous improvement process for the identification, follow-up, and control of patients with HTN in public healthcare facilities, with the aim of optimizing care and managing resources more efficiently. Within the Ministry, the Operational Management of Information and Statistics is responsible for identifying patients based on electronic health records (EHR) of the information system used in public facilities, and for building monitoring and control indicators. The identification process began with a set of simple rules, prioritizing sensitivity over specificity. Over time, improvements were applied to increase the positive predictive value and specificity.

The aim of this work is to describe the progress made in generating new rules for the identification of people with HTN, and the establishment of a periodic monitoring and evaluation circuit. The new version of the process incorporates reasons for referral related to HTN that were initially excluded, and also includes a periodic update to validate new expressions entered in the information system.

Text processing tools and quantitative techniques were used to evaluate the results. False positives were eliminated and terms associated with the disease were incorporated.

Keywords: hypertension, health information system, text processing, reason for referral.

1. Introducción

Según la Organización Mundial de la Salud, la hipertensión arterial es una de las enfermedades crónicas no transmisibles más prevalentes a nivel mundial y su gestión efectiva es clave para la prevención de complicaciones cardiovasculares y otras condiciones clínicas asociadas.

En Argentina, el aumento de la prevalencia de hipertensión arterial, reflejado en los datos del Ministerio de Salud, destaca la necesidad de mejorar los procesos de identificación y seguimiento de los pacientes con diagnóstico de la misma.

El sistema de información en salud (SIS) del Gobierno de la Ciudad de Buenos Aires (GCBA), que comenzó a digitalizarse en 2016, incluye un módulo de Historia Clínica Electrónica (HCE) y es la plataforma utilizada para gestionar la información de los pacientes. A medida que fue avanzando la implementación y el uso de la HCE, el volumen de datos sanitarios aumentó considerablemente, así como la diversidad en los modos de registro. En este contexto, se generó la necesidad de identificar y caracterizar líneas de salud priorizadas por el Ministerio a partir de los registros en el SIS, una de ellas la HTA. El paso inicial consistió en la identificación de personas con esta condición a partir de reglas sencillas, que incluyeron los motivos de consulta (problemas) registrados en sus HCE y medicación entregada. El uso de la HCE en los servicios sanitarios es altamente prevalente, por lo que el mayor volumen de datos se obtiene a partir de esta fuente de información. Asimismo, considerando que el campo 'problema' es de carácter semiestructurado, esto trae como consecuencia una diversidad creciente en la terminología utilizada al registrar un problema. El objetivo principal de este trabajo ha sido mejorar la identificación de términos referidos a hipertensión arterial en el SIS de los efectores públicos de salud del GCBA y optimizar el uso de los recursos para el seguimiento de pacientes.

El artículo pone de manifiesto los avances alcanzados en la creación de un dataset de problemas relacionados con HTA, utilizando un enfoque basado en el procesamiento de texto mediante expresiones regulares, complementado por un proceso manual para validar términos y excluir falsos positivos. Presentando un proceso de trabajo para mejorar la detección de términos referidos a hipertensión arterial en el SIS de los efectores públicos de salud del GCBA.

2. Metodología y Resultados

El proceso de identificación de problemas relacionados con HTA fue orientado al objetivo final de identificación de personas con HTA. Se utilizaron herramientas de proceso de texto en torno a un conjunto de reglas: una combinación de expresiones regulares y operaciones manuales, tanto de inclusión como de exclusión. La fuente de datos es el listado de expresiones únicas de problemas históricos registrados en la HCE, un tesauro de motivos de consulta.

Desafíos Iniciales y Generación del Dataset

Uno de los principales desafíos fue la alta variabilidad terminológica para referirse a una misma condición clínica. Esta heterogeneidad complicaba la agregación de datos, afectando la posibilidad de realizar un seguimiento longitudinal confiable y análisis poblacionales robustos. Diferentes profesionales registran el mismo concepto de manera diversa, lo que fragmenta la información y disminuye la capacidad de generar indicadores precisos. Para abordar esta dificultad, se diseñó un conjunto de expresiones regulares que permiten detectar términos específicos y patrones de texto incluidos en los registros. Se establecieron:

- **Expresiones de inclusión**, orientadas a captar distintas formas de mención de la condición ("hipertensión", "presión alta", "HTA", entre otras).

- **Expresiones de exclusión**, aplicadas a casos que, aunque contenían los términos clave, no se referían directamente al diagnóstico de HTA. Esto incluyó:

    - Otros diagnósticos (por ejemplo, "hipertensión ocular", "retinopatía hipertensiva"),

    - Referidas a antecedentes y evaluaciones preliminares (como "sospecha de hipertensión", "descartar HTA", "familiar con HTA"),

    - Situaciones o actividades (por ejemplo: "charla sobre hipertensión", "taller", "post vacunación").

Luego de aplicar estas reglas, se obtuvieron dos productos:

1. Un listado inicial de problemas referidos a HTA con fines de identificación de personas.

2. Un listado de términos referidos a HTA pero que resultó positivo para algún criterio de exclusión.

Esta limpieza sistemática del dataset permitió aumentar la precisión del proceso de detección, eliminando falsos positivos y optimizando el valor predictivo positivo del conjunto resultante.

Clasificación del dataset y generación del producto final

Para garantizar la precisión de los problemas detectados como indicativos de un diagnóstico de HTA y minimizar la incidencia de falsos positivos, se llevó a cabo una lectura detallada de ambos listados con el fin de clasificarlos como verdadero/falso positivos/negativos, según correspondiera. Este análisis derivó en un ajuste de las expresiones regulares, tanto de inclusión como de exclusión. En particular, se incorporaron términos de exclusión a partir de la detección de falsos positivos (problemas que contenían términos de inclusión, pero no referían a la condición de interés o no cumplían con un fin diagnóstico). Los nuevos términos de exclusión incluyeron otras condiciones (“intracraneal”, “encefalopatía”, “gastropatía”, “renovascular”, “cardíaca”, “pulmonar”), situaciones agudas (“post vacunación”, “inducida”), prácticas en torno a la condición de interés (“taller”, “charla”) o referencias a la condición (“miedo”, “temor”).

Luego de la clasificación manual y el ajuste de las expresiones regulares, comenzamos con una base de 1161 problemas, de los cuales se identificaron 858 como verdaderos positivos y 303 como falsos positivos. El dataset actualizado ahora cuenta con 881 problemas identificatorios. Además, se generó un dataset separado de 701 exclusiones, que agrupa problemas no asociados con HTA, sino relacionados con condiciones médicas que incluyen el término "hipertensión", pero no hacen alusión a la hipertensión arterial propiamente dicha. Este conjunto ayuda a depurar continuamente el proceso de identificación, garantizando la precisión del dataset actualizado.

Tabla 1: Resultados del Proceso de Clasificación y Depuración de Datos

Descripción	Cantidad
Total de problemas en la base inicial	1161 (100 %)
Verdaderos Positivos (VP)	858 (73.9%)
Falsos Positivos (FP)	303 (26.1%)
Dataset actualizado	881
Dataset de exclusiones	701

Nota: El conjunto de exclusiones incluye problemas que incorporan el término "hipertensión" pero no están específicamente relacionados con hipertensión arterial (HTA). La separación de este conjunto mejora la precisión del dataset actualizado.

Proceso automático de actualización y revisión

El tesauro de problemas utilizado en la HCE es dinámico y creciente. Esto significa que el listado de problemas HTA no es estático y debe ser actualizado de manera periódica. Teniendo esto en cuenta, diseñamos un proceso de revisión para mejorar la precisión y exhaustividad del proceso de detección. El mismo se ejecuta cada seis meses y aplica el conjunto de reglas del producto final al tesauro actual de la HCE. En un segundo paso, compara los listados resultantes con aquellos que forman parte del producto final. En caso de encontrar alguna expresión única nueva, se genera una alerta, a partir de la cual un analista debe validar ese problema de manera manual, ya sea a incluir o excluir. De ser necesario, se ajustan las expresiones regulares. Si esto no fuera posible, se incluirá o excluirá ese problema de manera particular. Así, el producto final se actualiza y por lo tanto constituye un producto dinámico. Esto, asegura la depuración continua. Categorizando nuevos términos según su relevancia, se proporcionan métricas que reflejan el grado de confiabilidad del dataset actualizado.

Este proceso no solo reduce los falsos positivos, sino que también optimiza el valor predictivo positivo, permitiendo un monitoreo más preciso y eficiente de los pacientes con HTA.

3. Discusión y conclusión

Este proceso de identificación y validación continua permite realizar evaluaciones periódicas de su efectividad. A medida que se actualizan los datos, se realizan análisis estadísticos para determinar la precisión de la detección de términos relacionados con HTA, permitiendo ajustar las expresiones regulares y las reglas manuales para mejorar los resultados. Entre las limitaciones actuales, encontramos la dependencia de la validación manual para evitar falsos positivos, lo que implica una carga de trabajo que podría ser optimizada en futuras etapas con herramientas de NLP. El dataset con los problemas validados proporciona una fuente confiable de datos que facilita el análisis para el equipo de Gestión de Información y Estadísticas de Salud, el cual genera información que puede ser utilizada para la gestión de recursos y la planificación de políticas de salud pública relacionadas con HTA. Con el enfoque dinámico de este proceso, se espera lograr una mejora continua en la precisión de la identificación de pacientes con HTA, lo que redundará en un mejor seguimiento de la enfermedad y en la optimización de los recursos de salud pública.

El proceso de identificación y validación de problemas relacionados con HTA en el SIS del GCBA ha demostrado ser un enfoque eficaz para mejorar la gestión de los pacientes diagnosticados con esta condición. La implementación de expresiones regulares junto con un sistema dinámico de revisión y actualización permite mantener la relevancia y precisión del dataset, lo que contribuye a un mejor uso de los recursos y una atención más eficiente a los pacientes con HTA. Este proceso, en constante evolución, proporciona una base sólida para la toma de decisiones en la gestión de la salud pública en la Ciudad Autónoma de Buenos Aires.

4. Referencias

Loscalzo, J., Fauci, A., Kasper, D., Hauser, S., Longo, D., y Jameson, J. (2022). Harrison. Principios de Medicina Interna, 21e. McGraw-Hill Education.
Shortliffe, E. H., & Cimino, J. J. (Eds.). (2014). Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer.

Ministerio de Salud de la Nación. (2016). Manual para el cuidado integral de personas adultas en el primer nivel de atención: Capítulo 7: Hipertensión arterial (HTA).

Ministerio de Salud de la Nación. (2024). Guía de Práctica Clínica Nacional sobre Prevención, Diagnóstico y Tratamiento de la Hipertensión Arterial (HTA). https://www.argentina.gob.ar/sites/default/files/salud-guia-practica-clinica-nacional-hta-2024.pdf

Rajkomar, A., Dean, J., y Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358. https://doi.org/10.1056/NEJMra1814259

Sociedad Argentina de Hipertensión Arterial, Sociedad Argentina de Cardiología, y Federación Argentina de Cardiología. (2018). Consenso Argentino de Hipertensión Arterial. https://www.saha.org.ar/uploads/pdf/CONSENSO-SAHA-1.pdf