

COMUNICACIÓN BREVE

Innovación en la recolección de datos en los registros de cáncer de base poblacional de Colombia: implementación y desafíos

Innovation in data collection for population-based cancer registries in Colombia: implementation and challenges

Constanza Pardo¹

¹ Grupo Vigilancia Epidemiológica del Cáncer, Instituto Nacional de Cancerología, Bogotá, D. C., Colombia.

Fecha de sometimiento: 25/07/2025
Fecha de aceptación: 04/12/2025
Disponible en internet: 02/03/2026

Abstract

Population-based cancer registries (PBCR) enable the measurement of cancer incidence and survival rates. The challenge is to reduce data collection times, which affects operating costs. Therefore, the *Instituto Nacional de Cancerología* (INC; National Cancer Institute) seeks to implement information technologies in PBCR throughout Colombia. The Cali PBCR, using natural language processing, optimized information extraction from unstructured pathology formats and reduced data collection time. This methodology is being transferred to the PBCR in Barranquilla and Neiva. Advances include automated data extraction, relational database creation, and real-time data visualization. The collaboration among the INC, the Cali PBCR, and the regional node of the Global Initiative for Cancer Registry Development will enable capacity pooling and strengthen cancer surveillance in Colombia and Latin America.

Keywords: information systems; machine learning; natural language processing; technological innovation; neoplasms; population-based cancer registries; public health surveillance; Colombia.

Conflictos de interés

Los autores declaran no presentar conflictos de interés.

Citación

Pardo C. Innovación en la recolección de datos en los registros de cáncer de base poblacional de Colombia: implementación y desafíos. Rev Col Cancerol. 2026;30(1):90-5.
<https://doi.org/10.35509/01239015.1136>

Autor de correspondencia

Elda Constanza Pardo-Ramos
Grupo Vigilancia Epidemiológica del Cáncer,
Instituto Nacional de Cancerología, Bogotá, D. C.,
Colombia.

Correo electrónico:

cpardo@cancer.gov.co

Resumen

Los registros de cáncer de base poblacional (RCBP) permiten medir la incidencia y la supervivencia de cáncer. El desafío es reducir los tiempos de recolección de datos con impacto en los costos de operación, por lo que, el Instituto Nacional de Cancerología (INC) busca implementar tecnologías de información en los RCBP de Colombia. El RCBP Cali, con procesamiento de lenguaje natural, optimizó la extracción de información a partir de formatos no estructurados de patología y redujo el tiempo de recolección de datos. Esta metodología se está transfiriendo a los RCBP de Barranquilla y Neiva.

Entre los avances están: automatización en extracción de datos, creación de bases de datos relacionales y visualización de datos en tiempo real. La colaboración entre INC, RCBP Cali y nodo regional de la Iniciativa Mundial para el Desarrollo de Registros de Cáncer permitirá sumar capacidades y fortalecer la vigilancia del cáncer en Colombia y América Latina.

Palabras clave: sistemas de información; aprendizaje automático; procesamiento de lenguaje natural; innovación tecnológica; neoplasias; registros poblacionales de cáncer; vigilancia en salud pública; Colombia.

Introducción

La vigilancia del cáncer define a los registros de cáncer de base poblacional (RCBP) como el estándar de oro para medir la incidencia y supervivencia de cáncer. Esta información es la base para la evaluación de los planes nacionales de control del cáncer, el diseño de programas de prevención y detección temprana, y la orientación de la investigación en cáncer (1).

Colombia actualmente tiene siete RCBP distribuidos en distintas zonas geográficas del país, con cobertura del 25 % de la población nacional. La información de los RCBP es insumo para las estimaciones de incidencia y mortalidad por cáncer en Colombia (2) y para las estimaciones mundiales de la Agencia Internacional para la Investigación sobre el Cáncer (3). Cuatro RCBP son reconocidos internacionalmente por su alta calidad: Bucaramanga, Cali, Manizales y Pasto; y tres están en proceso de mejora: Barranquilla, Medellín y Neiva (4). El Instituto Nacional de Cancerología (INC) de Colombia coordina las acciones de los RCBP a nivel nacional con asistencia técnica, capacitación y financiación parcial.

El crecimiento poblacional, el aumento en la carga del cáncer, el modelo de salud y la organización de la prestación de los servicios han generado un aumento de las instituciones oncológicas (que en Colombia ascienden a 2421 servicios habilitados). Esta dinámica llevó a los RCBP a una “transición tecnológica”, con crecimiento de volumen, velocidad y variabilidad de los datos reportados en las distintas fuentes de información, relacionadas con la atención del cáncer. Actualmente, los procesos son más complejos y lentos, con aumento de distintos tipos

de formatos (estructurados y no estructurados) y mayor tiempo de recolección, desafíos similares que enfrentan otros registros en países de ingresos medios y bajos.

La no integración y estandarización de la información entre los registros de cáncer y las fuentes hace ineficiente la extracción de la información, con un subregistro de casos y, en consecuencia, subestimación del riesgo de cáncer en la población. Los laboratorios de patología, principal fuente de información para los RCBP, generan altos volúmenes de informes de patología en archivos digitales individuales no estructurados y no unificados para el país; y el informe como tal contiene los datos precisos de diagnóstico, con información relevante para un RCBP.

Como solución a esta situación, se han desarrollado varias iniciativas de trabajo que utilizan el procesamiento de lenguaje natural (PLN), con el propósito de mejorar la eficiencia en la recolección de casos en tiempo real (5). Entre algunos desarrollos están los registros centrales de cáncer, que son parte del Programa Nacional de Registros de Cáncer de Estados Unidos, los cuales implementaron una plataforma informática en la nube para la vigilancia del cáncer. En la región de América Latina y el Caribe, actualmente existen diversas iniciativas que usan PLN, sin embargo, no están documentadas. Se destacan dos casos de innovación tecnológica, uno de carácter regional (6) y el otro local (7), que implementaron algoritmos a datos clínicos y en reportes patológicos para un registro de cáncer de base hospitalaria.

El RCBP de Cali, Colombia, con más de 60 años de funcionamiento, hace una década dinamizó el proceso de recolección de la información con el uso de la informática. Su desafío fue mejorar la oportunidad y

disminuir posibles retrasos en el análisis y uso de la información (8). Durante este periodo, logró implementar una metodología para optimizar la extracción de la información de los informes de patología con el uso de PLN y el sistema unificado de lenguaje médico (UMLS, por sus siglas en inglés) (9-10), convirtiendo el lenguaje natural en datos estructurados y clasificarlos en cáncer y no cáncer, con reducción del tiempo recolección de 6 meses a 48 horas (8). Además, dicho RCBP desarrolló las metodologías para el uso de la base de datos relacional y la visualización de datos.

El objetivo de esta comunicación breve es presentar el programa “Optimización de los métodos de extracción en los RCBP de Colombia”, con fases establecidas y la implementación inicial en Neiva y Barranquilla, a partir de la transferencia técnica del RCBP de Cali, lo cual resulta relevante para la vigilancia del cáncer en el país.

Implementación del programa «Optimización de los métodos de extracción en los RCBP de Colombia»

Con la innovación metodológica llevada a cabo por el RCBP de Cali, se inició en el año 2024 la implementación del programa «Optimización de los métodos de extracción para el aseguramiento en la calidad de información en los RCBP de Colombia», dentro del marco del programa «Misiones transformativas para el control integral del cáncer» del INC. El propósito fue realizar la transferencia de capacidades técnicas a los otros RCBP del país, en el que inicialmente se incluyeron los registros de Barranquilla y Neiva. El desarrollo incluyó distintas fases orientadas a mejorar las capacidades y apropiar nuevas metodologías para la extracción inteligente de datos, construcción de bases de datos relacionales y visualización de datos (tabla 1, figura 1).

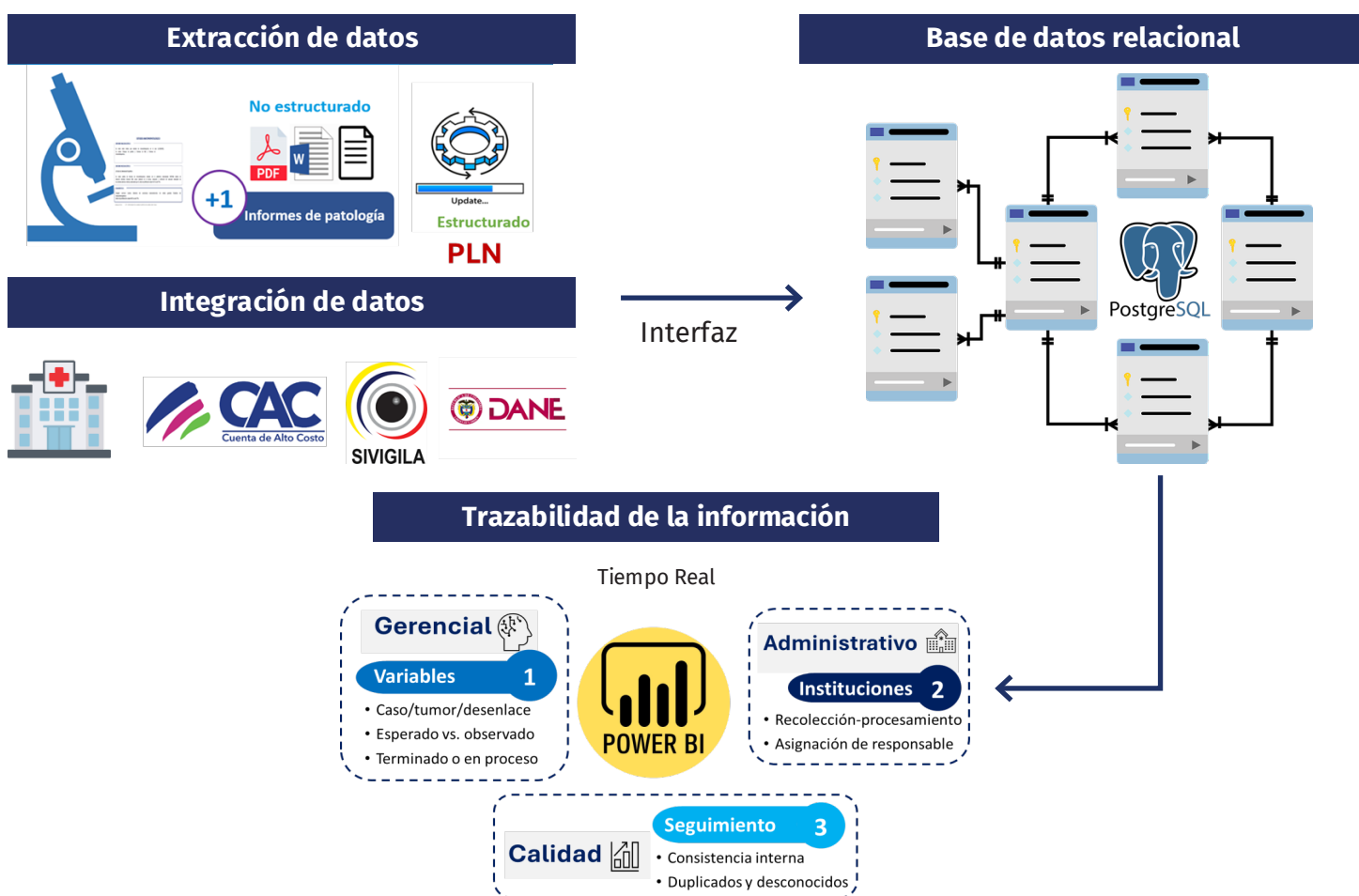
Tabla 1. Fases de implementación para la incorporación de nuevas tecnologías de recolección y visualización de información en los registros de cáncer de base poblacional de Colombia

Fases	Actividades	Herramientas
Censo y caracterización de fuentes y sistemas de información del RCBP	<ul style="list-style-type: none"> Revisión y caracterización de las fuentes de información y de los sistemas de información y notificación Revisión de los procedimientos operativos estándar 	<ul style="list-style-type: none"> Red Especial de Prestadores de Servicios de Salud (REPS) Sistemas de notificación: Sistema Nacional de Vigilancia en Salud Pública (Sivigila) del Instituto Nacional de Salud, enfermedades de alto costo (cáncer) y mortalidad (DANE) Métodos de recolección Tipos de formatos: datos estructurados y datos no estructurados Cuestionario de caracterización en REDCap
Entrenamiento	Talleres virtuales: tres en total, modalidad mixta con conexión sincrónica por tres días, dos horas diarias en cada taller y actividades asincrónicas	<ul style="list-style-type: none"> Identificación de las fuentes de información: tipos y características Implementación del proceso de extracción de datos y creación de la base de datos relacional Visualización de indicadores de seguimiento con <i>Power BI</i>
Proceso automatizado de manejo de la información: extracción de datos	Datos no estructurados: configuración de archivos, estandarización, consolidación y complementación de información faltante	<ul style="list-style-type: none"> Códigos fuente Plantillas Parametrización de variables Lenguaje de programación <i>Python</i> <i>Scripts</i> de extracción
Proceso automatizado de manejo de la información: integración de datos	Creación de la estructura de base de datos relacional: extracción, procesamiento y automatización	<ul style="list-style-type: none"> <i>Scripts</i> Adaptación de diccionarios Sistema de base de datos objeto-relacional de código abierto en <i>PostgreSQL</i>

Fases	Actividades	Herramientas
Visualización de datos y trazabilidad de la información	<ul style="list-style-type: none"> • Uso de herramientas para conectar con la base de datos relacional, transformar y visualizar en tiempo real • Diseño e implementación de tableros dinámicos y paneles visuales 	<ul style="list-style-type: none"> • Access • Plataforma Power BI • Información gerencial, administrativa y de calidad • Indicadores

DANE: Departamento Administrativo Nacional de Estadística (Colombia); RCBP: Registros de cáncer de base poblacional.

Fuente: elaboración propia con base en los documentos soporte bajo el Convenio Interadministrativo 2024-0102, INC - Universidad del Valle.



DANE: Departamento Administrativo Nacional de Estadística (Colombia); PLN: procesamiento del lenguaje natural; Sivigila: Sistema Nacional de Vigilancia en Salud Pública del Instituto Nacional de Salud.

Fuente: elaboración propia con base en los documentos soporte bajo el Convenio Interadministrativo 2024-0102, INC - Universidad del Valle.

Figura 1. Tecnologías de recolección y visualización de datos en implementación de los registros de cáncer de base poblacional de Colombia

Se realizó el censo y la caracterización de las fuentes de información para cada RCBP. Se priorizó a los laboratorios de patología y a las instituciones oncológicas con información no estructurada, con los cuales se inició la implementación del proceso de extracción. Se realizaron tres talleres en modalidad mixta, sincrónicos y asincrónicos para conocer y apropiarse de las metodologías. Los dos RCBP avanzaron en la extracción de información con el desarrollo de *scripts* en el lenguaje de programación *Python*, para leer archivos estructurados y no estructurados (informes de patología), mediante librerías específicas de *UMLS*, lo que permitirá llevar estos datos a la base de datos relacional mediante una interfaz.

En Neiva, el procesamiento de un archivo de datos estructurados en Excel demoró 20 segundos para 2800 registros y entre 30-40 segundos para 1800 archivos no estructurados en Word. En Barranquilla, para un archivo estructurado en Excel, con 6600 registros, el proceso transcurrió en 3 minutos y en uno no estructurado en Word, 40 segundos, donde cada archivo representó a una persona. Las variaciones en los tiempos de procesamiento evidenciaron los retos y avances que surgen al no tener un formato estándar con las variables mínimas.

Actualmente, se implementa la parte funcional de la base de datos relacional en *PostgreSQL*, para el almacenamiento consolidado de la información. En cuanto al desarrollo de la visualización de datos con *Power BI*, se completó el diseño y se está avanzando en la implementación y los ajustes de los paneles visuales funcionales para la trazabilidad, el monitoreo y el control de calidad de la información. La interfaz gráfica de usuario para la base de datos relacional se encuentra en desarrollo, la cual facilitará la interacción del personal del registro en la visualización y actualización de los casos.

Desafíos

Existen muchos desafíos por abordar con los dos registros de cáncer seleccionados, por lo que es necesario continuar con la implementación de los componentes de «clasificador de cáncer» y «calidad» para obtener las métricas de desempeño del modelo de PLN (precisión, sensibilidad y especificidad) y alcanzar la operatividad integral en ambos RCBP. Obtener la interoperabilidad

con otros sistemas nacionales de información en cáncer, como lo son el sistema de seguimiento a las enfermedades de alto costo, entre ellas cáncer, el Sistema Nacional de Vigilancia en Salud Pública (Sivigila) del Instituto Nacional de Salud y la mortalidad nacional, con el fin de mejorar la eficiencia en la recolección, la validación y el análisis de los datos oncológicos. Además, mejorar la oportunidad de la información para los indicadores de incidencia y supervivencia, y gestionar nuevos recursos para la apropiación de las metodologías en los otros RCBP en Colombia.

Esta experiencia de trabajo permitirá, en principio, destacar retos comunes y tal vez, en un futuro, promover la cooperación con otros RCBP de la región, ya sea con entrenamiento o asistencia técnica. El papel del nodo regional en América Latina de la Iniciativa Mundial para el Desarrollo de Registros de Cáncer es importante en la alineación de las distintas iniciativas.

Conclusiones

La colaboración del INC y el Registro Poblacional de Cáncer de Cali ha permitido iniciar la transferencia de capacidades técnicas a dos RCBP del país (Neiva y Barranquilla). La implementación de la automatización del proceso de extracción de la base de datos relacional y del tablero de visualización de datos permitirá, en un corto tiempo, mostrar resultados en la reducción de los tiempos de recolección, la oportunidad de presentación de informes y la mejora de la calidad basada en alertas sobre el comportamiento de los datos.

Reconocimientos

Como autora y líder del programa «Optimización de los métodos de extracción para el aseguramiento en la calidad de información en los RCBP de Colombia», agradezco la contribución y el trabajo de los RCBP y las instituciones a las que están adscritos:

- Integrantes del RCBP de Cali: trabajo realizado mediante el Convenio Interadministrativo 2024-0102, celebrado entre el INC y la Universidad del Valle.

- Integrantes de los RPC de Barranquilla (Universidad del Norte) y de Neiva (Universidad Surcolombiana).
- Equipo técnico y administrativo del INC.

Financiamiento

El programa «Optimización de los métodos de extracción en los RCBP en Colombia» fue financiado por el Ministerio de Ciencia, Tecnología e Innovación, a través del Fondo de Investigación en Salud (H190503003805) y recursos propios asignados al Programa Vigilancia Epidemiológica del Cáncer del INC.

Referencias

1. The Lancet. Cancer registries: the bedrock of global cancer care. *Lancet*. 2025;405(10476):353. [https://doi.org/10.1016/S0140-6736\(25\)00189-8](https://doi.org/10.1016/S0140-6736(25)00189-8)
2. Pardo-Ramos C, Cendales-Duarte R. Estimaciones de incidencia y mortalidad para los cinco principales tipos de cáncer en Colombia, 2017-2021. *Rev Col Cancerol*. 2024;28(4):162-76. <https://doi.org/10.35509/01239015.1061>
3. Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, et al. Global Cancer Observatory: Cancer Today (version 1.1). [internet]. Lyon, Francia: International Agency for Research on Cancer. [citado 2025 jul. 23]. Disponible en: <https://gco.iarc.who.int/today>
4. Navarro E, Caballero H, Cortés A, Arias N, Casas H, de Vries E, et al. Sistema de información de cáncer en Colombia - SICC (Versión 1.0). [internet]. Bogotá, Colombia: Instituto Nacional de Cancerología; 2024. [citado 2025 jun. 21]. Disponible en: <https://www.infocancer.co>
5. Jones D, Alimi T, Pordell P, Tangka F, Blumenthal W, Jones S, et al. Pursuing data modernization in cancer surveillance by developing a cloud-based computing platform: real-time cancer case collection. *JCO Clin Cancer Inform*. 2021;5:24-9. <https://doi.org/10.1200/cci.20.00082>
6. Villena F, Dunstan J. Obtención automática de palabras clave en textos clínicos una aplicación de procesamiento del lenguaje natural a datos masivos de sospecha diagnóstica en Chile. *Rev Med Chile*. 2019;147(10):1229-38. <http://dx.doi.org/10.4067/s0034-98872019001001229>
7. Mendoza-Urbano D, Garcia J, Moreno J, Bravo-Ocaña J, Riascos A, Zambrano A, et al. Automated extraction of information from free text of Spanish oncology pathology reports. *Colomb Med*. 2023;54(1):e2035300. <https://doi.org/10.25100/cm.v54i1.5300>
8. Portilla N-A, Solarte-Pabón O, Bravo L-E. Automatic classification of cancer pathology reports written in Spanish: a machine-learning approach. *Rev Fac Ing*. 2024;33(68):e18080. <https://doi.org/10.19053/01211129.v33.n68.2024.18080>
9. Munzone E, Marra A, Comotto F, Guercio L, Sangalli C, Lo Cascio M, et al. Development and validation of a natural language processing algorithm for extracting clinical and pathological features of breast cancer from pathology reports. *JCO Clin Cancer Inform*. 2024;8:e2400034. <https://doi.org/10.1200/cci.24.00034>
10. Hochheiser H, Finan S, Yuan Z, Durbin E, Jeong J, Hands I, et al. DeepPhe-CR: natural language processing software services for cancer registrar case abstraction. *JCO Clin Cancer Inform*. 2023;7:e2300156. <https://doi.org/10.1200/CCI.23.00156>