

BRIEF COMMUNICATION

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

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

Constanza Pardo¹

¹ Cancer Epidemiological Surveillance Group, Instituto Nacional de Cancerología, Bogotá, D. C., Colombia.

Submission date: 25/07/2025
Acceptance date: 04/12/2025
Available online: 02/03/2026

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.

Conflicts of interest

The authors declare no conflicts of interest.

Citation

Pardo C. Innovation in data collection for population-based cancer registries in Colombia: implementation and challenges. Rev Col Cancerol. 2026;30(1):96-101. <https://doi.org/10.35509/01239015.1136>

Corresponding author

Elda Constanza Pardo-Ramos
Cancer Epidemiological Surveillance Group,
Instituto Nacional de Cancerología, Bogotá, D. C.,
Colombia.

Email:

cpardo@cancer.gov.co

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.

Introduction

Cancer surveillance defines population-based cancer registries (PBCR) as the gold standard for measuring cancer incidence and survival rates. This information forms the basis for evaluating national cancer control plans, designing prevention and early detection programs, and guiding cancer research (1).

Colombia currently has seven PBCR distributed across different geographic areas of the country, covering 25% of the national population. Data from these PBCR are used to estimate cancer incidence and mortality in Colombia (2) and for global estimates by the International Agency for Research on Cancer (3). Four PBCR are internationally recognized for their high quality: Bucaramanga, Cali, Manizales, and Pasto; while three are undergoing improvements: Barranquilla, Medellín, and Neiva (4). The *Instituto Nacional de Cancerología* (INC; National Cancer Institute) coordinates PBCR activities nationwide in Colombia, providing technical assistance, training, and partial funding.

Population growth, the increasing burden of cancer, the healthcare model, and the organization of service delivery have contributed to a rise in oncology institutions (which now number 2,421 in Colombia). This shift has prompted PBCR to undergo a “technological transition,” leading to increased volume, speed, and variability in data reported from various information sources related to cancer care. Currently, processes are more complex and slower, with an increase in different types of formats (structured and unstructured) and longer data collection times—challenges similar to those faced by other registries in low- and middle-income countries.

The lack of integration and standardization of information among cancer registries and data sources makes data extraction inefficient, leading to underreporting of cases and,

consequently, an underestimation of the cancer risk in the population. Pathology laboratories, which are the primary source of information for PBCR, generate large volumes of pathology reports in individual, unstructured digital files that are not standardized nationally, and each report contains precise diagnostic data and information relevant to a PBCR.

To address this situation, several initiatives have been developed using natural language processing (NLP) to enhance the efficiency of real-time case collection (5). Among these developments are the central cancer registries, part of the U.S. National Cancer Registry Program, which implemented a cloud-based platform for cancer surveillance. In Latin America and the Caribbean, some NLP-based initiatives currently exist, but they are not well documented. Two notable examples of technological innovation stand out: one regional (6) and one local (7), both of which implemented algorithms on clinical data and pathology reports for a hospital-based cancer registry.

The Cali PBCR (Colombia), with over 60 years of operation, modernized its data collection process a decade ago through information technology. Its challenge was to improve timeliness and minimize potential delays in analyzing and utilizing information (8). During this period, it successfully implemented a methodology to optimize data extraction from pathology reports using NLP and the Unified Medical Language System (UMLS) (9-10), converting natural language into structured data and classifying it as cancer or non-cancer, reducing collection time from 6 months to 48 hours (8). Additionally, this PBCR developed methods for relational database use and data visualization.

The purpose of this brief communication is to introduce the program “Optimization of extraction methods for PBCR in Colombia,” with defined phases and initial implementation in Neiva and Barranquilla, based on the technical transfer from the Cali PBCR, which is important for cancer surveillance in the country.

Implementation of the program “Optimization of extraction methods for PBCR in Colombia”

With the methodological innovation implemented by the Cali PBCR, the program “Optimization of extraction methods for ensuring the quality of information in PBCR in Colombia” was

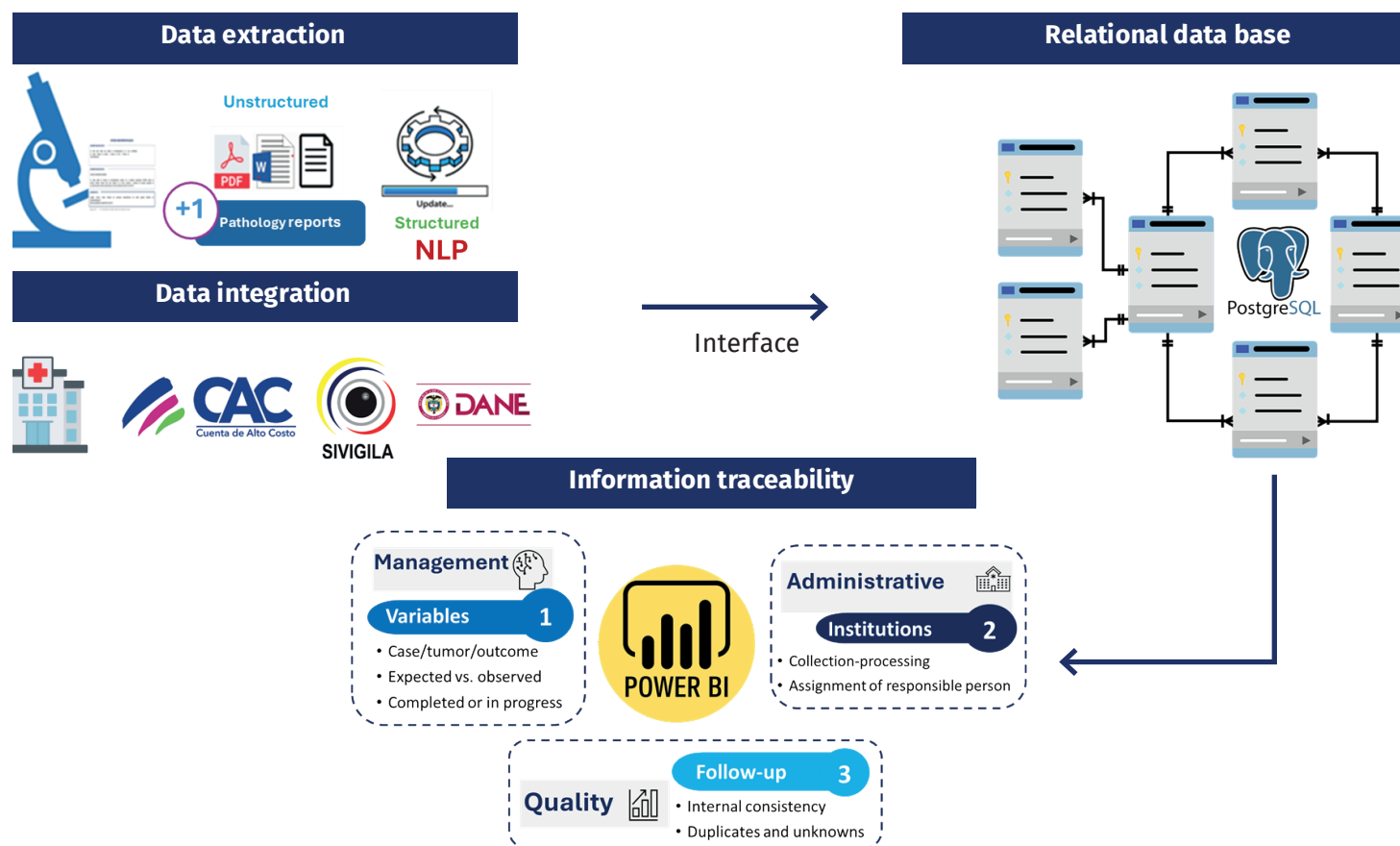
launched in 2024 as part of the INC’s “Transformative missions for comprehensive cancer control” initiative. Its goal was to transfer technical capabilities to other PBCR nationwide, initially including the registries in Barranquilla and Neiva. The development involved multiple phases focused on enhancing capabilities and adopting new methods for intelligent data extraction, relational database construction, and data visualization ([Table 1](#), [Figure 1](#)).

Table 1. Implementation phases for integrating new technologies for data collection and visualization in population-based cancer registries in Colombia

Phases	Activities	Tools
Census and characterization of PBCR information sources and systems	<ul style="list-style-type: none"> Review and characterization of data sources, as well as information and notification systems Review of standard operating procedures 	<ul style="list-style-type: none"> <i>Red Especial de Prestadores de Servicios de Salud</i> (REPS; Special Network of Health Service Providers) Notification systems: <i>Sistema Nacional de Vigilancia en Salud Pública</i> (Sivigila; National Public Health Surveillance System) of the National Institute of Health, high-cost diseases (cancer), and mortality (DANE) Data collection methods Format types: structured and unstructured data Characterization questionnaire in REDCap
Training	Virtual workshops: three in total, mixed format with synchronous connection for three days, two hours daily in each workshop, and asynchronous activities	<ul style="list-style-type: none"> Identification of data sources: types and characteristics Implementation of the data extraction process and creation of the relational database Visualization of tracking indicators with Power BI
Automated information management process: data extraction	Unstructured data: file configuration, standardization, consolidation, and supplementation of missing information	<ul style="list-style-type: none"> Source codes Templates Variable parameterization Python programming language Extraction scripts
Automated information management process: data integration	Creation of the relational database structure: extraction, processing, and automation	<ul style="list-style-type: none"> Scripts Dictionary adaptation Open-source object-relational database system in PostgreSQL
Data visualization and information traceability	<ul style="list-style-type: none"> Use of tools to connect to the relational database, transform, and visualize in real time Design and implementation of dynamic dashboards and visual panels 	<ul style="list-style-type: none"> Access Power BI platform Managerial, administrative, and quality information Indicators

DANE: *Departamento Administrativo Nacional de Estadística* (National Administrative Department of Statistics); PBCR: Population-based cancer registries.

Source: Own elaboration based on the supporting documents under Inter-administrative Agreement 2024-0102, INC - *Universidad del Valle*.



DANE: *Departamento Administrativo Nacional de Estadística* (National Administrative Department of Statistics); NLP: natural language processing; Sivigila: *Sistema Nacional de Vigilancia en Salud Pública* (National Public Health Surveillance System) of the National Institute of Health.

Source: Own elaboration based on the supporting documents under Inter-administrative Agreement 2024-0102, INC - *Universidad del Valle*.

Figure 1. Data collection and visualization technologies in the implementation of population-based cancer registries in Colombia

A census and characterization of information sources was conducted for each PBCR. Pathology laboratories and oncology institutions with unstructured data were prioritized, and data extraction was initiated with them. Three blended learning workshops, combining syn-chronous and asynchronous formats, were held to familiarize participants with the methodologies and facilitate their adoption. Both PBCR made progress in data extraction by developing scripts in the Python programming language to read structured and unstructured files (pathology reports) using specific UMLS libraries. This will enable data transfer to the relational database via an interface.

In Neiva, processing a structured data file in Excel took 20 seconds for 2,800 records, while processing 1,800 unstructured files in Word took 30 to 40 seconds.

In Barranquilla, processing a structured Excel file with 6,600 records took 3 minutes, whereas processing an unstructured Word file took 40 seconds, each file representing one person. These differences in processing times highlight the challenges and progress caused by the lack of a standardized format for the basic variables.

Currently, the functional part of the relational database is being implemented in PostgreSQL for consolidated data storage. Regarding the development of data visualization with Power BI, the design is complete, and progress is being made on implementing and adjusting functional visual dashboards for traceability, monitoring, and data quality control. The graphical user interface for the relational database is under development, which will facilitate the interaction of registry staff in viewing and updating cases.

Challenges

There are many challenges to address with the two selected cancer registries, so it is necessary to continue implementing the “cancer classification” and “quality” components to obtain the NLP model’s performance metrics (accuracy, sensitivity, and specificity) and achieve comprehensive operational capacity in both PBCR. Achieving interoperability with other national cancer information systems, such as the high-cost disease monitoring system (including cancer), the *Sistema Nacional de Vigilancia en Salud Pública (Sivigila*; National Public Health Surveillance System) of the National Institute of Health, and national mortality data, is also essential to enhance the efficiency of cancer data collection, validation, and analysis. Furthermore, the timeliness of information for incidence and survival indicators must be improved, and new resources must be secured to facilitate the adoption of these methodologies in other PBCR in Colombia.

This collaborative experience will, in principle, highlight common challenges and, perhaps, in the future, promote cooperation with other PBCR in the region, whether through training or technical assistance. The role of the Latin American regional node of the Global Initiative for Cancer Registry Development is vital in coordinating the various initiatives.

Conclusions

The collaboration between the INC and the Cali Population-Based Cancer Registry has enabled the transfer of technical capabilities to two other PBCR in the country (Neiva and Barranquilla). Implementing automated data extraction, a relational database, and a data visualization dashboard will enable, in a short period, the demonstration of reduced data collection times, improved reporting timeliness, and enhanced data quality through alerts on data behavior.

Acknowledgments

As author and leader of the program “Optimization of extraction methods for ensuring the quality of information in PBCR in Colombia,” I appreciate the contribution and work of the PBCR and the institutions to which they are attached:

- Members of the Cali PBCR: work carried out through Inter-administrative Agreement 2024-0102, signed between the INC and *Universidad del Valle*.
- Members of the Barranquilla (*Universidad del Norte*) and Neiva (*Universidad Surcolombiana*) PBCR.
- Technical and administrative team at the INC.

Funding

The program “Optimization of extraction methods for PBCR in Colombia” was financed by the Ministry of Science, Technology and Innovation, through the Health Research Fund (H190503003805) and own resources allocated to the Cancer Epidemiological Surveillance Program of the INC.

References

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). Lyon, Francia: International Agency for Research on Cancer [cited 2025 Jul 23]. Available from: <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 (Version 1.0). Bogotá, Colombia: Instituto Nacional de Cancerología; 2024 [cited 2025 Jun 21]. Available from: <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, García 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>