ORIGINAL

Toward Efficiency and Accuracy: Implementation of a Semiautomated Data Capture and Processing Model for the Construction of a Hospital-based Tumor Registry in Chile

Hacia la eficiencia y precisión: implementación de un modelo automatizado parcial de captura y procesamiento de datos para la construcción de un Registro Hospitalario de Tumores en Chile

CarolinaVillalobos^{[1](https://orcid.org/0009-0003-0607-4684) ©} ⊠, Carla Cavallera^{1 ©} ⊠, Matías Espinoza² ⊠, María Francisca Cid¹ ⊠, Inti Paredes^{2 ©} ⊠

1 Arturo López Pérez Oncology Institute (FALP), Hospital-based Tumor Registry, Cancer Research Department, Santiago, Chile. 2 Arturo López Pérez Oncology Institute (FALP), Medical Informatics and Data Science Unit, Cancer Research Department, Santiago, Chile.

Cite as: Villalobos C, Cavallera C, Espinoza M, Cid MF, Paredes I. Toward Efficiency and Accuracy: Implementation of a Semiautomated Data Capture and Processing Model for the Construction of a Hospital-based Tumor Registry in Chile. Data and Metadata 2023;2:124. [https://](https://doi.org/10.56294/dm2023124) [doi.org/10.56294/dm2023124.](https://doi.org/10.56294/dm2023124)

Submitted: 26-07-2023 **Revised:** 28-10-2023 **Accepted:** 10-12-2023 **Published:** 11-12-2023

Editor: Prof. Dr. Javier González Argot[e](https://orcid.org/0000-0003-0257-1176)

ABSTRACT

Introduction: the innovative implementation of a Hospital-based cancer registry (HBCR) at the Arturo López Pérez Oncology Institute (FALP), showcasing the transition from a manual data extraction model to a semiautomation of the process. The purpose of this publication is to compare both methodologies by assessing their efficiency and accuracy.

Methods: the analysis was conducted by comparing the complete dataset of the FALP HBCR from 2017 to 2021. The efficiency variable is analyzed, taking into account the total execution time of the registration process, and the precision variable was measured through the internal data consistency method using the IARCcrg Tools Software

Results: in terms of efficiency, the analysis reveals that in 2017, employing a manual approach without automation, it was necessary to analyze 13 061 cases over 144 weeks with an average of 4 registrars to achieve a total of 3 211 cases fully registered. In contrast, over the subsequent 4 years (2018 to 2021), with varying degrees of automation, 65 088 cases were analyzed within 115 weeks, employing an average of 8 registrars, resulting in 13 537 fully registered. This method demonstrated to be 3 times more efficient. Regarding precision, the manual approach exhibited a 5 % error rate in registered cases, whereas the automated approach showed a 0,6 % error rate during the 2018-2021 period.

Conclusion: the obtained results highlight the significant impact of semi-automating the tumor registration process through the utilization of tools for data capture and processing, achieving a threefold increase in efficiency and reducing errors to 0,6 %.

Keywords: Hospital-Based; Cancer Registry; Automation.

RESUMEN

Introducción: la innovadora implantación de un Registro Hospitalario de Cáncer (RHCC) en el Instituto Oncológico Arturo López Pérez (FALP), muestra la transición de un modelo manual de extracción de datos a una semiautomatización del proceso. El objetivo de esta publicación es comparar ambas metodologías evaluando su eficiencia y precisión.

Métodos: el análisis se ha realizado comparando el conjunto de datos completo del HBCR de la FALP desde 2017 hasta 2021. Se analiza la variable eficiencia, teniendo en cuenta el tiempo total de ejecución del proceso de registro, y la variable precisión se midió a través del método de consistencia interna de datos utilizando el Software IARCcrg Tools**.**

© 2023; Los autores. Este es un artículo en acceso abierto, distribuido bajo los términos de una licencia Creative Commons ([https://](https://creativecommons.org/licenses/by/4.0) [creativecommons.org/licenses/by/4.0\)](https://creativecommons.org/licenses/by/4.0) que permite el uso, distribución y reproducción en cualquier medio siempre que la obra original sea correctamente citada

Resultados: en términos de eficiencia, el análisis revela que en 2017, empleando un enfoque manual sin automatización, fue necesario analizar 13,061 casos durante 144 semanas con un promedio de 4 registradores para lograr un total de 3,211 casos registrados en su totalidad. En cambio, durante los 4 años siguientes (2018 a 2021), con diversos grados de automatización, se analizaron 65 088 casos en 115 semanas, empleando una media de 8 registradores, lo que dio como resultado 13 537 casos registrados en su totalidad. Este método demostró ser 3 veces más eficiente. En cuanto a la precisión, el enfoque manual exhibió una tasa de error del 5 % en los casos registrados, mientras que el enfoque automatizado mostró una tasa de error del 0,6 % durante el período 2018-2021.

Conclusiones: los resultados obtenidos destacan el impacto significativo de la semiautomatización del proceso de registro de tumores mediante la utilización de herramientas para la captura y procesamiento de datos, logrando triplicar la eficiencia y reducir los errores al 0,6 %.

Palabras clave: Registro Hospitalario De Cáncer; Automatización.

INTRODUCTION

Cancer registries are information systems designed for the collection, storage, and management of oncological data.**(1)** Generally, the gathered data serve the purpose of monitoring cancer incidence, tracking treatment schemes, and evaluating the effectiveness of public policies aimed at prevention and improved survival rates.**(2)**

The collection and analysis of accurate and up-to-date data are pivotal for advancing scientific research in healthcare. In this context, the development of a partially automated data capture and processing model for constructing a Hospital-based tumor registry in Chile emerges as an innovative and efficient tool.

In the year 2018, the Arturo López Pérez Foundation Oncological Institute (FALP) implemented a Hospitalbased cancer registry (HBCR) based on a manual registration model utilizing Microsoft Excel spreadsheets. Subsequently, in 2020, this evolved into a model that highlights a semi-automated registry process and a custom-developed software.

This approach enhances the efficiency and precision of data capture and processing, while valuing human resource involvement. This article aims to compare both methodologies employed in the development process of the FALP HBCR, by analyzing their efficiency and precision, offering valuable insights for oncology centers interested in implementing partially automated registration systems.

METHODS

The study was based on comparing the complete set of incident cancer cases with full registration in the FALP HBCR from 2017 to 2021. The sample comprises 16,748 cases, including 3 211 cases in 2017, 2 959 in 2018, 3 299 in 2019, 2 868 in 2020, and 4 411 in 2021. No cases were excluded from the analysis. See table 1.

Before delving into the comparative analysis, a description of the FALP HBCR process and the manual and partially automated working models for all incident cases registered from 2017 to 2021 is provided.

The definition of the FALP Hospital-based cancer registry process arose from an institutional decision, aligned with the patient care and admissions at the Institute. Three major stages were established:

1. Construction of a matrix of information sources: this process is overseen by a data engineer, who generates a matrix from cross-referencing data from FALPs electronic health record (EHR) systems (such as medical consultations, pathology reports, oncology committees, hospital discharges, among others). Subsequently, rules for case selection, tested within the institution, are established and applied in collaboration with the FALP HBCR team to identify potentially registrable cases with utmost

3 Villalobos C, *et al*

specificity.

- 2. Case identification and capture: during this stage, all registrable incident cases contained within the matrix are identified and classified.
- 3. Case registration: this phase involves registering the set of data for each registrable case, obtained through the collection of information from the EHR systems. The variables to be registered are internationally established by the International Agency for Research on Cancer.**(3)** This stage is subdivided into three steps, managed by distinct data entry profiles based on the level of data complexity and interpretation.

For the registration of incident cases in 2017, a manual working model was implemented, utilizing a Microsoft Excel spreadsheet for registration. The registration team consisted of an average of 4 registry technicians with a technical nursing profile (TENS) working full-time. The registration process was divided into three stages: selected Database, Identification, and Registration. Monitoring and management control were manually executed by the FALP HBCR coordination team, consisting of two nurses supervising assigned cases and daily completions per registrar. The Minimum Data Set (MDS) utilized, comprised a total of 50 variables. See table 2.

For the registration of incident cases from 2018 to 2021, a semi-automated data extraction model was gradually implemented, transitioning to a partially automated system with three registration profiles through locally developed software with specific rules for case identification and capture, as well as the provision of relevant clinical data. To enhance efficiency in data capture and processing, additional solutions developed by the FALP Medical Informatics and Data Science team (IMDS) were employed, including natural language processing (NLP) tools, Topo-Morfo and Oncotext. Topo-Morfo is homemade software designed for automatic detection and coding of morphology and topography from non structured data. Oncotext is a homemade software that searches and identifies clinical terms in a vast volume of documents with non structured data. The registration team consisted of an average of 8 registry technicians with nursing and medical profiles working full-time. The registration process was divided into three stages: Selected Database, Identification, and Registration. A data engineer was incorporated to support a cancer repository for this project. Power BI was used for monitoring and management control. Finally, the Minimum Data Set (MDS) utilized, comprised a total of 100 variables. See table 2.

After describing the manual and semi automated working models, two relevant variables were identified to conduct a comparative descriptive analysis of both working models. To compare the two models, two variables were established: Efficiency (execution time) and Precision (data quality).

Efficiency refers to the total execution time of the registration process. Time measurement was standardized in weeks, starting from the stage of case identification and capture until 95% of registrable cases were fully registered.

Precision relates to data quality and reliability. This variable was measured through internal consistency and pertains to data plausibility and compatibility across datasets. The IARCcrg Tools Software were employed for this purpose.^(4,5,6) This tool enables the identification of unlikely and impossible coding combinations, designed for validity and internal consistency analysis among the following datasets:

- Age and incidence dates and birthdates
- Age and topography
- Age and morphology
- Sex and topography
- Sex and morphology
- Topography and morphology
- Behavior and morphology
- Differentiation grade and morphology
- Diagnostic basis and morphology

For data processing through the IARCcrg Tools Software, each case was assigned a unique identification number for subsequent verification. The dataset included variables such as sex, topography, morphology/ behavior, differentiation grade, incidence and birth dates, diagnosis and age at diagnosis. Years with missing differentiation grade data in their MDS, was treated as unknown. This applies to the years 2017, 2018, and 2019. The IARCcrg Tools Software generated warnings and errors for each registration year. Subsequently, the dataset was recoded according to IARCcrg Tools Software standards for analysis.

RESULTS

In this section, we will be showcasing the outcomes of our analysis related to the efficiency and precision variables, comparing both data extraction and processing models: manual and semi-automatic.

Efficiency Variable

Tables 3 and 4 display the comparison of different variables for the efficiency analysis (2017 - 2021), using different degrees of automation in the process.

Regarding execution time, a reduction in the number of weeks required to complete the annual registration process is evident when comparing data extraction models. To complete the registration for the year 2017, it was necessary to analyze 13,061 cases, over 144 weeks with an average of 4 registry technicians. In contrast, the years 2018 to 2021, there was an increase in the average number of registry technicians from 4 to 8. Consequently, we see an increase in the overall volume of cases analyzed with a reduction in the time required to complete the registration process. For example, in the year 2018, when analyzing 16,489 cases (21 % more than in 2017), a reduction in the required time to finalize the process is evident, decreasing to 36 weeks compared to the 144 weeks in 2017.

In 2019, with an average of 8 registry technicians, 16,610 cases were analyzed, completing the total registration for that year in 28 weeks. Similarly, data analysis reveals that for 2020, 13 721 cases were completed with an average of 8 registry technicians in 22 weeks. Finally, in 2021, with 18 268 cases analyzed, the process was completed in 29 weeks with 8 registry technicians.

When grouping data according to the data extraction model, the results show that in 2017, using a manual extraction model without any degree of automation, it took 144 weeks with an average of 4 registry technicians to analyze 13,061 cases. In contrast, during the period from 2018 to 2021, employing a data extraction model with varying degrees of automation, the analysis of 65,088 cases was completed within a total of 115 weeks, with an average of 8 registry technicians. The entire registration process for the 2018-2021 period took less time compared to the complete registration process for the year 2017.

Table 4 illustrates the total number of incident cases per year from the registration period 2017 to 2021. For the registration of incident cases in 2017, a manual data extraction model was used. In this period, a total of 105 127 medical consultations were conducted, identifying 9,850 potentially registrable cases. As a result of this process, a total of 3 211 incident cases were validated and registered.

In the registration process for the year 2018, the data extraction model transitioned to a semi-automated approach. Throughout this year, a total of 97 811 medical consultations were conducted, resulting a total of 7 612 potentially registrable cases. Consequently, 2 959 incident cases were validated and registered.

During the years 2019 and 2020, a semi-automated data extraction model was fine-tuned. In 2019, a total of 113 258 medical consultations were conducted, identifying 6 713 potentially registrable cases. As a result of this process, 3 299 incident cases were validated and registered. In 2020, 89 599 medical consultations were conducted, identifying 5 117 potentially registrable cases. As a result of this process, 2 868 incident cases were

5 Villalobos C, *et al*

validated and registered.

For the registration process in 2021, a more advanced automation model was introduced, allowing the complete identification of 843 cases automatically in addition to the previous semi automatic model. Throughout this year, a total of 121,803 medical consultations were conducted, yielding 5 035 potentially registrable cases. Consequently, 4 411 incident cases were registered.

The varying degrees of automation during the period 2017-2021 displayed a positive reduction in the pool of potentially registrable cases, thanks to the implementation of precise rules for data mining and capture. Additionally, the proportion of registered cases in relation to the total pool of potentially registrable cases experienced a significant increase upon introducing new algorithms for case processing and capture.

Precision Variable

Using de IARCcrg Tools Software, out of the total of 16 748 cases, the program analyzed 16 746 cases. The program did not analyze the entirety of cases, as in 2 instances, the diagnostic code conversion from the International Classification of Diseases for Oncology, Third Edition (ICD-O3) to the International Classification of Diseases, Tenth Edition (ICD-10) was not carried out.

The IARCcrg Tools Software reports errors containing invalid variable combinations and warnings corresponding to improbable combinations.**(5)**

In terms of errors found in the analysis, out of the 16,746 cases evaluated, the IARCcrg Tools Software reported 241 individual errors, representing 1,4 % of the total cases. The errors were categorized by type: 121 behavior code errors, accounting for 0,72 %; 115 invalid age errors, accounting for 0,70 %; and 5 diagnosis date errors preceding birth date, accounting for 0,03 % of the total analyzed cases.

Analyzing errors by year, in 2017, the tool reported 160 errors, representing 5 % of the total cases registered that year. In 2018, 10 errors were reported, equivalent to 0,3 % of the cases registered that year. For 2019, only 1 error was reported, representing 0,03 % of the total cases registered that year. In 2020, 10 errors were identified, representing 0,5 % of the fully registered cases analyzed for that year. Lastly, in 2021, 56 errors were found, representing 1,3 % of the cases registered that year.

Regarding the registration model, for 2017, where a manual data extraction model was used within an Excel spreadsheet, 160 errors were reported, accounting for 5 % of the total cases registered under that modality. Conversely, the model with different degrees of automation, and the use of our custom made HBCR software, covering the period 2018 to 2021, reported only 81 errors, representing 0,6 % of the total number cases registered.

Table 6 shows that 319 warnings were generated from the 16,746 cases analyzed, corresponding to 1,9 % of the total cases. The IARCcrg Tools Software reports 198 warnings for improbable combinations between topography and morphology, accounting for 1,2 % of the total cases; 110 warnings for improbable combinations between diagnostic basis and morphology, accounting for 0,7 % of the total cases; 6 warnings for improbable combinations between sex and morphology, accounting for 0,4 % of the total cases; 3 warnings for improbable combinations between age, topography, and morphology, accounting for 0,02 % of the total cases; and 2 warnings for improbable combinations between behavior and morphology, accounting for 0,01 % of the total cases.

Analyzing warnings by year of diagnosis, in 2017, the tool reported 90 warnings, representing 2,8 % of the total cases registered that year. In 2018, 34 warnings were reported, equivalent to 1,1 % of the cases registered that year. For 2019, 55 warnings were observed, representing 1,6 % of the total cases registered that year. In 2020, 79 warnings were identified, representing 2,7 % of the fully registered cases analyzed for that year. Lastly, in 2021, 61 warnings were found, representing 1,3 % of the cases registered that year.

Regarding the registry model, for 2017, where a manual data extraction model was used without technical profiles and within an Excel spreadsheet, 90 warnings were reported, accounting for 2,8 % of the total cases registered under that modality for 1 year. On the other hand, the registry model with varying levels of automation, utilizing technical profiles for registration and our HBCR software, during the period of 2018 to 2021, recorded a total of 229 warnings through the IARCcrg Tools Software, constituting 1,6 % over the fouryear period.

DISCUSSION

The impact of a Hospital Based Cancer Registries depends on the quality of its data.

Efficiency, data-driven decision-making, and cancer-related considerations are valuable resources for clinicians, researchers, and policymakers, aiding in the understanding of cancer patterns, treatment outcomes, and the effectiveness of various interventions. It is essential to generate credible and timely data that enable well-informed decisions.(**9,10,11,12,13)** In response to this requirement, we have developed an anonymized and publicly accessible web viewer. This information is updated every 30 days and can be accessed throw the following link: <https://rht.oncodata.org>

By employing NLP and data mining tools for data capture and processing within a semi-automated framework, it offers significant advantages compared to manual methods. These benefits offers substantial time savings and an enhancement in result accuracy.**(8)**

For FALP, the results demonstrate that when comparing both methods, the total registration process for the years 2018 to 2021 was completed in less time than the complete registration for the year 2017.

The partial automation framework has a positive impact on all stages of the registration process, as it establishes rules for search and capture of specific data, increasing the detection of potentially registrable cases per year. Additionally, the availability of clinical data automatically integrated from de EHR into the HBCR software allows the registry technicians to act as a validators, speeding the registration process. This reaffirms the importance that although partial automation improves efficiency, human resource participation is still required for validation.**(8)**

Ensuring data quality in a cancer registry is an ongoing process, including routine validation and coherence checks. The most robust validation method involves comparing registered data with original source documents to evaluate the accuracy of data extraction from these documents. Nevertheless, it is important to control the internal coherence of registry data, i.e., detecting impossible or improbable combinations of codes for different data elements.**(5)**

The results obtained in FALP when analyzing internal consistency of information through the IARCcrg Tools Software, show that the introduction of a semi-automated registration framework with the use of technical profiles to capture data according to its complexity has significantly reduced errors. Likewise, upon analyzing

7 Villalobos C, *et al*

warnings generated by the IARCcrg Tools Software, it is concluded that they do not always correspond to errors but can represent improbable or unusual combinations. Furthermore, the use of IARCcrg Tools Software and the cross-referencing of new variables strengthen data quality control.

In the future, tumor registries should move towards developing systems with some degree of automation, incorporating rules for internal consistency validation at the time of data entry, facilitating immediate case registration review rather than post-editing.

CONCLUSION

This study supports the adoption of a partially automated model for tumor registration, which involves the creation of custom-made software utilizing NLP tools for handling unstructured data and integration with the electronic health record. This model has shown superior efficiency and precision in data when compared to the manual approach achieving a threefold increase in efficiency and reducing errors to a 0,6 %. Additionally, the significant importance of systematically verifying quality and coherence in the registration of oncological cases is highlighted, employing tools such as IARCcrg Tools Software to enhance data quality.

REFERENCES

1. NIH. Cancer Registry - What is Cancer Registry? [Internet]. National Cancer Institute. [https://seer.cancer.](https://seer.cancer.gov/registries/cancer_registry/cancer_registry.html) [gov/registries/cancer_registry/cancer_registry.html](https://seer.cancer.gov/registries/cancer_registry/cancer_registry.html).

2. White MC, Babcock F, Hayes NS, Mariotto AB, Wong FL, Kohler BA, et al. The History and Use of Cancer Registry Data by Public Health Cancer Control Programs in the United States. Cancer. 2017;123 Suppl 24:4969- 4976. <https://doi.org/10.1002/cncr.30905>.

3. International Agency for Research on Cancer. International Classification of Diseases for Oncology. 3rd ed. Geneva: WHO; 2000.

4. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer. 2009;45(5):747-755. [https://doi.org/10.1016/j.](https://doi.org/10.1016/j.ejca.2008.11.032) [ejca.2008.11.032](https://doi.org/10.1016/j.ejca.2008.11.032)

5. Ferlay J, Burkhard C, Whelan S, Parkin DM. Check and conversion programs for cancer registries (IARC/ IACR Tools for Cancer Registries). IARC Technical Report No. 42, Lyon, France. 2005. [https://cri.tums.ac.ir/2/](https://cri.tums.ac.ir/2/pbcr/References/Check and Conversion Programs for Cancer Registries.pdf) [pbcr/References/Check%20and%20Conversion%20Programs%20for%20Cancer%20Registries.pdf](https://cri.tums.ac.ir/2/pbcr/References/Check and Conversion Programs for Cancer Registries.pdf)

6. Hilsenbeck S, Glaefke G, Feigl P, Lane W. Quality for cancer registries. US Department of Health and Human Services. 1985;16-23. https://seer.cancer.gov/archive/manuals/historic/quality_control.pdf

7. al-Haddad BJ, Jedy-Agba E, Oga E, et al. Comparability, diagnostic validity and completeness of Nigerian cancer registries. Cancer Epidemiol. 2015;39(3):456-464.<https://doi.org/10.1016/j.canep.2015.03.010>.

8. Ribes J, Gálvez J, Melià À, Clèries R, Messeguer X, Bosch FX. Automatización de un registro hospitalario de tumores. Gaceta Sanitaria. 2005;19(3):221-228.

9. Epstein D, Bermúdez-Tamayo C, Cantarero D, et al. Special edition of Gaceta Sanitaria on evidence-based decision making in public health. Gac Sanit. 2018;32:403-4.

10. Canova-Barrios C, Machuca-Contreras F. Interoperability standards in Health Information Systems: systematic review. Seminars in Medical Writing and Education 2022;1:7–7. <https://doi.org/10.56294/mw20227>

11. Herrera IMR, Castañeda MEG, Henrion CT. La toma de decisiones en salud pública: Una revisión del procedimiento desde el enfoque racional. RESPYN Revista Salud Pública y Nutrición. 2012;13(2).

12. Inastrilla CRA. Data Visualization in the Information Society. Seminars in Medical Writing and Education 2023;2:25–25. <https://doi.org/10.56294/mw202325>

13. Saesen R, Van Hemelrijck M, Bogaerts J, Wilson R, Lacombe D, van der Graaf WT, et al. Definición del papel de los datos del mundo real en la investigación clínica del cáncer: la posición de la Organización Europea para la Investigación y el Tratamiento del Cáncer. Eur J Cancer. 2023.186:52-61. [https://doi.org/10.1016/j.](https://doi.org/10.1016/j.ejca.2023.03.013) [ejca.2023.03.013](https://doi.org/10.1016/j.ejca.2023.03.013)

FUNDING

The authors did not receive any funding for the development of this research.

CONFLICT OF INTERESTS

The authors declare no conflicts of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Carolina Villalobos, Carla Cavallera, Inti Paredes. *Data curation:* Carolina Villalobos, Carla Cavallera, María Francisca Cid, Matías Espinoza. *Formal analysis:* Carolina Villalobos, Carla Cavallera. *Research:* Carolina Villalobos, Carla Cavallera. *Methodology:* Carolina Villalobos, Carla Cavallera. *Project management:* Carolina Villalobos, Carla Cavallera. *Resources:* Carolina Villalobos, Carla Cavallera. *Software:* Carolina Villalobos, Carla Cavallera, Inti Paredes. *Supervision:* Inti Paredes. *Validation:* Carolina Villalobos, Carla Cavallera. *Display:* Carolina Villalobos, Carla Cavallera. *Drafting - original draft:* Carolina Villalobos, Carla Cavallera. *Writing - proofreading and editing:* Carolina Villalobos, Carla Cavallera, Inti Paredes.