Data and Metadata. 2023; 2:123
doi: 10.56294/dm2023123

DATA & METADATA

ORIGINAL

Comparative Analysis of Classification Models for Predicting Cancer Stage in a Chilean Cancer Center

Análisis comparativo de modelos de clasificación para predecir el estadio del cáncer en un centro oncológico chileno

Marcela Aguirre[1], Sergio Peñafiel[1], April Anlage[2], Emily Brown[2], Cecilia Enriquez-Chavez[2], Inti Paredes[1]

[1]Instituto Oncológico Fundación Arturo López Pérez, Medical Informatics and Data Science Unit, Department of Cancer Research, Santiago, Chile.
[2]Massachusetts Institute of Technology, Global Health Informatics Course, Boston, USA.

Cite as: Aguirre M, Peñafiel S, Anlage A, Brown E, Enriquez-Chavez C, Paredes I. Comparative Analysis of Classification Models for Predicting Cancer Stage in a Chilean Cancer Center. Data and Metadata 2023;2:123. https://doi.org/10.56294/dm2023123.

Submitted: 02-09-2023 **Revised:** 18-10-2023 **Accepted:** 07-12-2023 **Published:** 08-12-2023

Editor: Prof. Dr. Javier González Argote

ABSTRACT

This study aimed to develop a predictive model for cancer stage using data from a Chilean cancer registry. Several factors, including cancer type, patient age, medical history, and time delay between diagnosis and treatment, were examined to determine their association with cancer stage. Multiple supervised multi-class classification methods were tested, and the best-performing models were identified. The results showed that the random forest, SVM polynomial, and composite models performed well across different stages, although distinguishing between Stages II and III was more challenging. The most important features for predicting cancer stage were found to be cancer type, TNM variables, and diagnostic extension. Variables related to treatment timing and sequence also showed some importance. It was emphasized that the results of predictive models should be interpreted carefully to avoid overprediction or underprediction. Clinical context and additional information should be considered to enhance the accuracy of predictions. The small dataset and limitations in data availability posed challenges in accurately predicting cancer stage for different cancer types. Implementing the predictive model can have various benefits, including informing treatment decisions, assessing disease severity, and optimizing resource allocation. Further research and expansion of the model's scope were recommended to improve its performance and impact. Overall, the study emphasized the potential of predictive models in cancer staging and highlighted the need for ongoing advancements in this field.

Keywords: Cancer Staging; Machine Learning; Decision Support Techniques.

RESUMEN

El objetivo de este estudio fue desarrollar un modelo predictivo del estadio del cáncer a partir de datos de un registro chileno de cáncer. Se examinaron varios factores, entre ellos el tipo de cáncer, la edad del paciente, los antecedentes médicos y el tiempo transcurrido entre el diagnóstico y el tratamiento, para determinar su asociación con el estadio del cáncer. Se probaron múltiples métodos supervisados de clasificación multiclase y se identificaron los modelos con mejores resultados. Los resultados mostraron que los modelos de bosque aleatorio, SVM polinómico y compuesto funcionaban bien en los distintos estadios, aunque distinguir entre los estadios II y III era más difícil. Las características más importantes para predecir el estadio del cáncer fueron el tipo de cáncer, las variables TNM y la extensión del diagnóstico. Las variables relacionadas con el momento y la secuencia del tratamiento también mostraron cierta importancia. Se hizo hincapié en que los resultados de los modelos predictivos debían interpretarse con cautela para evitar predicciones excesivas o insuficientes. El contexto clínico y la información adicional deben tenerse en cuenta para mejorar la exactitud de las predicciones. El pequeño conjunto de datos y las limitaciones en la disponibilidad de los mismos plantearon retos a la hora de predecir con exactitud el estadio del cáncer para diferentes tipos de cáncer. La aplicación del modelo de predicción puede tener varias ventajas, como informar las decisiones de tratamiento, evaluar la gravedad de la enfermedad y optimizar la asignación de recursos. Se recomendó seguir investigando y ampliar el alcance del modelo para mejorar su rendimiento e impacto. En conjunto, el estudio pone de relieve el potencial de los modelos predictivos en la estadificación del cáncer y subraya la necesidad de seguir avanzando en este campo.

Palabras clave: Estadificación Del Cáncer; Aprendizaje Automático; Técnicas De Apoyo A La Toma De Decisiones.

INTRODUCTION

Driven by a growing and aging population and social determinants of health, cancer morbidity and mortality have increased exponentially around the world. Worldwide, an estimated 19,3 million new cancer cases and almost 10,0 million cancer deaths occurred in 2020.(1) The region of the Americas represents 20,9 % of incidence and 14,2 % of mortality worldwide.(2) In Chile, a country with a population of 19 million, there were 54 227 new cancer cases and 28 584 deaths according to Globocan. However, in the face of this burden, access to cancer care can be a challenge in Chile. Exacerbated by COVID-19 when resources were diverted to pandemic related efforts, access to care is challenging due to limited human resources, long wait times for specialized care, and high out-of-pocket costs.

Timely detection and effective treatment are key to increasing survival rates. The diagnosis and treatment of cancer has advanced significantly in recent years with the advent of new imaging technologies, histopathology, advancements in existing therapies, and with new therapies such as interventional radiology and immunotherapy.(3) However, despite these advancements, many cancers remain undetected until the disease has progressed to advanced stages due to systemic challenges faced in creating access to care. Many methods and tools have been developed to try to more effectively distribute the limited resources available in this space. Prevailing clinical methods include interventions such as routine screening in target populations and nomograms for prostate cancer. Additionally, many machine learning models have started to crop up to meet this challenge as well. Tools like "Predict" for breast cancer or AI-enabled analysis of dermoscopy images have been developed to help inform clinicians' care decisions.(4,5) They have developed algorithms after studying thousands of cases and their characteristics as well as their treatments along with statistical models to predict the best treatment for women with breast cancer.

The diagnosis and prognosis of cancer are influenced by different factors including type of cancer, organ affected, stage, dissemination, histopathological grade of cancer, age, baseline health of the individual and response to treatment.(6) Stage can be used as a particularly helpful predictor for prognosis and can help inform the appropriate treatment for a patient. A common system for determining the stage of cancer is the TNM, which describes the location of the tumor and its extent, the nearby nodes that have cancer, and whether it has metastasized.(6) Despite its usefulness, staging data may not be collected or be inaccessible in Chile given the disjointed nature of care. Providers must rely on siloed sources for information about a patient's diagnosis and treatment plan, which presents meaningful barriers to care provision as well as patient prioritization given the resource-constrained environment.

The goal of this study is to review the process for creating a predictive model for cancer stage, based on curated data from a cancer registry available from Fundación Arturo López Pérez (FALP), a non-profit cancer center in Chile. Certain variables, such as cancer type, patient age, medical history, time delay between diagnosis and treatment and other factors may be key indicators of disease stage. Using these variables as input, several supervised multi-class classification methods are tested in order to evaluate their performance on a labeled data set. It is expected that one or several of these models will be able to predict cancer stage reliably and accurately.

This model could be used as a tool that helps physicians make informed decisions about treatment and patient care. By predicting the stage of the cancer, the severity of the disease can be determined, and its likely progression over time. Additionally, the model can be useful for patient prioritization, allowing the medical and management team to design more effective and suitable planning and resource allocation, enabling institutions to get care to the right people at the right time.

METHODS

A retrospective cross-sectional study was conducted during the first semester of 2023, using a dataset with incident cancer cases collected between 2018 and 2021 in FALP. A dashboard visualization of the current data avalaible the dataset is available in https://rht.oncodata.org/.

Data Description

This analysis leverages a proprietary dataset maintained by FALP's cancer registry unit. The dataset contains ~11 000 records curated and validated by a team of registry experts, based on electronic health records of the institution. The records were anonymized and de-identified for the purpose of this analysis. Any patient that was in the database was included in the analysis except for clear outliers that represented obvious errors (i.e. patient classified as female with prostate cancer). The main variables contained in the dataset are: demographic data (age, sex, region, insurance type, oncological insurance (yes or no)), cancer category, subcategory and stage, tumor topography and morphology, clinical tumor size (cT), clinical node status (cN), clinical metastasis status (cM), pathological tumor size (pT), pathological node status (pN), pathological metastasis status (pM), diagnostic extension, vital status (alive or dead), date of death, treatment type (i.e. surgery, chemotherapy, etc.) and date.

Variable Identification

Before building a model, an exploratory analysis is done to identify variables with predictive value relative to cancer staging. After developing a series of graphs and figures comparing distinct variables to stage, some variables were isolated as potentially predictive variables from others that demonstrated minimal predictive potential. The selected variables were used as input characteristics so that fair and direct comparisons could be made between models.

In addition to evaluating the pre-existing variables in the dataset, new variables were created to reflect potentially significant clinical dynamics. Prolonged time intervals between symptom onset and treatment can lead to poorer clinical outcomes and are associated with worse patient experience of subsequent cancer care.[7] Additionally, certain types of cancer have a high chance of cure if they are detected at an early stage and adequately treated.[8] Using the date of diagnosis and date receiving treatment in our data set we were able to build a composite variable to get the number of days between this to dates and included it in our model. The type of treatment depends on the type and progression of the cancer. Most people need a combination of treatments, such as surgery with chemotherapy and/or radiation therapy,[9] to effectively treat their cancer. Treatment sequencing can be another powerful predictor of cancer stage. For example, more advanced cancer might be treated with chemotherapy before surgery, in order to increase the chances of a successful surgery. This could be associated with the stage of the cancer at diagnosis. Including a variable that captures this can provide additional levels of retrospective insight.

Model Testing and Performance Evaluation

A variety of multi-class classification models were implemented and evaluated, including random forest, decision tree, logistic regression, K-nearest neighbors, support vector machine (linear), support vector machine (polynomial), and naive Bayes. The parameters of each model were adjusted in order to find the best performance, as evaluated by accuracy, precision, recall, F1 score, confusion matrices, and the area under the curve in receiver operating characteristic curves. A standard 70 % training, 30 % test split was used on the data set. Additionally, categorical variables (cancer category, sex, treatment, etc.) were processed using one-hot encoding, necessary for use of several classification models. A composite approach was tested, in which individual classifier models were trained on specific cancers types (restricted to the top 4 most common cancers in the data set, to avoid excessively small samples). The composite model used a polynomial SVM model for breast and skin cancer, a random forest model for digestive organ cancer, and a logistic regression model for male genital organ cancer.

In order to check for overfitting on the relatively small data set, cross validation was performed. K-folds cross validation was used to examine accuracy scores, using 10 folds. Evaluation statistics were computed in bulk and compared to the models that were trained on the combined data. Using only the cases labeled with cancer stage, these models are trained on a total of 3087 cases and then tested on 1324 cases.

Feature Importance

Feature importance is an important tool in classification models since it allows the identification of variables or characteristics that have the greatest influence on the classification result. This information is valuable since it allows the optimization of model performance and improves model accuracy. This can lead to greater accuracy and efficiency in classification.

RESULTS

Before modeling, basic descriptive analysis was performed to gain a deeper understanding of the composition of the data. The sample is split 60 % to 40 % female to male and breast cancer, skin cancer, digestive organs, cancer, masculine genitals cancer, and thyroid cancer compose ~80 % of the cases included. Table 1 below provides a basic overview of the breakdown by cancer type for the top 5 forms of cancer.

Table 1. Overview of data set

	Sample Size	Sex Distribution		Mean Age (years)
		Female	Male	
Total Dataset	11 023	6 589 (60 %)	4 434 (60 %)	60,1
Cancer Type Prevalence				
Breast Cancer	19,3 %	99,5 %	0,5 %	58,4
Skin Cancer	17,7 %	52,3 %	47,6 %	66,9
Digestive Organ Cancer	16,7 %	48,5 %	51,5 %	64,1
Male Genital Organs	11,8 %	0,76 %	99,2 %	62,9
Thyroid and other endocrine glands cancer	10,6 %	86,1 %	13,8 %	47,0

The following variables were isolated as potentially predictive variables: cancer type, insurance type, and extension. Age and region of the patient were other variables that were tested that demonstrated minimal predictive potential. The following variables were used as input: cancer category (i.e. skin), cancer subcategory (i.e. melanoma), sex, insurance type, oncological insurance (yes or no), clinical tumor size (cT), clinical node status (cN), clinical metastasis status (cM), pathological tumor size (pT), pathological node status (pN), pathological metastasis status (pM), diagnostic extension, vital status (alive or dead), treatment (i.e. surgery, chemotherapy, etc.). In addition, treatment sequence and days between diagnosis and treatment were included.

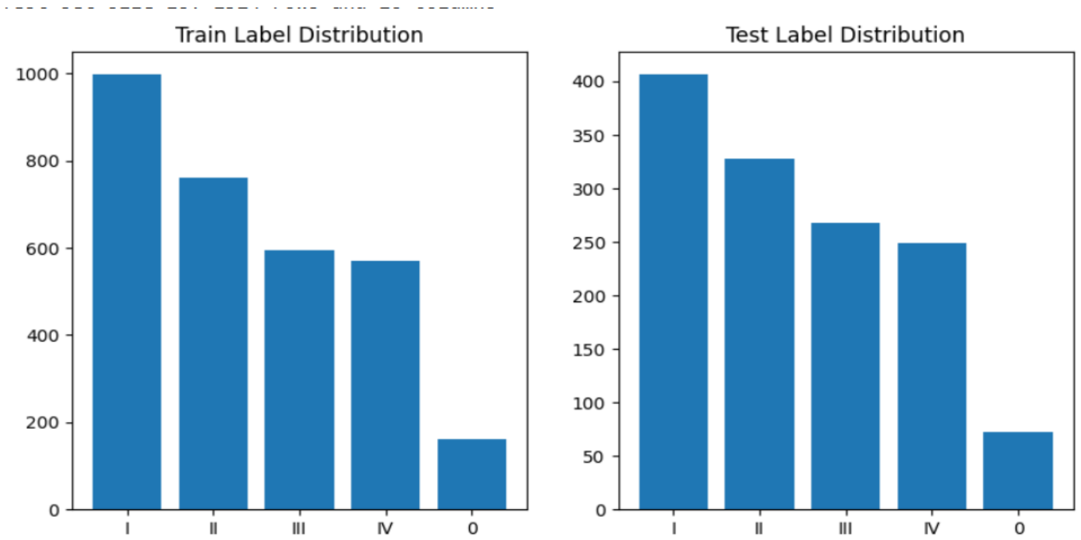


Figure 1. Label (cancer stage) distribution of train and test dataset after split

The training and testing randomized split was done ensuring the distribution of the cancer stage variable (predicted variable). Figure 1 shows the distribution of the label distribution.

In table 2 and 3, the results for the tested modela are reported, considering the performance parameters on the test dataset, ROC curve and confusion matrix. The best three models are shown (Random Forest, SVM polynomial and composite model).

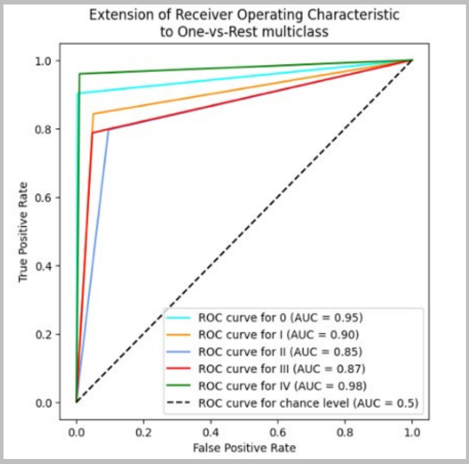
Table 2. Performance metrics of selected models

Model	Accuracy	Precision	Recall	F1
Random Forest (max depth = 20)	0,854	0,857	0,854	0,855
SVM (polynomial order 3)	0,857	0,860	0,857	0,858
Composite Model	0,856	0,858	0,856	0,857

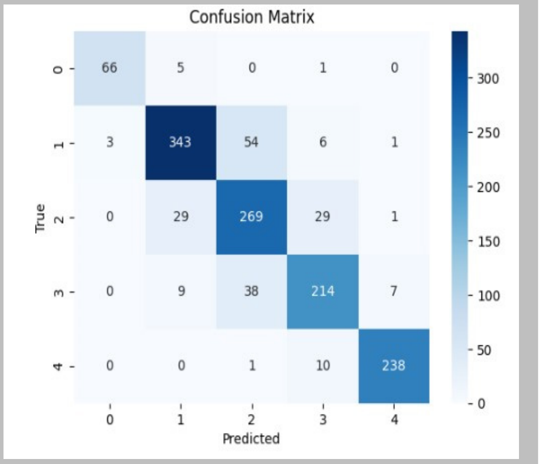
https://doi.org/10.56294/dm2023123

Table 3. ROC and Confusion Matrix of selected models

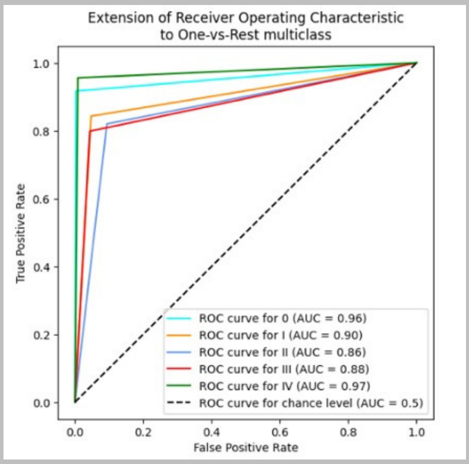
<table>
<tr><th>Model</th><th>AUC Graph</th><th>Confusion Matrix</th></tr>
<tr><td>Random Forest
(max depth = 20)</td><td>



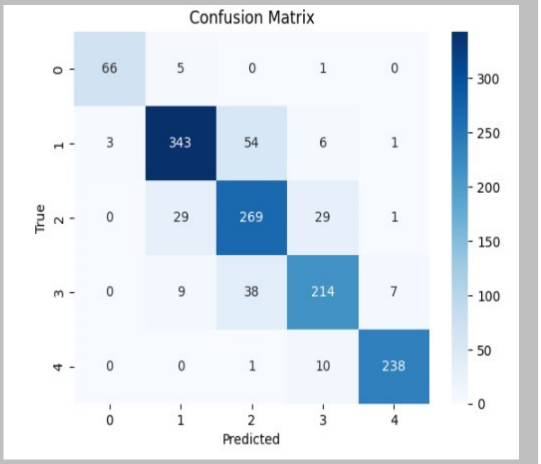
</td><td>



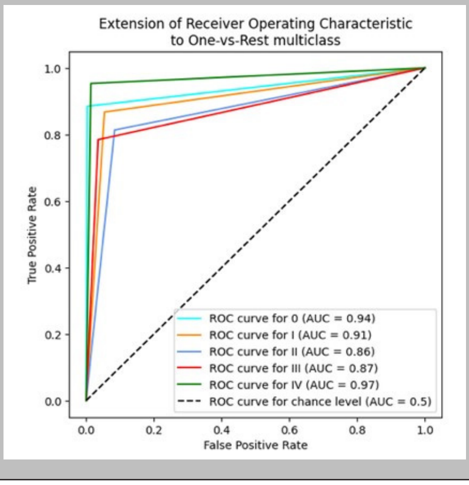
</td></tr>
<tr><td>SVM
(polynomial order 3)</td><td>



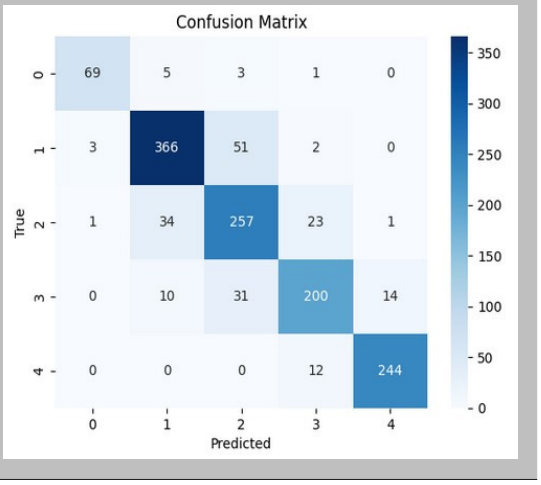
</td><td>



</td></tr>
<tr><td>Composite Model</td><td>



</td><td>



</td></tr>
</table>

The 10-fold cross-validation resulted for the random forest model in the average accuracy across the 10 folds of 84,4 %, with a standard deviation of 6,1 %. For the support vector machine polynomial model, the average accuracy was 85,0 %, with a standard deviation of 5,9 %. The composite model was more difficult to cross validate, but the individual models for each cancer type reported similar standard deviations, indicating stable model performance, regardless of the training/test division.

Generally, the most important features (across models) were type of cancer, TNM (both clinical and pathological), and diagnostic extension. This is an indicator that the models are performing as expected, as these are the same factors that influence a clinician's diagnosis.

With respect to the new variables that were created, in the random forest model, time between diagnosis and treatment appears as an important feature with the same weight as some categories from the TNM system.

However, the variable of the sequence of treatment, that is, whether they received surgery as a first intervention, did not score in the top 20 of most important features (out of 679 one-hot encoded variables). Nevertheless, including it improves model performance by 1-2 %. Including the age of the patient generally decreased the performance of the models. This may be due to the one-hot encoding scheme used, since age is inherently an ordinal value. However, no correlation was seen between age and stage at diagnosis during exploratory data analysis, so age was ultimately not included in the model. Additionally, variables such as diagnosis date and treatment dates were not included in the model, as there was expected to be no predictive value.

DISCUSSION

The use of predictive models in the treatment and diagnosis of cancer based on stage can offer many benefits. Given the burgeoning complexity of diagnostic and prognostic information there is no realistic alternative to incorporating multiple variables into a single prediction model. As such, the question should not be whether but how prediction models should be used to aid decision making.[10,11,12,13]

In the case of this analysis, generally, the random forest, SVM, and composite models perform fairly well and reliably. They do not differ significantly, exhibiting similar accuracy rates across different stages consistently. The results of our predictive models of cancer stage indicate that the predictions are good for Stage I and IV, which is consistent with the published literature. However, Stages II and III are not as good, suggesting that these stages are harder to distinguish. There are several possible reasons for this, including differences in the clinical presentation of cancer, variability in diagnostic criteria used by different clinicians, and lack of clear classification for these intermediate stages.

Removing the date variable from the predictive model can be essential to ensure that the model does not misassociate cancer detection dates with the patient's stage. In many cases, cancer screening dates can be skewed by several factors, such as the frequency of medical examinations, the ability of the health care system to perform tests, and the detection of asymptomatic cancers. If these dates are included in the predictive model, there may be an incorrect association between the stage of the cancer and the date of detection. In addition, removing the date variable from the predictive model can help ensure patient privacy. The inclusion of specific dates can make it easier for third parties to identify the patient, which can have legal and ethical implications.

Finally, although these models can provide a useful tool for clinicians and patients in making informed decisions, the results must be interpreted carefully. When predicting stage, there is a possibility for overprediction or underprediction. Overprediction can result in more aggressive and unnecessary medical care, while underprediction can result in a lack of treatment or inadequate treatment. To mitigate these risks in practice, it is important to consider the clinical context and use additional information available, such as diagnostic images and laboratory test results.

Another risk is that a predictive model can generate overly simplistic results that do not consider the individual complexities of each patient. If you rely too much on a predictive model, you can overlook important information and make decisions that are not optimal for an individual patient. Finally, with this dataset, another limitation is the small data set, especially when broken down by different types of cancer. There are not enough examples of non-stage 1 thyroid cancer, so we specified the model only for the 4 most common cancers in the data. Challenges like this can make accurate predictions across cancer types challenging, increasing the need to further build out the dataset.

CONCLUSIONS

A predictive model for cancer stage provides multiple advantages to any management team that works with cancer cases, especially a comprehensive cancer center. From an individual patient perspective, predicting this data point can inform clinician care plans. Stage is helpful for understanding a patient's prognosis. Using this information, the clinician and patient can decide how best to move forward. This will improve institution's ability to meet the needs of its beneficiaries. From a broader organization perspective, this predictive model provides a valuable resource management and prioritization tool. By predicting the stage of cancer, this model can provide critical information for planning and resource allocation in the health system, which can help ensure that patients receive the right treatment at the right time. Over time, there is potential to extend the use of this model beyond the scope of the initial cancer center that provided the dataset and leverage other related datasets such as hospital databases to add variables into the model and improve its performance while also scaling the advantages of the model to a broader audience. The model discussed above provides a helpful tool that clinicians and administrative personnel can use to increase their effectiveness in taking care of Chile's oncology patients. Furthering this work can have positive impacts for the broader Chilean health system.

BIBLIOGRAPHIC REFERENCES

1. The International Agency for Research on Cancer (IARC). Iarc.fr. Global Cancer Observatory. https://gco.iarc.fr/

2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49. http://dx.doi.org/10.3322/caac.21660

3. Inastrilla CRA. Big Data in Health Information Systems. Seminars in Medical Writing and Education 2022;1:6–6. https://doi.org/10.56294/mw20226

4. Memorial Sloan Kettering Cancer Center. Types of cancer treatments. https://www.mskcc.org/cancer-care/diagnosis-treatment/cancer-treatments

5. The Breast Cancer Risk Assessment Tool [Internet]. [cited 2023 May 12]. Breast Cancer Risk Assessment Tool: Online calculator (The Gail Model). https://www.cancer.gov/bcrisktool

6. Shimizu H, Nakayama KI. Artificial intelligence in oncology. Cancer Sci. 2020;111(5):1452–60. http://dx.doi.org/10.1111/cas.14377

7. National Cancer Institute. 2014. Understanding cancer prognosis. https://www.cancer.gov/about-cancer/diagnosis-staging/prognosis

8. Basu A, Ghosh D, Mandal B, Mukherjee P, Maji A. Barriers and explanatory mechanisms in diagnostic delay in four cancers – A health-care disparity? South Asian J Cancer. 2019;08(04):221–5. http://dx.doi.org/10.4103/sajc.sajc_311_18

9. Al-Azri MH. Delay in Cancer Diagnosis: Causes and Possible Solutions. Oman Med J. 2016;31(5):325–6. http://dx.doi.org/10.5001/omj.2016.65

10. National Cancer Institute [Internet]. 2015 [cited 2023 May 5]. Treatment for cancer. https://www.cancer.gov/about-cancer/treatment

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

12. 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

13. Vickers AJ. Prediction models in cancer care. CA Cancer J Clin. 2011. http://dx.doi.org/10.3322/caac.20118

FINANCING

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

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Marcela Aguirre, Sergio Peñafiel.
Data curation: Sergio Peñafiel.
Formal analysis: April Anlage, Emily Brown, Cecilia Enriquez-Chavez.
Research: April Anlage, Emily Brown, Cecilia Enriquez-Chavez.
Methodology: Marcela Aguirre, April Anlage, Emily Brown, Cecilia Enriquez-Chavez.
Project management: Marcela Aguirre.
Resources: Inti Paredes, Sergio Peñafiel.
Software: Sergio Peñafiel.
Supervision: Inti Paredes, Marcela Aguirre.
Validation: Marcela Aguirre, Sergio Peñafiel.
Display: Marcela Aguirre.

Drafting - original draft: April Anlage, Emily Brown, Cecilia Enriquez-Chavez.
Writing - proofreading and editing: Marcela Aguirre.