Calculations of Regional Cancer Estimates Using the Incidence/Mortality Ratio: Tutorial with the Support of R Software Scripts

https://doi.org/10.32635/2176-9745.RBC.2025v71n1.4836

Cálculos de Estimativas de Câncer Regionais pela Razão Incidência/Mortalidade: Tutorial com Apoio de Scripts do Software R

Cálculos de Estimaciones Regionales de Cáncer Utilizando la Razón Incidencia/Mortalidad: Tutorial con el Apoyo de *Scripts* del *Software* R

Gil Patrus Pena[1]; Sara Oliveira Ribeiro[2]

ABSTRACT

Introduction: One of the main roles of cancer surveillance is to provide the manager a scenario of the cancer status, predicting expectations about the incidence of different types of cancer in the territories, enabling rational planning of prevention, early detection and treatment actions for the main types of cancer occurring in that region. The preparation of cancer estimates is one of the main tools in this regard, but their detailed elaboration for different types of cancer and regions is complex and laborious. Software R is a powerful data manipulation tool and its use is still not widespread in health surveillance activities. **Objective:** To describe the methodology for calculating cancer incidence estimates and make available the set of software R scripts used, with the objective of disseminating and promoting its use in cancer surveillance activities. **Method:** Description of the preparation of 2024 Cancer Estimates in Minas Gerais by Health Macro-regions. The entire manipulation and calculation process was structured in R scripts, which are available as a supplement. **Results:** The preparation of estimates is enhanced by the use of software R, facilitating the handling of large amount of data and the necessary calculations, in addition to the possibility of presenting results in graphs and maps. **Conclusion:** With the availability of the scripts, it is expected that the entire process is adjusted by other states to prepare their own regional estimates.

Key words: Public Health Surveillance; Neoplasms/epidemiology; Incidence.

RESUMO

Introdução: Um dos principais papéis da vigilância do câncer é fornecer ao gestor um cenário da situação do câncer, projetando expectativas sobre a incidência dos diferentes tipos de câncer nos territórios, possibilitando o planejamento racional das ações de prevenção, detecção precoce e tratamento direcionados aos principais tipos de câncer incidentes naquela Região. A elaboração de estimativas de câncer constitui uma das principais ferramentas nesse sentido, mas sua elaboração detalhada pelos diferentes tipos de câncer e regiões é complexa e trabalhosa. O *software* R é uma ferramenta poderosa de manipulação de dados e seu uso ainda é pouco difundido nas atividades de vigilância em saúde. **Objetivo:** Descrever a metodologia do cálculo das estimativas de incidência do câncer e disponibilizar o conjunto de *scripts* do *software* R utilizado, com o objetivo de difundir e fomentar o seu uso nas atividades de vigilância do câncer. **Método:** É apresentada a elaboração das Estimativas de Câncer em Minas Gerais, para o ano de 2024, por Macrorregiões de Saúde. Todo o processo de manipulação e cálculos foi estruturado em *scripts* do R, que são disponibilizados como um suplemento. **Resultados:** A elaboração de estimativas é potencializada pelo uso do *software* R, pela facilidade da manipulação da grande quantidade de dados e dos cálculos necessários, além da possibilidade de apresentação dos resultados em gráficos e mapas. **Conclusão:** Com a disponibilização dos *scripts*, espera-se que todo o processo possa ser adaptado por outros Estados, para a elaboração de suas próprias estimativas regionais.

Palavras-chave: Vigilância em Saúde Pública; Neoplasias/epidemiologia; Incidência.

RESUMEN

Introducción: Una de las funciones principales de la vigilancia del cáncer es brindar al gestor un escenario de la situación del cáncer, proyectando expectativas sobre la incidencia de los distintos tipos de cáncer en los territorios, posibilitando la planificación racional de las acciones de prevención, detección temprana y tratamiento dirigidos a los principales tipos de cáncer que ocurren en esa región. La elaboración de estimaciones de cáncer es una de las principales herramientas en este sentido, pero su elaboración detallada para diferentes tipos de cáncer y regiones es compleja y laboriosa. El *software* R es una poderosa herramienta de manipulación de datos y su uso aún no está extendido en las actividades de vigilancia de la salud. **Objetivo:** Describir la metodología para el cálculo de estimaciones de incidencia de cáncer y poner a disposición el conjunto de *scripts* del *software* R utilizados, con el objetivo de difundir y promover su uso en actividades de vigilancia del cáncer. **Método:** Se presenta la elaboración de las Estimaciones de Cáncer en Minas Gerais, para el año 2024, por macrorregiones de salud. Todo el proceso de manipulación y cálculo se estructuró en *scripts* R, que están disponibles como suplemento. **Resultados:** La elaboración de estimaciones se ve favorecida por el uso del *software* R, dada la facilidad para manejar la gran cantidad de datos y los cálculos necesarios, además de la posibilidad de presentar los resultados en gráficos y mapas. **Conclusión:** Con la disponibilidad de los *scripts*, se espera que otros Estados puedan adaptar todo el proceso para preparar sus propias estimaciones regionales.

Palabras clave: Vigilancia en Salud Pública; Neoplasias/epidemiología; Incidencia.

[1] Registro de Câncer de Base Populacional de Belo Horizonte, Coordenação de Vigilância das Doenças e Agravos Crônicos não Transmissíveis e Câncer. Superintendência de Vigilância Epidemiológica, Subsecretaria de Vigilância em Saúde, Secretaria de Estado da Saúde de Minas Gerais. Belo Horizonte (MG), Brasil. E-mail: gil.pena@saude.mg.gov.br. Orcid iD: https://orcid.org/0000-0002-9395-6372

[2] Superintendência de Vigilância Epidemiológica, Subsecretaria de Vigilância em Saúde, Secretaria de Estado da Saúde de Minas Gerais. Belo Horizonte (MG), Brasil. E-mail: sara.oribeiro@yahoo.com.br. Orcid iD: https://orcid.org/0009-0008-1433-3056

Corresponding author: Gil Patrus Pena. Secretaria de Estado da Saúde. Rodovia João Paulo II, 4143, Edifício Minas, 13º andar – Serra Verde. Belo Horizonte (MG), Brasil. CEP 31630-900. E-mail: gil.pena@saude.mg.gov.br

INTRODUCTION

Cancer is a complex disease that encompasses numerous histological types and different topographies. Surveillance actions involve monitoring cancer incidence and mortality. In the Regions not covered by Population-Based Cancer Registries (PBCR), the incidence can be estimated by modeling from mortality information, using incidence/mortality ratios derived from Region-specific PBCR data.

The elaboration of cancer estimates, such as those published periodically by the National Cancer Institute (INCA) for the Federative Units (UF) and capitals of Brazil, for selected types of cancer and by sex, is laborious. For each type of cancer, gender, and region, calculations regarding the population and the number of deaths are necessary to obtain crude and adjusted mortality rates. To obtain adjusted mortality, it is necessary to calculate age-specific mortality rates and apply them to a standard population. It is also necessary to obtain the incidence rates by type of cancer and sex in the selected PBCR[1].

Although the calculations involved in the methodology are based on classical epidemiology, with a relatively accessible methodology, the large amount of data and calculations involved, multiplied for each type of cancer, region, and sex, certainly inhibits estimates initiatives with better regional detailing.

In Minas Gerais, the Cancer and its Risk Factors Surveillance Assessment Program (*Programa de Avaliação de Vigilância do Câncer e seus Fatores de Risco*, PAV-MG) published in 2013 the first estimate for the State Health macroregions[2]. An update of the estimates was fully calculated, but because of a change in the administrative regionalization of healthcare[3], it was not published. Recently, a new amendment to the Master Plan for Health Regionalization was instituted[4], exemplifying the need to have a structured analysis of populations and mortality that can be quickly adapted to changes in the administrative organization of the territory.

The knowledge of the cancer scenario in Minas Gerais, detailed by the health macroregions, is necessary, given the extent and regional diversity, with marked environmental, cultural, and economic variations. This information enables managers to define priorities and plan actions to cope with the disease according to regional particularities.

In addition to the previous estimates carried out under SES-MG[2], studies with cancer estimates for State subregions, calculated by the incidence/mortality ratio method, were carried out for the 17 Regional Health Care Networks (*Redes Regionais de Atenção à Saúde*, RRAS) of the State of São Paulo[5].

Incidence estimates can be produced based on different methods, according to the availability of information in each country. In global estimates by country, the Global Cancer Observatory: Cancer Today[6] (Globocan) applies different methods. The method used primarily involves the projection of PBCR incidence rates with large national or subnational coverage. In the unavailability of incidence rates with comprehensive coverage, national mortality estimates by modeling are used, using incidence/mortality ratios derived from cancer records specific to the Region or neighboring Regions.

In addition to the direct method for calculating the incidence/mortality ratio, statistical methods involving multilevel Poisson models for estimating the incidence/mortality ratio were employed by Jardim et al.[7], making it possible to consider the variation of random effects of PBCR that would not be captured by traditional fixed effects models.

The R software[8] is a tool for data manipulation and statistical analysis. One of the best features that R provides is the ability to keep all data manipulations and analyses performed in a script. Even in the most complex analyses, it was possible to quickly start over and redo after corrections. An organized script allows you to document work routines, process automation, and have task reproducibility[9]. In addition, R is free software, which can be installed without payment of licenses. The use of R is also enhanced by the numerous packages that add specific functions to the program, expanding data manipulation, analysis, and visualization resources (graphs and publications).

This article describes in detail the methodology for elaborating cancer estimates for the 16 health macroregions in the State of Minas Gerais, 21 specific types of cancer, and all neoplasms, for the male and female sexes. The entire course of data preparation and calculations was structured in R scripts that are made available as a supplement and can be adapted for the elaboration of estimates in other territories (Brazil, FU, and regional divisions by FU).

METHOD

All data manipulation was performed in the R software, version 4.3.2[8]. An Intel®Core™ i5-4210U CPU1.70GHz 2.40 GHz processor computer with 8 GB of RAM and 64-bit Windows 10 Pro operating system was used. The following packages with their respective dependencies have been installed: *dplyr*[10], *tidyr*[11], *readxl*[12], *ggplot2*[13], *read.dbc*[14], *car*[15], *gmodels*[16], *lmtest*[17], *rmarkdown*[18], *knitr*[19] and *sf*[20]. Except for the *rmarkdown* (.Rmd) scripts, the analyses were performed on the *Rgui.exe* interface, an executable program made available on the R Base installation that opens a console in a Windows window. The *rmarkdown* (.rmd) scripts were run on the *Rstudio* interface[21]. All scripts were fully prepared for this study, except for the linear regression of mortality rates script that was adapted from a previous version used in the preparation of estimates by INCA [1]. A brief description of the bases used is shown in the supplementary material.

The databases referring to the population were downloaded from the Department of Informatics of the National Health System (DATASUS)[22]. Between 1980 and 2012, information from the Census (1980, 1991, 2000, and 2010), counting (1996), and intercensal projections (1981 to 2012), prepared by the Brazilian Institute of Geography and Statistics (IBGE) and made available by DATASUS, according to age group, sex, and household situation were used. The bases from 1980 to 2012 were downloaded for the whole of Brazil through the Tabwin[22] page. For the years 2013 to 2021, the preliminary estimates prepared by the Ministry of Health/ Health and Environmental Surveillance Secretariat/ Department of Epidemiological Analysis and Surveillance of Non-Communicable Diseases/General Coordination of Information and Epidemiological Analysis (MS/ SVSA/DAENT/CGIAE) were used[23]. The information was tabulated year by year, for each sex, selecting the State of Minas Gerais and setting up municipalities in the line and age group in the column. The information for 2022 corresponds to the Census population and was obtained from IBGE[24]. With the assistance of the R script, the population information organized by the 853 municipalities of Minas Gerais was grouped by Health macroregions, according to the worksheet of the Minas Gerais Regionalization Master Plan (see map in Figure 1 of the supplementary material). In addition to the information by health macroregion, the population information of Belo Horizonte and Poços de Caldas, which are covered by PBCR, was prepared to perform the calculations of cancer incidence. Five-year age groups from 0 to 79 years and over 80 years were employed. A long *data.frame* was created with the columns Macroregion, year, sex, age group, and macroregion population (available as a supplement). Population pyramid graphs were elaborated for each year and macroregion with the assistance of scripts and R software.

Mortality data from 1979 to 2022 from the Mortality Information System (SIM) were downloaded from the Tabwin[22]page. Between 1996 and 2022, the basic causes are codified using the tenth revision of the International Classification of Diseases and Related Health Problems (ICD-10)[25] and, between 1979 and 1995, by the ninth revision (ICD-9)[26]. Deaths without sex information (n=373; 0.06%), age (n=834; 0.14%), and residence (n=546; 0.09%) were not included in the analysis. Deaths from cancer were grouped according to Table 1.

Data from cases classified as uterus, part unspecified (PU) (C55) was incorporated into the analysis in an attempt to produce more reliable rates for cervix (C53) and uterine body (C54) locations. This reallocation is recommended to produce rates that allow significant comparisons between populations. Reallocation methods were based on age and specific distributions of cases[27]. In the complete series, between 1979 and 2022, the proportion of cases "uterus PU (C55)" corresponded to 35.5% of the set of "12-cervix (C53)", "13-uterine body (C54)" and "uterus PU (C55)". When age was under 50 years old, all cases "uterus PU (C55)" were reallocated to "12-cervix (C53)". If age was greater than or equal to 50 years, cases "uterus PU (C55)" were reallocated to "12-cervix (C53)" in the proportion of cases:

$$\frac{\text{"12} - \text{Cervix (C53)"}}{\text{“12} - \text{Cervix (C53)”} + \text{“13} - \text{Uterine body (C54)”}}$$

for each macroregion and year;

and "for 13-uterine body (C54)", in the proportion of:

$$1 - \frac{\text{“12} - \text{cervix (C53)”}}{\text{“12} - \text{cervix (C53)”} + \text{“13} - \text{uterine body (C54)”}}$$

for each macroregion and year.

With the population and number of deaths information, specific mortality rates for neoplasm, age, and sex, crude rates and direct method-adjusted rates were calculated according to the world population (Table 1 of the supplementary material).

The specific mortality rate by age group (a) was calculated for each macroregion (m), neoplasm (n), sex (s) and year (y).

$$Specific\ mortality\ rate_{amnsy} = \frac{number\ of\ deaths_{amnsy}}{population_{amsy}}$$

Subtitles: a = age group; m = macroregion, n = neoplasm group; s = sex; y = year.

This rate was applied to the standard population, resulting in the number of deaths expected in each age group. The sum of the expected number of deaths was then divided by the total of the standard population, to obtain the adjusted rate by age.

$$Adjusted\ mortality\ rate_{mnsy} = \frac{\sum(Specific\ mortality\ rate_{amnsy} \times Standard\ population_{a})}{\sum Standard\ population_{a}}$$

The crude mortality rate was calculated for each macroregion (m), type of cancer (n), sex (s) and year (y):

Table 1. Neoplasm groups according to ICD-9 (1979-1995) and ICD-10 (1996-2022) coding

Neoplasm groups	ICD-9 codes (1979-1995)	ICD-10 codes (1996-2022)
First three digits	140, 141, 142, 144, 145, 146	C00, C01, C02, C03, C04, C05, C06, C07, C08, C09, C10
01-Oral cavity (C00-C10)	140, 141, 142, 144, 145, 146	C00, C01, C02, C03, C04, C05, C06, C07, C08, C09, C10
02-Esophagus (C15)	150	C15
03-Stomach (C16)	151	C16
04-Colon and rectum (C18-C21)	153, 154	C18, C19, C20, C21
05-Liver and intrahepatic bile ducts (C22)	155	C22
06-Pancreas (C25)	157	C25
07-Larynx (C32)	161	C32
08-Trachea, bronchi and lung (C33-C34)	162	C33, C34
09-Malignant skin melanoma (C43)	172	C43
10-Other malignant skin neoplasms (C44)	173	C44
11-Female breast (C50)	174	C50 and female sex
12-Cervix (C53)	180	C53
13-Uterine body (C54)	182	C54
Uterus PU (C55)*	179	C55
14-Ovary (C56)	183	C56
15-Prostate (C61)	185	C61
16-Bladder (C67)	188	C67
17-Central nervous system (C70-C72)	191, 192	C70, C71, C72
18-Thyroid gland (C73)	193	C73
19-Hodgkin lymphoma (C81)	201	C81
20-Non-Hodgkin lymphoma (C82-C85; C96)	200, 202	C82, C83, C84, C85, C96
21-Leukemias (C91-C95)	204, 205, 206, 207	C91, C92, C93, C94, C95
22-Other neoplasms (C00-C97; D46)	147, 148, 149, 152, 156, 158, 159, 160, 163, 164, 165, 170, 171, 175, 176, 177, 178, 181, 184, 186, 187, 189, 190, 194, 195, 196, 197, 198, 203, 238	C11, C12, C13, C14, C17, C23, C24, C26, C27, C28, C29, C30, C31, C35, C36, C37, C38, C39, C40, C41, C45, C46, C47, C48, C49, C51, C52, C57, C58, C59, C60, C62, C63, C64, C65, C66, C68, C69, C74, C75, C76, C77, C78, C79, C80, C86, C88, C90, C97, D46

Caption: (*) The location in uterus, part unspecified (PU) was reallocated to the cervical and uterine body locations[27].

$$Crude\ mortality\ rate_{mnsy} = \frac{\sum number\ of\ deaths_{amnsy}}{\sum population_{amsy}}$$

Crude and adjusted mortality rates for each type of cancer, sex, and macroregion, between 1979 and 2022, were submitted to linear regression, placing the year as an independent numerical variable. The script for the regressions was developed in *rmarkdown,* adapted from a script previously used in the estimates provided by INCA. The results of the regressions, with the respective coefficients of determination (r^2, a model adjustment measure), were analyzed in a Word document (.docx), along with dispersion graphs of the crude mortality rate *versus* year of death, of distribution of residues in relation to adjusted values and the crude rate, and a normal QQ chart (for each neoplasm, sex, and macroregion). If the regression was considered adequate (value of r^2=> 0.7 and random distribution of the residues), the rates calculated from the equation obtained with the regression were used. In other cases, the average rate (crude or adjusted) of the last five years was chosen (2018-2022).

To calculate the population projected for the macroregions, the population projections for the State of Minas Gerais as a whole were used, by age and gender (projection of the population of Brazil and FU by sex and age for the period 2010-2060, edition 2018, produced by IBGE and available at DATASUS)[23]. The population calculation for each macroregion was done according to IBGE's methodology[28], considering the trend of population growth between the 2010 and 2022[28] census:

$$Population_m = a_a \times Population_{MG} + b_a$$
$$a_a = \frac{Population_{m2022} - Population_{m2010}}{Population_{MG2022} - Population_{MG2010}}$$
$$b_a = Population_{m2010} - a_a \times Population_{MG2010}$$

Where:
Population_{m} = projected population of the macroregion
Population_{MG} = estimated population for Minas Gerais
$\text{Population}_{m2022}$ = macro population in the 2022 census
$\text{Population}_{MG2022}$ = MG population in the 2022 census
$\text{Population}_{m2010}$ = macro population in the 2010 census
$\text{Population}_{MG2010}$ = MG population in the 2010 census

In the INCA tabulator[29], the number of cancer cases per age group was tabulated, according to ICD-10, for each sex, in the PBCR of Belo Horizonte (information from 2015 to 2019) and Poços de Caldas (information from 2010 to 2014). The populations covered by the PBCR of Belo Horizonte (2,315,560 inhabitants) and Poços de Caldas (163,742 inhabitants) represent 12.1% of the population of the State (20,539,989 inhabitants), considering the results of the 2022 Census. The cases reported as C55 (uterus PU) were reallocated to codes C53 and C54, following the same methodology used for deaths, justifying the obtainment of the age groups information. The mortality incidence ratio was calculated for Belo Horizonte and Poços de Caldas according to the methodology used by INCA[1] for each type of neoplasm (subscript *n* in the equations) and sex (subscripts in the equations). No adjustments were provided for cases in which the incidence/mortality ratio resulted in a value lower than 1.

$$Ratio\frac{incidence}{mortality}ns = \frac{\dfrac{Number\ of\ new\ cases_{ns}}{\sqrt{Population}}}{\dfrac{Number\ of\ deaths_{ns}}{\sqrt{Population}}}$$

The incidence estimates were made using the mean incidence/mortality ratio calculated for the PBCR of Poços de Caldas and Belo Horizonte. This ratio was applied to crude and adjusted mortality rates estimated by regression for 2024, for each sex (indicated by subscript s), macroregion (subscript m), and neoplasm (subscript n). When the regression was not adequate, the mean mortality rates (crude and adjusted) between 2018 and 2022 (five years) were used.

$$Estimated\ incidence\ rate_{mns2024} = Mortality\ rate_{mns2024} \times Ratio\frac{incidence}{mortality}ns$$

Absolute values for the number of new cases were calculated by applying the estimated crude rate to the 2024 projected population, for each macroregion and gender.

$$Number\ of\ new\ cases_{mns2024} = \frac{Estimated\ incidence\ rate_{mns2024} \times Population_{ms2024}}{100{,}000}$$

Following the INCA estimation methodology[1], the results of the number of new cases are only presented if greater than or equal to 20. Absolute values were rounded to multiples of 10.

The estimated adjusted incidence rates were used for spatial representation based on the distributions of rates per quartile.

As the whole calculation methodology was structured in scripts, it was possible to replicate it for both sexes, for all neoplasms (including or not non-melanoma skin cancers) and the State as a whole. Thus, the number of cases estimated for both sexes does not necessarily correspond exactly to the arithmetic sum of the number of cases estimated for the male and female sexes. Similarly, the estimated case numbers for all neoplasms and all neoplasms except non-melanoma skin cancer, and for the whole State, do not necessarily correspond to the arithmetic sums of the number of cases estimated for each neoplasm (including or not non-melanoma skin cancer) or for each macroregion, respectively.

The consolidation of the results for publication was done in a *rmarkdown* script, presenting in text the values of the estimated cases for Minas Gerais as a whole and of the five macroregions with the highest crude rates, for each sex. The estimated rate results and the number of new cases for each of the State's macroregions are presented in a table (shown in Chart 2 of the supplementary material). Maps were also prepared with adjusted incidence rates by sex, based on distribution by quartile. In this article, only as a demonstration of the methodology used, information regarding all neoplasms, except non-melanoma skin cancers, is presented, also considering the importance of this information in the State planning of high complexity healthcare.

RESULTS

The population pyramids for the State macroregions show a large variation in the population contingent among the health macroregions (Figure 2 of the supplementary material). In the different State macroregions, the aging of the population is observed, with change in the graph shape between 1980 and 2022.

To illustrate the functioning of the script, the results obtained for all neoplasms, except non-melanoma skin cancer, for the State of Minas Gerais (female, male, and both sexes) are presented. In Figure 1, crude and adjusted mortality rates are found, between 1979 and 2022, for male, female, and both sexes.

The results of the rates projected by linear regression, the values of the r^2 for each regression and the average rate for the last five years are presented in Table 2.

For the calculations of incidence and mortality incidence ratio, information from the last five years available in the PBCR of Poços de Caldas (2010-2014) and Belo Horizonte (2015-2019) was used. The results of the calculations for all neoplasms, except for non-melanoma skin cancer, are presented in Table 3.

The incidence/mortality ratio calculated for the populations of Belo Horizonte and Poços de Caldas was applied on the mortality rates projected for 2024, allowing to estimate the crude and adjusted incidence rates for the other territories. The estimated crude rates, in turn, applied to the population of the territory result in the number of new cases estimated for that population. An example of these calculations is shown in Table 4.

The last step in the preparation of the estimates is the publication of the results. The *rmarkdown* package enables the combination of analysis results with text elements, charts, tables, and maps, with the preparation of a document (Word, *.pdf* or *html* format). Chart 1 of the supplementary material shows a snippet of the document obtained by running the script.

The prepared document also includes tables and maps. An example of a table produced in *rmarkdown* can be found in Chart 2 of the supplementary material. The spatial representation of the adjusted incidence rates estimated for the macroregions of the State, organized by the distribution in quartiles, is illustrated in Figure 2 of the supplementary material.

DISCUSSION

In the actions of health surveillance, one can no longer abdicate the resources of management and analysis of information, such as those made available by the R software and its packages. Mastering these technologies is essential to maintain the current status of situation analyses, incorporate the greatest amount of information into this study, and use statistical methods appropriate to each analysis. In addition, R software and its packages contribute a lot to the presentation of results, in the elaboration of graphs, maps, and even text, as was done in this study.

Unlike point-and-click software, learning R software is not intuitive, requiring study and practice. It is a free tool (free software) capable of handling an amount of data far beyond the 1,048,576 rows available in an Excel spreadsheet. In addition, all manipulation and analysis structured in scripts enables the whole process to be redone whenever necessary, with relative simplicity, enabling information update. As an example, the restructuring of regionalization can be quickly incorporated into the analyses.

The sharing of scripts is another advantage, allowing the analysis developed for Minas Gerais to be adapted to other states and to Brazil as a whole. The scripts used in this study are available (see link in the supplementary material), with the expectation that they are used and improved.

A previous experience in sharing R scripts, for preparing the PBCR databases for submission to an international study, facilitated the participation of Brazilian registries in the study. The study direction considered the strategy useful and will make scripts adapted to PBCR from other countries available[30]. Based on such experiences, it is believed that health surveillance actions can be enhanced with the use of R software, far beyond the elaboration of estimates.

Whenever a new updated version of R is installed, the packages must also be updated. When a package is not upgraded to the new version, it can leave the official repository in the Comprehensive R Archive Network (CRAN) and no longer be easily accessible to users. As a strategy to deal with this situation, it is possible to keep old versions with the packages previously installed.

The elaboration of R scripts involves the creation of strategies for solving problems in data management and analysis, in many aspects limited by the user's knowledge of the multiple potentialities provided by the tool. In this sense, many aspects can be improved and small imperfections that might exist can be corrected.

Currently, DATASUS/Tabwin is a great ally of the Health Surveillance System in providing information. Better information cohesion, however, would be desirable. In the population information, for example, it was possible to download the complete bases in Tabwin from 1980 to 2012, make tabulations in DATASUS of population estimates between 2013 and 2021, and download the information from the 2022 Census on the IBGE website. Although the files are well documented and the sources of information preserved, it would be desirable to have all data available in the same format and the same directory.

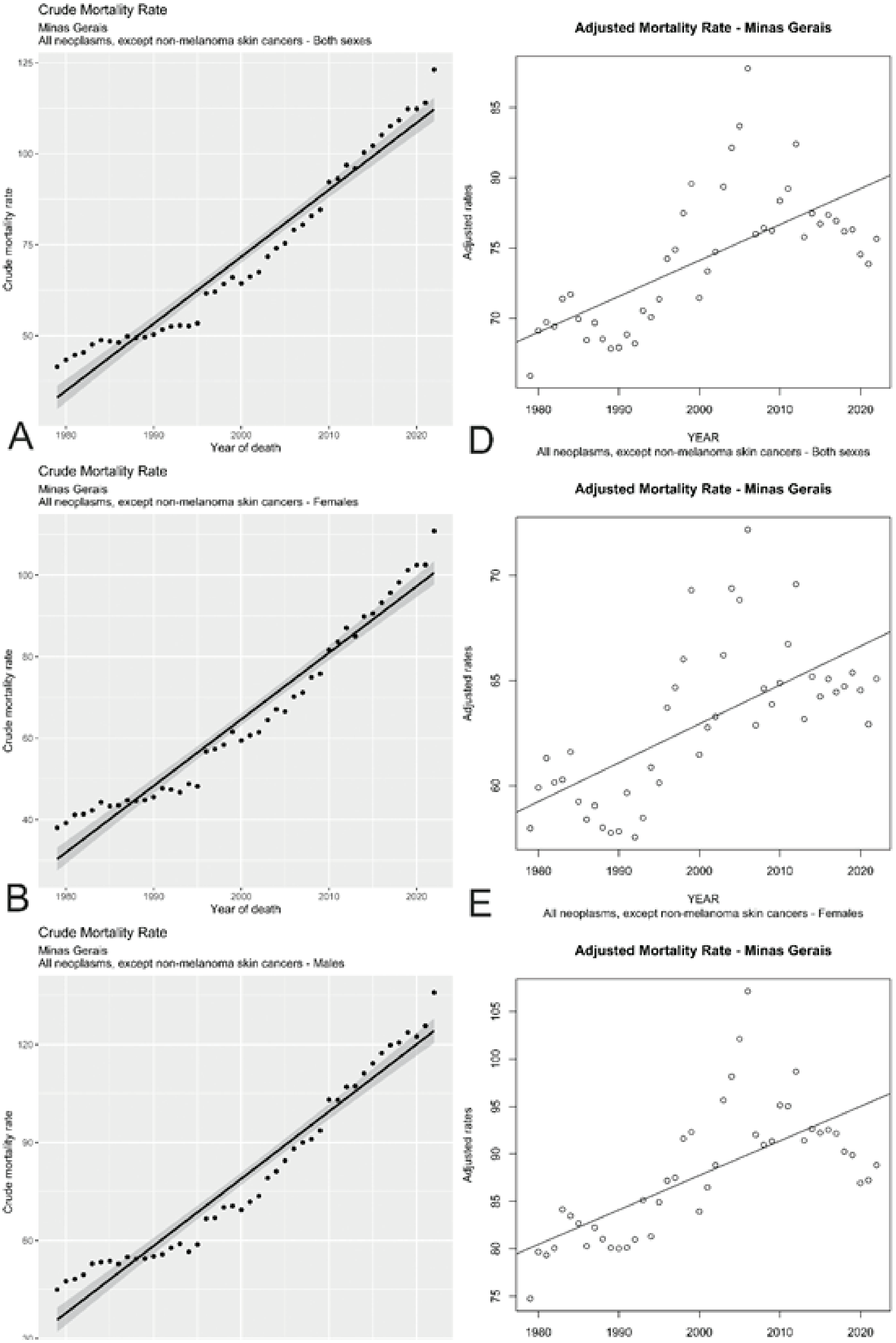

Figure 1. Crude (A, B, C) and adjusted (D, E, F) mortality rates for all neoplasms, except non-melanoma skin cancers in the State of Minas Gerais, for both sexes (A, D), female (B, E) and male (C, F) sexes. The straight lines indicate the estimated trend by linear regression. Y axes with different scales between the graphs

Table 2. Projected mortality rates for the year 2024, for all neoplasms, except non-melanoma skin cancers, using the linear regression line equation, with the r^2 measure. Mean rate of the last five years used as an alternative, in cases where regression was not considered satisfactory

Rate type	Sex	Rate projected by regression	r^2	Mean rate last five years	Rate chosen for calculation
Crude	Both	115.85	0.95	114.12	115.85
Crude	Female	103.82	0.95	103.04	103.82
Crude	Male	128.29	0.95	125.59	128.29
Adjusted	Both	80.25	0.45	75.28	75.28
Adjusted	Female	67.37	0.42	64.52	64.52
Adjusted	Male	96.47	0.47	88.58	88.58

Table 3. Calculation of the incidence/mortality ratio for all neoplasms, except non-melanoma skin cancers

Sex		Belo Horizonte (2015-2019)			Poços de Caldas (2010-2014)		
		Deaths	Cases	Population	Deaths	Cases	Population
Both	Absolute values	16,550	36,613	12,461,005	928	2,149	781,295
	I/M ratio	2.39			2.31		
	Median	2.35					
Female	Absolute values	8,417	20,503	6,627,012	410	978	403,729
	I/M ratio	2.43			2.38		
	Median	2.41					
Male	Absolute values	8,133	19,110	5,833,993	518	1,171	377,566
	I/M ratio	2.35			2.26		
	Median	2.31					

Caption: I/M = Incidence/mortality.

Table 4. Calculation of estimated crude and adjusted rates for all neoplasms, except non-melanoma skin cancers in Minas Gerais, in 2024

Rate type	Sex	Projected mortality rate (2024)	Incidence/ mortality ratio	Estimated incidence rate (2024)	Estimated population (2024)	Estimated new cases (2024)	Rounding
Crude rate	Both	115.85	2.35	272.79	21,737,475	59,298.43	59,300
	Female	103.82	2.41	250.28	11,037,396	27,624.84	27,620
	Male	128.29	2.31	295.72	10,700,079	31,642.01	31,640
Adjusted rate	Both	75.28	2.35	177.26			
	Female	64.52	2.41	155.54			
	Male	88.58	2.31	204.19			

However, there is no retrospective projection available for 1979 (year that starts the SIM series), leading us to use the population of 1980 also for 1979. There are currently no population projections from 2023 onwards distributed by municipality and age groups, so it is necessary to carry out calculations based on the population projected for the State of Minas Gerais.

Incidence estimates should be understood as estimates, not real-world data. Thus, some variability in the results is expected, not only by the incorporation of new bases but also by the results of the regression and projection analyses used. In the analyses carried out in this study, for example, the mortality data for the year 2022 and population data for the 2022 Census are incorporated, unlike the projections calculated by INCA for 2023[1], which used data until 2020. In comparison with the projections elaborated by INCA [1], the present estimate projects a greater number of cases, mainly for males – 52,090 cases in the projection of INCA, 59,300 in this estimate (Table 2 of the supplementary material). Most of this difference can be attributed to the number of estimated cases for prostate cancer – 7,970 new cases estimated by INCA and 11,800 new cases in this estimate. Specifically in this case, the INCA methodology uses the median of the Geographic Region to avoid the inflation of the incidence rate produced by the increase in the neoplasm screening through the advent of the prostate-specific antigen (PSA) test[1].

This comparison only aims to demonstrate the consistency and viability of the methodology proposed in the article, not establishing a direct comparison with other estimates.

The calculation of the estimates has limitations that may be related to the quality of incidence information, mortality, and the population itself. Regarding the incidence, the bases of the PBCR of Poços de Caldas were analyzed and approved for publication in Cancer Incidence in Five Continents, volumes XI and XII, respectively. A quality criterion for information in the registries is the occurrence of an incidence/mortality ratio lower than 1, indicating possible underreporting of cases, if death certificates are accurate and the incidence and survival of neoplasm are constant. Considering the median ratios used to calculate the estimates, this situation occurred for the liver and intrahepatic bile duct neoplasms (male, 0.97; female, 0.81 and both sexes, 0.93) and pancreas (female, 0.89; and both sexes, 0.94), known to be high lethality and short survival neoplasms. With the decision not to propose any adjustment in the incidence/mortality ratio, some underestimation may occur for these neoplasms. Taking the extreme example of liver and intrahepatic bile duct neoplasm, in females, if the ratio had been adjusted to 1,420 new cases would be expected in Minas Gerais, instead of 340 new cases calculated without this adjustment. Part of the deaths reported for these neoplasms may represent cases detected already in the terminal phase, without time to confirm the clinical diagnosis, making it difficult to include them in the Cancer Registry.

The linear regression model was used for the projection of mortality rates, applied to a long time series (1979-2022). This approach does not capture changes in the trend of growth or decrease in rates over time. This difficulty was circumvented with the use of the average rate of the last five years, in cases where the adjustment of the regression model was not considered adequate.

The population information must also be qualified; the results of the 2022 Census revealed a divergence between the previous population projections and the effective population count, indicating the need to reconcile this information. For this study, specifically, no adjustment was made in the population information provided by IBGE (Census of 1980, 1991, 2000, 2010, and 2022; count of 1996; and intercensal projections from 1981 to 2012) or by the Ministry of Health (preliminary estimates from 2013 to 2021). A possible under-enumeration effect of the population of the 2022 Census would be an increase in the mortality rate for this year. The population projections used for 2024, however, are based on information up to 2018 and do not incorporate data from the 2022 census.

In the information on mortality, the reallocation of the cases notified with uterus PU followed the methodology adopted in the estimates elaborated by INCA. Unlike the study by Jardim et al.[7], other codes of ill-defined malignant neoplasm were not reallocated (C26.0, C26.8, C26.9, C57.8, C57.9, C76.0, C76.2, C76.3, C78-C79, and C80) which, in the present study, were included in the other neoplasms group. Considering only the mortality between 1996 and 2022, the group of ill-defined neoplasms of the digestive tract (C26.0, C26.8, and C26.9) represented 6.9% of the set of deaths due to neoplasms C15-C26, which may lead to a small underestimation of cases attributed to these specific target groups.

In State planning, the health macroregions constitute the territorial basis for planning tertiary healthcare that encompasses health microregions with a population of around 700 thousand inhabitants and that offers to its population hospital health services of higher technological density[4]. Still, there are macroregions with relatively small population, such as the Jequitinhonha macroregion, with 385,593 inhabitants in 2022, leading to higher rate fluctuation, especially for lower frequency cancers. This is a limitation for calculating estimates of neoplasms with lower disease burden. There is, however, no other form of division into fewer groups. There was an attempt to circumvent this limitation with the decision to elaborate estimates for both sexes, which may be appropriate in the case of tumors with no clear predominance in one sex. The number of new cases was also omitted, when the estimate resulted in a number lower than 20. As cancer care tends to be more and more specialized, it may be appropriate to group correlated neoplasms according to the trend in oncology specialization (hematological tumors, digestive tract tumors, female genital system, etc.) as a subsidy for planning the organization of the healthcare network. On the other hand, neoplasms with specific characteristics regarding prevention or early detection (for example, cervical cancer, breast cancer, colorectal cancer) should be evaluated individually.

Brazil has a long tradition in the elaboration of cancer estimates, published regularly since the mid-1990s. Given the continental dimensions of the country, with only part of the population covered by the PBCR[31], the estimates play a key role in dimensioning the magnitude and impact of cancer in the different regions of the country. The quality of the estimates is highly dependent on the quality of the information used in its calculation, including the incidence information obtained from PBCR, mortality (from SIM) and the population itself. The challenge of providing information on the incidence of cancer in State Sub-Regions can provide the manager with subsidies to plan, monitor and evaluate cancer control actions and assist in the management of State cancer care networks both in the prevention aspect and early diagnosis and treatment as recommended by the National Cancer Prevention and Control Policy under the Brazilian National Health System (SUS) and the National Cancer Diagnosis Program[32].

CONCLUSION

In this study, the whole methodological path for the elaboration of cancer estimates for the State of Minas Gerais and its Health macroregions was presented, and all scripts used in the preparation of the information, necessary calculations and presentation of the results were made available as a supplement. The expectation is that this article can be used as a tutorial, along with the scripts provided, for the elaboration of estimates in other territories and populations, allowing to detail the distribution of different types of cancer at the intrastate level, where adequate information of incidence is available, that is, where existing PBCR information can be used, with satisfactory quality of information and coverage, specific to the Region or neighboring Regions.

ACKNOWLEDGMENTS

To the team of INCA's Surveillance and Situation Analysis Division, of the Prevention and Surveillance Coordination (*Coordenação de Prevenção e Vigilância* – Conprev), Rio de Janeiro (RJ), Brazil, in the person of Marceli de Oliveira Santos *(in memoriam)*, who contributed greatly to the realization of these regional estimates. To Fernanda Cristina da Silva de Lima, who contributed with a script version of *rmarkdown* to the regression of rates.

CONTRIBUTIONS

Gil Patrus Pena and Sara Oliveira Ribeiro have substantially contributed to the study design, data acquisition, analysis and interpretation, wording, and critical review. Gil Patrus Pena elaborated the R scripts. Both authors approved the final version to be published.

DECLARATION OF CONFLICT OF INTERESTS

There is no conflict of interest to declare.

FUNDING SOURCES

None.

REFERENCES

1. Santos MO, Lima FCS, Martins LFL, et al. Estimativa de incidência de câncer no Brasil, 2023-2025. Rev Bras Cancerol. 2023;69(1):e-213700. doi: https://doi.org/10.32635/2176-9745.RBC.2023v69n1.3700
2. Secretaria de Estado da Saúde (MG). Programa de Avaliação e Vigilância do Câncer e seus fatores de risco (PAV-MG). Situação do câncer em Minas Gerais e suas macrorregiões de saúde: estimativas de incidência e mortalidade para o ano 2013, válidas para 2014: perfil da mortalidade: perfil da assistência na alta complexidade. Belo Horizonte: SES-MG; 2013. v. 1.
3. Secretaria de Estado de Saúde (MG). Subsecretaria de Gestão Regional. Ajuste do Plano Diretor de Regionalização de Saúde de Minas Gerais (PDR/MG) [Internet]. 1. ed. Belo Horizonte: SES-MG; 2020. [Acesso 2024 jul 7]. Disponível em: https://www.saude.mg.gov.br/images/1_noticias/06_2023/2-jul-ago-set/regionalizacao/1-PDR%202020.pdf
4. Secretaria de Estado de Saúde (MG) [Internet]. Belo Horizonte: SEM-MG; 2023. Plano Diretor de Regionalização, 2023 ago 17. [Acesso 2024 jul 7]. Disponível em: https://www.saude.mg.gov.br/gestor/regionalizacao
5. Luizaga CTM, Buchalla CM. Estimativa da incidência de câncer no Estado de São Paulo, Brasil, a partir de dados reais. Cad Saúde Pública, 2023;39(2):e00134222. doi: https://doi.org/10.1590/0102-311XPT134222
6. Ferlay J, Ervik M, Lam F, et al. Global Cancer Observatory: cancer today [Internet]. Lyon: International Agency for Research on Cancer; 2024 [acesso 2024 out 9]. Disponível em: https://gco.iarc.who.int/today/en/data-sources-methods
7. Jardim BC, Junger WL, Daumas RP, et al. Estimativa de incidência de câncer no Brasil e regiões em 2018: aspectos metodológicos. Cad Saúde Pública. 2024;40(6):e00131623. doi: https://doi.org/10.1590/0102-311XPT131623
8. R: The R Project for Statistical Computing [Internet]. Versão 4.3.2 Viena: The R foundation. 2021 nov 2 - [acesso 2022 set 6]. Disponível em: https://www.r-project.org
9. Lemes V, Baccaro FB. Introdução ao uso do programa R [Internet]. Manaus: Instituto Nacional de Pesquisas da Amazônia, 2011. [acesso 2024 out 9]. Disponível em: https://www.researchgate.net/profile/Victor-Landeiro/publication/275035302_Introducao_ao_uso_do_

programa_R/links/553041070cf20ea0a06f67ca/Introducao-ao-uso-do-programa-R.pdf

10. Wickham H, François R, Henry L, et al. dplyr: a grammar of data manipulation [Internet]. Versão 1.1.3. [sem local]: The R foundation. 2023 nov 17. [acesso 2024 out 9]. Disponível em: https://CRAN.R-project.org/package=dplyr
11. Wickham H, Vaughan D, Girlich M. Tidyr: Tidy Messy Data [Internet]. Versão 1.3.1, [sem local]: The R foundation. 2024 jan 1. [acesso 2024 out 9]. Disponível em: https://CRAN.R-project.org/package=tidyr
12. Wickham H, Bryan J. readxl: Read Excel Files [Internet]. Versão 1.4.3, [sem local]: The R foundation. 2023 nov 17. [acesso 2024 out 9]. Disponível em: https://CRAN.R-project.org/package=readxl
13. Wickham H. ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag; 2016.
14. Petruzalek D. read.dbc: Read Data Stored in DBC (Compressed DBF) Files [Internet]. Versão 1.0.6. [sem local]: The R foundation. 2024 jul 6. [acesso 2024 out 9]. Disponível em: https://CRAN.R-project.org/package=read.dbc
15. Fox J, Weisberg S. An R companion to applied regression [Internet]. 3 ed. Thousand Oaks: Sage; 2023. [acesso 2024 out 9]. Disponível em: https://www.john-fox.ca/Companion/
16. Warnes GR, Bolker B, Lumley TL, et al. Package 'gmodels', 2024. Versão 1.0.6. [sem local]: The R foundation. 2024. doi: https://doi.org/10.32614/CRAN.package.gmodels
17. Zeileis A, Hothorn T (2002). "Diagnostic Checking in Regression Relationships" [Internet]. R News [Internet]. 2002[acesso 2024 jul 17];2(3):7-10. Disponível em https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf
18. Allaire J, Xie Y, Dervieux C, et al. rmarkdown: Dynamic Documents for R [Internet]. Versão 2.25. São Francisco: GitHub, Inc.; 2023. [Acesso 2024 jul 17]. Disponível em https://github.com/rstudio/rmarkdown
19. Xie Y. knitr: A General-Purpose Package for Dynamic Report Generation in R. R package [Internet]. version 1.44. [Acesso 2024 jul 17]. Disponível em https://yihui.org/knitr/
20. Pebesma E. "Simple features for R: standardized support for spatial vector data." R Journal. 2018;10(1):439-46. doi: https://doi.org/10.32614/RJ-2018-009
21. RStudio [Internet]. Versão 2023.12.1+402. Boston: Posit Software, PBC. 2024 abr 1 - [acesso 2024 mar 1]. Disponível em: http://www.rstudio.com/ide
22. TABWIN [Internet]. Brasília (DF): DATASUS. c2008 – [acesso 2024 jul 28]. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/
23. TABNET [Internet]. Brasília (DF): DATASUS. c2008. [acesso 2024 jul 28]. Disponível em: http://tabnet.datasus.gov.br/cgi/deftohtm.exe?ibge/cnv/popsvsbr.def
24. SIDRA: Banco de Tabelas Estatísticas [Internet]. Rio de Janeiro: IBGE; [sem data]. [acesso 2024 jul 28]. Disponível em: https://sidra.ibge.gov.br/tabela/9514
25. Organização Mundial da Saúde. CID-10: Classificação Estatística Internacional de Doenças e problemas relacionados à saúde. São Paulo: Edusp; 2008.
26. Organização Mundial da Saúde. CID-9: Classificação Estatística Internacional de Doenças: manual de lesões e causas de óbito. São Paulo: Centro Brasileiro para Classificação de Doenças em Português; 1979.
27. Loos AH, Bray F, McCarron P, et al. Sheep and goats: separating cervix and corpus uteri from imprecisely coded uterine cancer deaths, for studies of geographical and temporal variations in mortality. Eur J Cancer. 2004;40(18):2794-803. doi: https://doi.org/10.1016/j.ejca.2004.09.007
28. Instituto Brasileiro de Geografia e Estatística. Estimativas da população 2021 nota metodológica nº 1. Rio de Janeiro: IBGE; 2021
29. Registro de Câncer de Base Populacional [Internet]. Rio de Janeiro: INCA. [2012] – [acesso 2024 jul 20]. Disponível em: https://www.inca.gov.br/BasePopIncidencias/Home.action
30. Pena Gil P, Pongnikorn D, Khan Baloch F, et al. Sharing R scripts to re-code cancer registry data for international research. In: 12º Congresso Brasileiro de Epidemiologia; 2024 nov 23-27; Rio de Janeiro. Rio de Janeiro: Abrasco; 2024.
31. Pena GPM, Santos ATC, Pezzotti CM, et al. Registros de câncer de base populacional no Brasil: relevância, desafios e oportunidades. Rev Bras Cancerol. 2025;71(1):e-104878. doi: https://doi.org/10.32635/2176-9745.RBC.2025v71n1.4878
32. Presidência da República (BR). Lei nº. 14.758 de 19 de dezembro de 2023. Institui a Política Nacional de Prevenção e Controle do Câncer no âmbito do Sistema Único de Saúde (SUS) e o Programa Nacional de Navegação da Pessoa com Diagnóstico de Câncer; e altera a Lei nº 8.080, de 19 de setembro de 1990 (Lei Orgânica da Saúde). Diário Oficial da União, Brasília, DF. 2023 dez 20; Seção 1.

Recebido em 29/7/2024
Aprovado em 11/11/2024

Associate-editor: Jeane Tomazelli. Orcid iD: https://orcid.org/0000-0002-2472-3444
Scientific-editor: Anke Bergmann. Orcid iD: https://orcid.org/0000-0002-1972-8777