


Article

A Novel Approach to Modeling and Forecasting Cancer Incidence and Mortality Rates through Web Queries and Automated Forecasting Algorithms: Evidence from Romania

Cristiana Tudor 

International Business and Economics Department, The Bucharest University of Economic Studies, 010374 Bucharest, Romania; cristiana.tudor@net.ase.ro

Simple Summary: Cancer remains a global burden, currently causing nearly one in six deaths worldwide. Accurate projections of cancer incidence and mortality are needed for effective and efficient policymaking, accurate resource allocation, and to assess the impact of newly introduced policies and measures. However, the COVID-19 pandemic disrupted public health systems and caused a significant number of cancers to remain undiagnosed, thus affecting the quality of official statistics and their usefulness for health studies. This paper addresses this issue by proposing novel cancer incidence/cancer mortality models based on population web-search habits and historical links with official health variables. The models are empirically estimated using data from one of the most vulnerable European Union (EU) members, Romania, a country that consistently reports lower survival rates than the EU average, and are further used to forecast cancer incidence and mortality rates in the country. Research findings have important policy implications, and the novel framework, owing to its generalizability, can be applied to the same task in other countries. Overall, the results indicate a continuation of the increasing trends in cancer incidence and mortality in Romania and thus underline the urgency to change the status quo in the Romanian public-health system.



Citation: Tudor, C. A Novel Approach to Modeling and Forecasting Cancer Incidence and Mortality Rates through Web Queries and Automated Forecasting Algorithms: Evidence from Romania. *Biology* **2022**, *11*, 857. <https://doi.org/10.3390/biology11060857>

Academic Editors: Shibiao Wan, Yiping Fan, Chunjie Jiang and Shengli Li

Received: 9 April 2022

Accepted: 30 May 2022

Published: 3 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Cancer remains a leading cause of worldwide mortality and is a growing, multifaceted global burden. As a result, cancer prevention and cancer mortality reduction are counted among the most pressing public health issues of the twenty-first century. In turn, accurate projections of cancer incidence and mortality rates are paramount for robust policymaking, aimed at creating efficient and inclusive public health systems and also for establishing a baseline to assess the impact of newly introduced public health measures. Within the European Union (EU), Romania consistently reports higher mortality from all types of cancer than the EU average, caused by an inefficient and underfinanced public health system and lower economic development that in turn have created the phenomenon of “oncotourism”. This paper aims to develop novel cancer incidence/cancer mortality models based on historical links between incidence and mortality occurrence as reflected in official statistics and population web-search habits. Subsequently, it employs estimates of the web query index to produce forecasts of cancer incidence and mortality rates in Romania. Various statistical and machine-learning models—the autoregressive integrated moving average model (ARIMA), the Exponential Smoothing State Space Model with Box-Cox Transformation, ARMA Errors, Trend, and Seasonal Components (TBATS), and a feed-forward neural network nonlinear autoregression model, or NNAR—are estimated through automated algorithms to assess in-sample fit and out-of-sample forecasting accuracy for web-query volume data. Forecasts are produced with the overperforming model in the out-of-sample context (i.e., NNAR) and fed into the novel incidence/mortality models. Results indicate a continuation of the increasing trends in cancer incidence and mortality in Romania by 2026, with projected levels for the age-standardized total cancer incidence of 313.8 and the age-standardized mortality rate of 233.8 representing an increase of 2%, and, respectively, 3% relative to the 2019 levels. Research findings thus indicate that, under the no-change hypothesis, cancer will remain a significant burden in Romania and highlight the need and urgency to improve the status quo in the Romanian public health system.

Keywords: cancer; incidence; mortality; modeling; forecasting; Google Trends; Romania; ARIMA; TBATS; NNAR

1. Introduction

Cancer remains a primary cause of death worldwide [1] and acknowledged as a growing global burden [2]. Moreover, many healthcare systems in less developed countries are ill-equipped to adequately deal with this burden, and a huge percentage of cancer patients worldwide lack access to timely, high-quality diagnosis and treatment [3]. As of 2020, cancer accounted for approximately 10 million deaths worldwide, or nearly one in six deaths. Furthermore, cancer maintains its place as the second leading cause of death in many nations, trailing only cardiovascular disease [3–6]. Additionally, the number of cancer diagnoses and fatalities is expected to significantly increase over the next decade, with projections for 2030 indicating 26 million new cancer cases and 17 million cancer deaths per year [7].

Concurrently, cancer is one of the most critical economic and financial burdens that the globe faces today [8]. In the United States alone, national costs of cancer totaled USD 183 billion as of 2015, with projections that include only population growth indicating an increase of 34% by 2030, reaching USD 246 billion [9].

Consequently, with this escalating global burden, cancer prevention and cancer mortality reduction are counted among the most serious public health concerns of the twenty-first century [10]. In particular, the term “primary prevention” refers to measures to reduce the incidence of the disease, whereas “secondary prevention” refers to efforts to diagnose cancer early or to reduce second cancers among cancer survivors [11]. Accurate cancer projections for future time points are paramount for both primary and secondary prevention and are additionally critical for planning future services and resource allocation, as well as establishing and evaluating cancer control programs [12]. However, time series forecasting is a challenging task [13], whereas producing accurate estimates for the future rates of cancer incidence and mortality is additionally complicated due to the short time series available. For example, at the time of the study the Eurostat (i.e., the statistical office of the European Union) database provides statistics for cancer deaths at a European level spanning the period 2011–2018, whereas the World Development Indicators (WDI) database of the World Bank offers data on the mortality rate from cardiovascular disease, cancer, diabetes, or chronic respiratory disease, and thus does not individualize cancer. Additionally, with short series, out-of-sample forecasting accuracy is hard to assess, and time series cross-validation can be difficult to implement [14].

To solve such research obstacles, monitoring health-seeking behavior in the form of public interest indicated by online search queries has emerged as an essential technique for early identification of health problem occurrences throughout certain periods and geographies [15]. This in turn is based on the fact that the internet has grown in importance as a source of health information accessed by the world population [16,17]. As a direct result, Google Trends has become increasingly popular in health and medical research over the past decade [15,18].

Our data confirm the relevance of web searches for highlighting real occurrences of health problems. Thus, [1] indicates that the three most common types of cancer in 2020, in terms of new cases, were breast cancer with 2.26 million cases, lung cancer with 2.21 million cases, and colon and rectum cancer with 1.93 million cases. Concurrently, as reflected in Figure 1, these were the exact web queries related to the term “cancer” over recent years at the world level, confirming that Internet searches are an accurate reflection of health issue incidences.

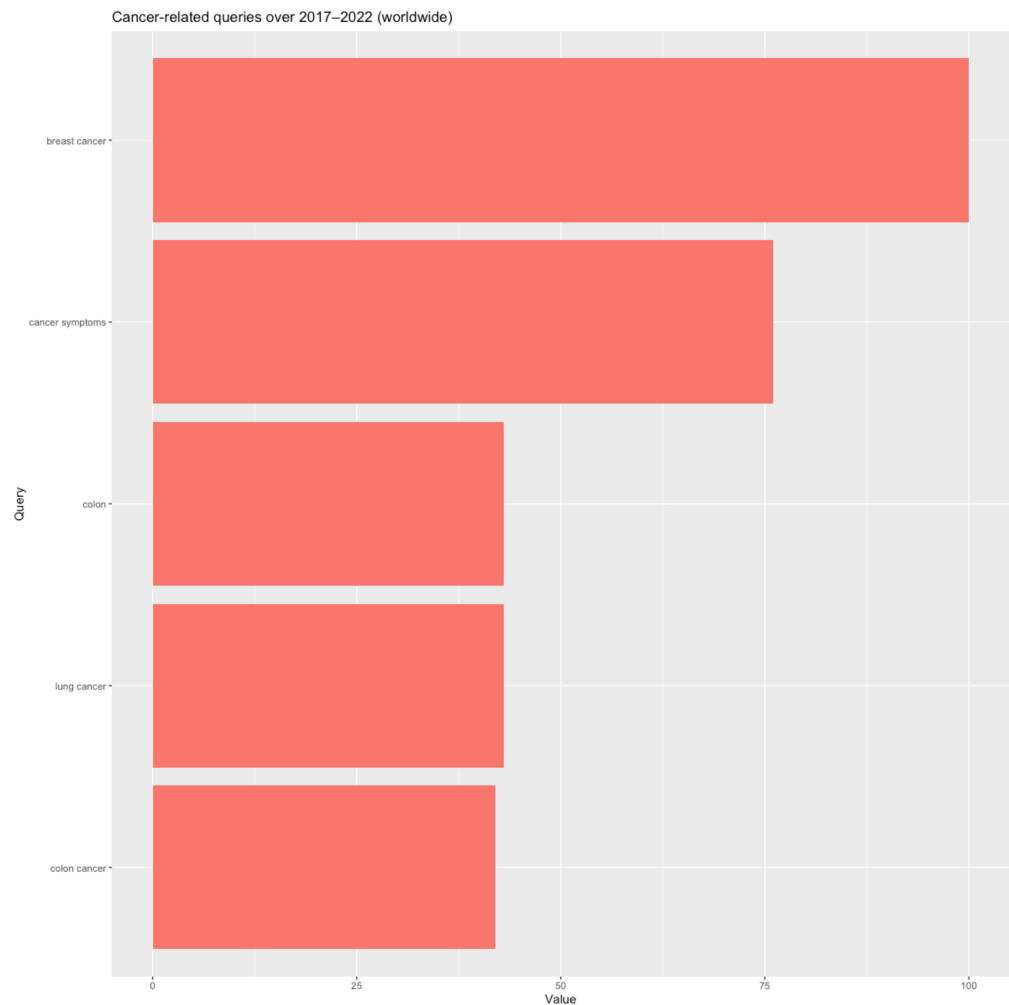


Figure 1. Most common queries related to the search term “cancer”: worldwide (April 2017–March 2022). Source of data: Google Trends. Estimation results using the “gtrendsR” package [19] in R software.

Moreover, a visualization of the global web-search interest reveals that most normalized searches emerged in countries that also reported the highest age-standardized cancer rates. Of note, to accurately comprehend the geography of search interest for a given keyword, the term should be searched across all the world’s languages. However, Google Trends provides a specific tool capable of dealing with this issue, i.e., “Topics,” which collects all related words, variant spellings, and names in other languages under a single label to help with comprehending topics in a multilingual setting. Topics can thus be particularly effective in combining translations into multiple languages under a single subject [20]. As such, we specified the topic “cancer” when sourcing Google Trends data. Thus, Figure 2 reflects the normalized number of internet searches over the most recent five years, confirming that the highest population interest in the topic “cancer” was encountered in countries including Australia, the US, and Ireland. On the other hand, WHO data confirm that Australia registered the world’s highest age-standardized cancer rate at 452.4 cases per 100,000 people in 2020, followed by New Zealand (422.9), Ireland (372.8), and the United States (362.2).

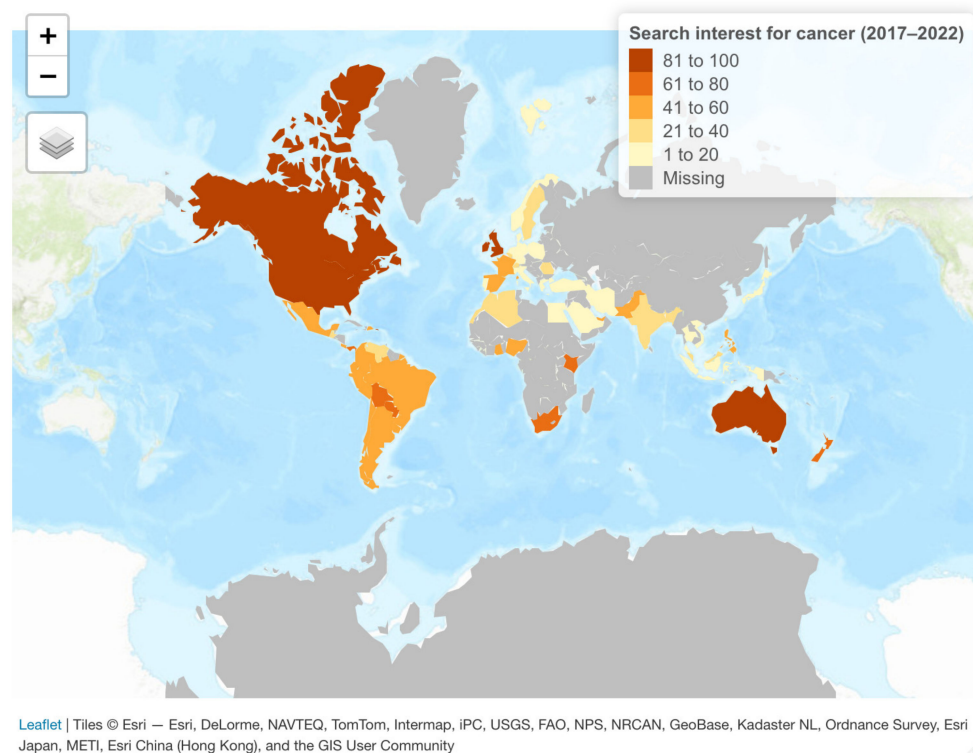


Figure 2. Internet search interest for “cancer” at the world level: (April 2017–March 2022). Source of data: Google Trends. Map is based on estimation results and uses the packages “gtrendsR” [19] and “tmap” [21] in R software.

Additionally, studies increasingly confirm that many cancers remained undiagnosed as a result of healthcare system disruptions caused by the COVID-19 outbreak [22–24]. In this context, with official statistics failing to accurately capture the variation in incidence, people’s search interest for specific symptomatology emerges as the most relevant indication of the health problem’s occurrence.

In light of the above considerations, this study sourced Google Trends data to extract information on internet searches for the word “cancer” and employed it as a proxy to forecast cancer incidence rates. Google Trends (www.trends.google.com, accessed on 30 March 2022) is a web-based tool that shows the popularity of a search phrase in a certain location over time. It provides a time series index of the number of Google queries submitted in a given location. The query weight or share is calculated by dividing the overall query volume for a specific search term within a geographic region by the total number of searches in that region throughout the period in question. Following that, the result is scaled from 0 to 100. As a result, the maximum query share of a search phrase for a given period is normalized to 100, reflecting the point when the search was at its most popular. In conclusion, on a scale of 0–100, Google Trends calculates relative search interest (RSI), with 100 reflecting peak interest [25]. Ref. [26] explore the utility of Google Trends data to examine population web searches for cancer screening and conclude that web queries can capture awareness and interest in cancer screening. Thus, Google Trends data may complement traditional data collection and analysis about cancer screening and related interests, providing important scientific possibilities. However, given the aforementioned disruption of public health systems caused by COVID-19 that altered official statistics, we argue that Google Trends data can now be used as a substitute for traditional statistics, which further expands its scientific value.

Of note, among European countries, cancer survival is significantly lower in newer and less developed EU members from Central and Eastern Europe [27]. Higher death rates at the CEE level are caused by two main factors: delayed diagnosis and suboptimal treat-

ment [28], which are in turn related to inefficient and underfinanced public health systems and lower overall economic development [29]. The situation is particularly challenging in Romania, which continues to register a divergent trend in mortality rates relative to its EU counterparts, including in the CEE area [30]. Figure 3 shows that the age-standardized mortality rate from all cancers follows an increasing trend in Romania from 2011 to 2018, reflecting the inefficiency of the public healthcare system in the country, whereas most other CEE countries have managed to reverse the trend.

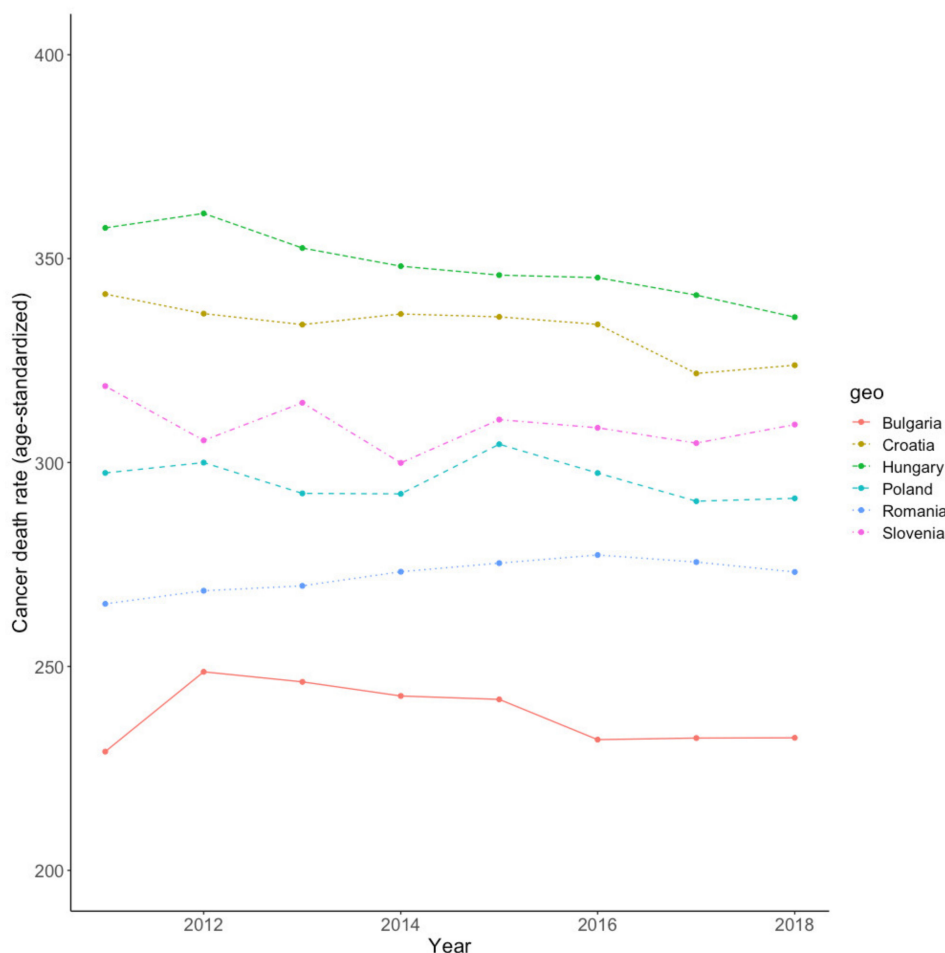


Figure 3. Trends in cancer mortality rates in selected CEE countries (2011–2018). Estimation results. Plot created in R software (“ggplots” function). Source of data: Eurostat.

Moreover, the excess mortality from the main types of cancer registered in Romania relative to the EU average is reflected by the difference in the five-year survival rates presented in Table 1. For example, whereas recent statistics show that the survival rate of breast cancer patients has rapidly increased over recent years, due to the availability of early diagnostic tools and treatment [31], Romania still reports significant health gaps, which is heavily influenced by the fact that there is no organized population screening for breast cancer in the country [30].

Table 1. The five-year survival rates from main types of cancer (Romania versus EU26).

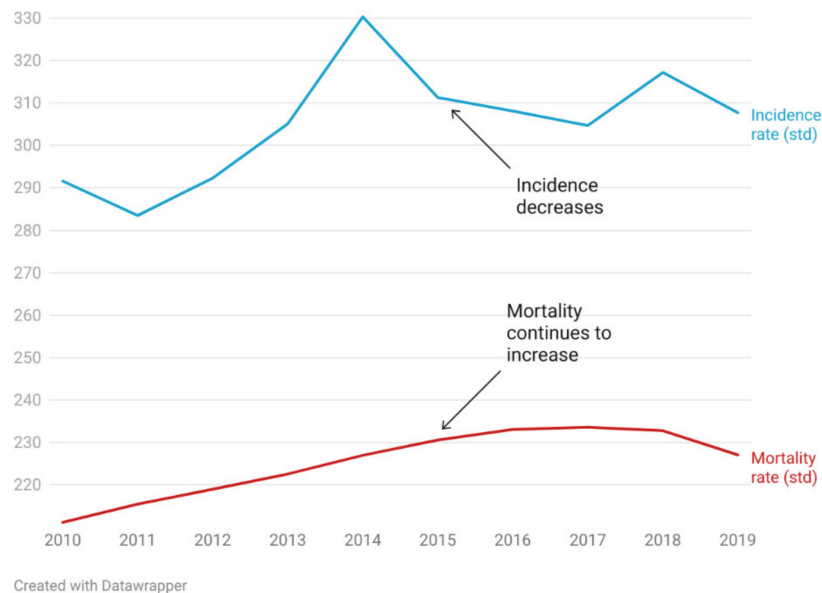
Type of Cancer	5-Year Survival Rate	
	Romania	EU26
Lung	11%	15%
Breast	75%	83%
Prostate	77%	87%

Source of data: Romanian Ministry of Health (2021) [32].

A worrisome disaggregation also occurs between cancer incidence and mortality rates registered in Romania, which has not managed to reduce cancer mortality despite periods of decreased incidence and which further highlights the necessity and urgency of better policies aimed at providing an efficient and inclusive public health system (Figure 4, Panel a). Moreover, the implementation of join-point regression analysis, also known as change-point regression or segmented regression analysis [33] to detect changing trends in cancer incidence (see for example [34–39]) further confirms that the incidence rate in Romania presents two join points in 2014 and 2017, leading to three periods with a different trend over the analysis period, as follows: a positive trend with a slope coefficient of 14.68 until 2014, a negative trend with a slope coefficient of -8.83 during 2014–2017, and a slightly increasing trend with a slope of 2.05 after 2017 (Figure 4, Panel b). Hence, whereas the incidence follows an increasing trend until the first join-point (i.e., 2014), a reversal is detected thereafter, and a decreasing trend is confirmed over the second segment (i.e., 2014–2017). However, the decreasing trend is reversed thereafter, as the incidence rate presents a subsequent rise. On the other hand, the join-point regression analysis found no join-point in the mortality rate, confirming the disaggregation between the two series.

Patient migration from CEE countries has grown with the implementation of a European Union Directive issued in 2011. According to this directive, European nationals are eligible to use European healthcare services in any of the European member states, and their treatments are covered (at least partially) by their home country's health insurance system [40]. As a result, a phenomenon called “oncotourism” or “cancer tourism” has emerged, whereas diagnosed patients move away from inefficient Eastern and Central European public healthcare systems, particularly from Romania, toward the private system or the healthcare systems of more developed EU countries [41].

Trends in incidence and mortality rates (age-standardised, all cancers, Romania)



(a)

Figure 4. Cont.

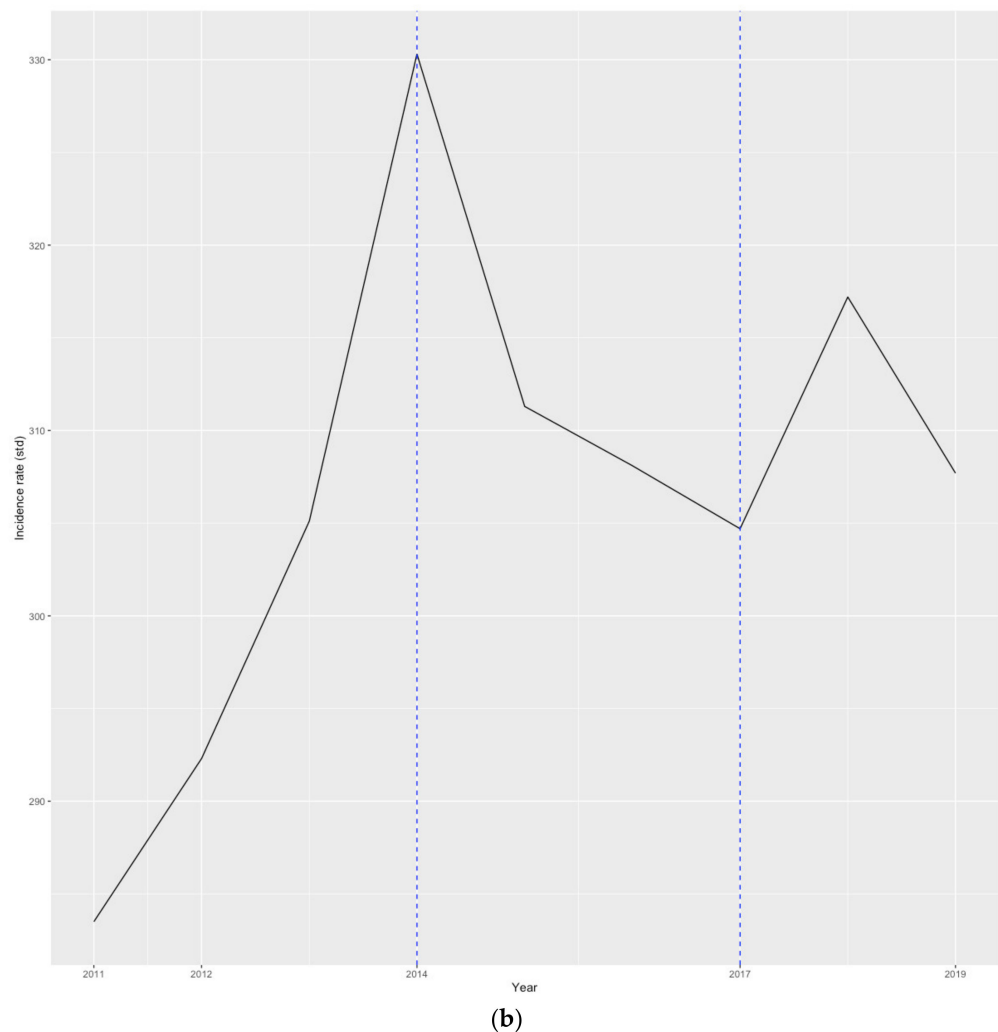


Figure 4. Trends in age-standardized cancer incidence and mortality rates in Romania (2010–2019) (panel a); joint-points in cancer incidence rate (panel b). Source of data: Romanian Ministry of Health (2021) [32]. Chart in panel (a) is produced in Datawrapper. Chart in panel (b) is produced with the “ggplot” function in R software; joint-point regression analysis is performed with the “segmented” package within R software.

Consequently, accurate predictions for cancer incidences are paramount for early detection and for issuing effective and more inclusive public health policies, especially in the most vulnerable EU members in the CEE area. Additionally, cancer incidence projections are also useful for planning health services and establishing a baseline for evaluating the impact of public health measures [42]. Thus, the main goal of this study is to develop cancer incidence/cancer mortality models, and subsequently to make use of web-search data extracted from Google Trends and its point estimates issued through an array of statistical and machine-learning models to ultimately produce accurate forecasts of cancer incidence and mortality rates, while taking a special focus on a vulnerable CEE country significantly plagued by this disease, i.e., Romania. From a methodological perspective, the robustness of results is assured through various approaches, such as: (i) the estimation of alternative predictive models (statistical and machine-learning); (ii) the assessment of the relative out-of-sample forecasting accuracy through the hold-out forecasting technique; (iii) the estimation of the Diebold-Mariano test for superior forecasting accuracy, and (iv) resampling Google Trends data and employing the sampling average for the web-query index.

Of note, the vast majority of previous research either employs one forecasting method or assesses the predictive ability of concurrent methods by estimating forecasting accuracy metrics. This study implements alternative predictive models, both statistical and machine learning, through automated forecasting algorithms. Moreover, the random sampling issue that arises from using Google Trends data is mitigated through resampling and averaging. Additionally, the forecasting results are defended against the Diebold-Mariano (DM) predictive accuracy test. Hence, various robustness checks confirm the reliability of current findings. Furthermore, the strand of literature on cancer research, particularly with a focus on Central and Eastern Europe, remains thin. Hence, whereas most related studies focus on developed countries, the current research contributes to filling the literature void and is thus concerned with a rather under-investigated EU member, Romania, a country that constitutes an interesting playing field for cancer research due to divergent trends relative to its EU counterparts and plagued by the worrisome phenomenon of “oncotourism”. The proposed method is novel and carries the generalizability advantage, being suitable to further investigate other countries for which official statistics have been heavily affected by the coronavirus pandemic.

Thus, compared to previous studies, the contributions of the current research are threefold: (i) we develop two novel models to explain cancer incidence and cancer mortality rates that embed both official statistics and data on population health-seeking behavior as reflected in internet search habits, whereas most previous studies employ some version of the age-period-cohort model (APC) for the same task; (ii) we propose a robust and integrated approach for web query volume forecasting that includes statistical and machine-learning forecasting methods and assures the robustness of results through multiple model calibration on training and test datasets and estimation of multiple accuracy metrics; and (iii) we apply this novel framework to data from one of the most vulnerable EU members, Romania, a country increasingly defined by the phenomenon of “oncotourism”, whereby diagnosed patients avoid the inefficient national public health system. We additionally provide evidence on the link between internet-seeking behavior and the incidence and mortality of the disease in Romania, thus contributing to the extent of infodemiological literature. Research findings have important policy implications, and the framework, owing to its generalizability, can be applied to the same task in other countries. The novel approach is particularly relevant in the aftermath of the COVID-19 pandemic, which has disrupted public health systems and caused a significant number of cancers to remain undiagnosed, thus affecting the quality of official statistics and their usefulness for health studies.

Results overall indicate a continuation of the increasing trends in cancer incidence and mortality in Romania, with a standardized cancer incidence rate of 313.8 by 2026 and a standardized cancer mortality rate of 233.8 by the same horizon, and thus underline the urgency to change the status quo in the Romanian public health system.

The paper continues as follows. Section 2 presents the data used in model development and explains the integrated method. Section 3 describes the empirical findings that emerge from implementing the novel-forecasting framework on the Romanian data. Section 4 discusses the main findings and, finally, Section 5 concludes the study.

2. Materials and Methods

In this study, we sourced annual data on Romanian cancer incidence and mortality rates spanning 2010–2019 from the Romanian Ministry of Health. Next, to develop a model capable of explaining and forecasting these relevant health indicators in the absence of reliable official statistics (which is a worldwide issue caused by the significant number of undetected cancer cases after the onset of the COVID-19 pandemic), we relied on previous infodemiological studies that acknowledge the population’s internet search habits as a reliable proxy for the incidence of a health problem.

The Google Trends platform is a handy tool for determining the popularity of a specific search keyword among a particular demographic. In this study, we extracted the monthly volume of Google queries issued from Romania for “cancer” for the period spanning

January 2005–March 2022. It should be acknowledged that Google Trends implements random sampling and uses only a fraction of the entire search data to construct a search index [43]. Thus, to overcome the sample instability issue, multiple samples (i.e., 12) were sourced and the average of samples was used to construct the web-query index, instead of only one sample (see [44] for relevant details on the sample bias and its correction). However, it has also been recognized that the Google Trends sampling procedure produces reasonably precise estimates, and consequently, there is often no need to employ more than a single sample [45]. The web query time series contained 207 monthly observations.

The relationship between web searches and the health indicators of interest for Romania was assessed through the linear model given by Equation (1).

$$\hat{y} = bX + a \quad (1)$$

where \hat{y} is alternatively the cancer incidence rate, and subsequently the cancer mortality rate, and the independent variable is the web-query index.

Additionally, we used both linear and nonlinear statistical and machine-learning techniques, which allowed us to capture most of the properties of the web-query time series and further contributed to avoiding unreliable forecasts. Predictive models can be delineated into two main categories [46–48]: statistical and machine learning methods (self-learning systems that can learn from data and continuously increase performance), respectively. Thus, in this study, the autoregressive integrated moving average (ARIMA) model (Equation (2)), the Exponential Smoothing State Space Model with Box-Cox Transformation, ARMA Errors, Trend, and Seasonal Components (TBATS) given by Equation (3), and the neural network autoregression (NNAR) model reflected in Equation (4) were alternatively fitted.

An ARIMA(p,d,q)(P,D,Q)s model, first developed by [49], is given by:

$$(1 - \varphi_1 B - \dots - \varphi_p B^p)(1 - \Phi_1 B^s - \dots - \Phi_P B^{sP})(1 - B)^d(1 - B^s)^D Y_t = (1 - \theta_1 B - \dots - \theta_q B^q)(1 - \Theta_1 B^s - \dots - \Theta_P B^{sQ}) \varepsilon_t \quad (2)$$

where s is the seasonal period, the lowercase and the capital letters represent nonseasonal and seasonal parameters, and ε_t is a random variable with mean zero and the standard deviation σ .

A TBATS model [50] can accommodate complex seasonal behaviors of data [51] and is written as:

$$\text{TBATS}(\omega, p, q, \varphi, \{m_1, k_1\}, \{m_2, k_2\}, \dots, \{m_T, k_T\}), \quad (3)$$

where ω is the Box-Cox transformation, k is the number of harmonics used for the seasonal trait, and φ is the dampening parameter.

Artificial neural networks (ANNs) are capable of simulating complicated real-world systems while properly accounting for nonlinearities [52]. Lagged values of time series are frequently utilized as inputs in an ANN structure when fitting time series data, which is then known as neural network autoregression (NNAR) [53] (Munim et al., 2019). As in [29,54], the NNAR model is written as:

$$Y = f(H) = f(W * X + B), X = [y(t-1), y(t-2), \dots, y(t-p)] \quad (4)$$

where Y stands for the output vector, f is the activation function, H is the vector of n nodes in the hidden layer, W is the weight matrix between the input and hidden layers, X is the vector of inputs (i.e., the lagged values of the actual observations), and B is a bias vector.

All estimations are automated and performed in R software via dedicated algorithms included in the “forecast” package [55]. To implement the method robustly, the series of length $N = 207$ was first split into a training set (containing 187 observations) for in-sample fit purposes and a testing set (containing the last 20 observations) on which the models that reported the best fit on the training set were further estimated and their out-of-

sample forecasting ability assessed. Lastly, we assessed the forecast accuracy of alternative predictive models by estimating both scale and scale-free accuracy metrics.

First, let us define the forecast error of a candidate model as:

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T} \quad (5)$$

where $\{y_1, \dots, y_T\}$ is the training set data and $\{y_{T+1}, y_{T+2}, \dots\}$ is the test-set data.

Then, the following forecasting accuracy metrics are computed as:

Mean absolute error:

$$MAE = \sqrt{\frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|} \quad (6)$$

Root mean squared error:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (7)$$

Mean absolute percentage error:

$$MAPE = \text{mean}(|p_t|) \quad (8)$$

where: $p_t = \frac{100e_t}{y_t}$

Mean absolute percentage error:

$$MASE = \text{mean}(|q_j|) \quad (9)$$

where q_j is given by as: $q_j = \frac{e_t}{\frac{1}{N-1} \sum_{i=2}^N |y_i - y_{i-1}|} f$ when the series is non-seasonal and by:

$q_j = \frac{e_t}{\frac{1}{N-m} \sum_{i=m+1}^N |y_i - y_{i-m}|}$ when the time series is seasonal.

Lastly, all predictive models were fitted to the entire series of length N and point forecasts for the web query index for the following 4 years (i.e., a 48-month forecasting horizon) were produced by the overperforming method in the out-of-sample setting. Forecasted values were then fitted into the incidence/mortality models developed by estimating Equation (1), which then issued the expected values for standardized cancer incidence and mortality rates corresponding to the forecasting horizon. Figure 5 reflects the integrated method employed in this study and implemented in R software.

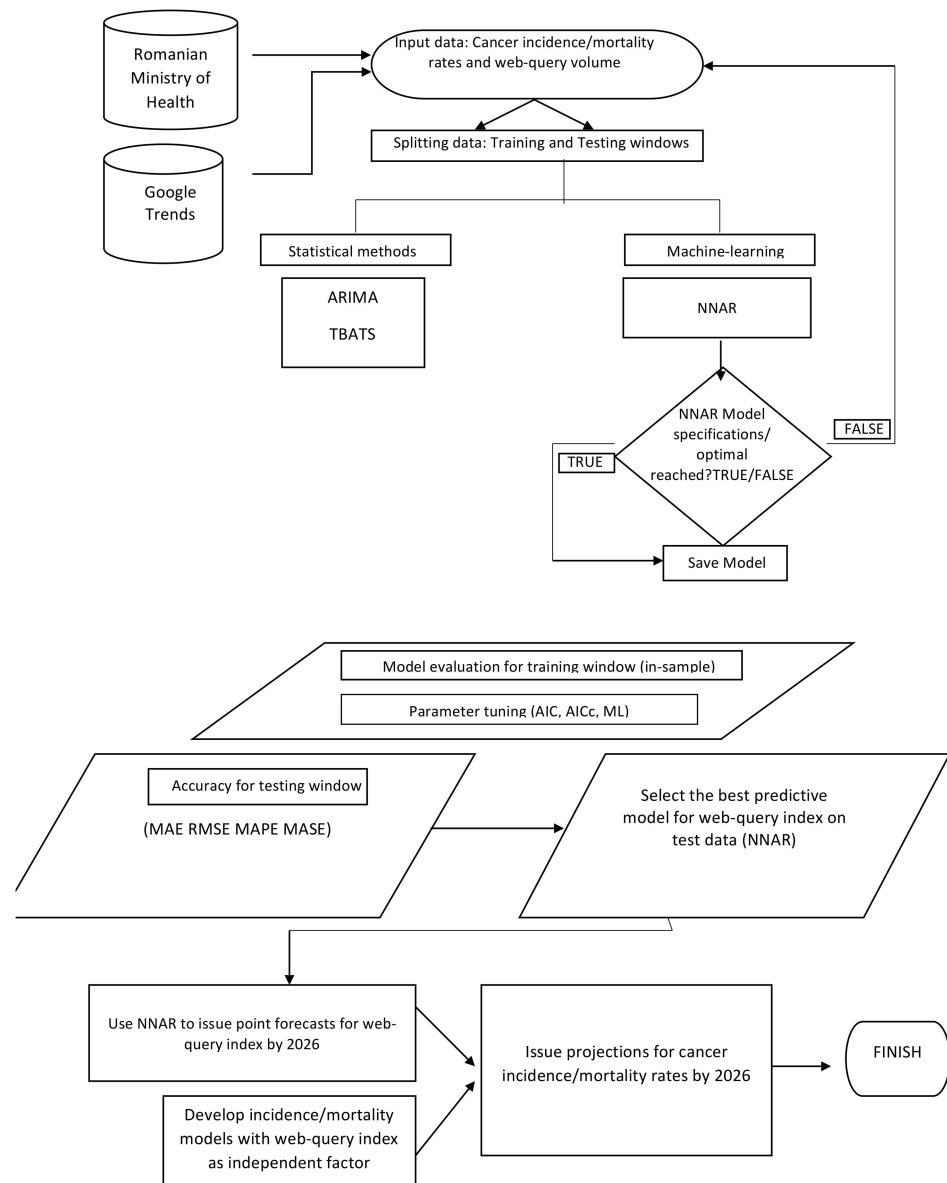


Figure 5. The integrated framework for modeling and forecasting cancer incidence and mortality rates.

3. Results

3.1. Relationship between Related Web Queries and the Age-Standardized Cancer Incidence/Cancer Mortality Rate in Romania

The best-fit linear model between the vector of web-query volume and the cancer incidence/cancer mortality rate is reflected in Figure 6, panels (a) and (b), respectively. Both representations highlight a positive link between the web-search interest and the variables reflecting the incidence and mortality of the disease. Additionally, both equations show a similar slope coefficient, equal to 0.47 in the incidence rate model, and equal to 0.46 in the mortality rate model.

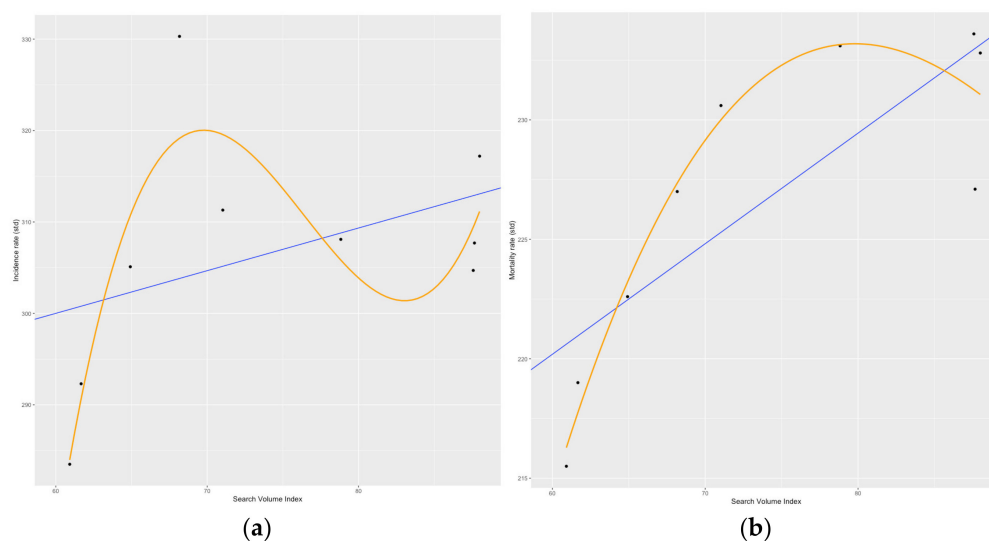


Figure 6. The relationship (linear—blue line, polynomial—orange line) between related web queries and the age-standardized cancer incidence rate in Romania (panel a). The relationship (linear—blue line, polynomial—orange line) between related web queries and the age-standardized cancer mortality rate in Romania (panel b). Source of data: Romanian Ministry of Health [32]. All estimations were performed in R software; plots were created in R software (i.e., “ggplot” function).

3.2. Results from Modeling and Forecasting the Web-Query Index

Table 2 reports the estimated accuracy measures for the test-set data containing topic searches for “cancer” submitted in Romania that were issued through the statistical and machine-learning predictive models. Results indicate that NNAR has been able to accurately capture variations in data and thus provide the best forecast for the web query index over the testing window.

Table 2. Accuracy measures for the out-of-sample (test-set) forecasting performance.

Predictive Model	MAE	RMSE	MAPE	MASE
ARIMA	4.21	5.16	5.76	0.61
NNAR	3.96	4.71	5.32	0.57
TBATS	4.60	5.73	6.54	0.74

To assess the forecasting superiority of the feed-forward neural network autoregression model, we estimated the Diebold-Mariano (DM) test [56,57] to examine any significant differences between forecasts produced by NNAR and the second best-performing model (ARIMA). The DM test result (estimated with the “dm.test” function within the “forecast” package in R software) confirmed that there was a significant difference between the distribution of errors from ARIMA and NNAR, thus ensuring the forecasting superiority of the machine-learning method.

We next employed the best-performing model in terms of out-of-sample forecasting accuracy (i.e., NNAR) to produce the expected web-query volume in Romania for the next 48 months (4 years), corresponding to the period spanning April 2022 to March 2026. Figure 7 reflects the estimation results, showing (in blue color) the point estimates produced by NNAR for April 2022–March 2026. Of note, estimations corresponding to the 48-month forecasting horizon indicated a continuation of the increasing trend in web searches related to “cancer” issued from Romania.

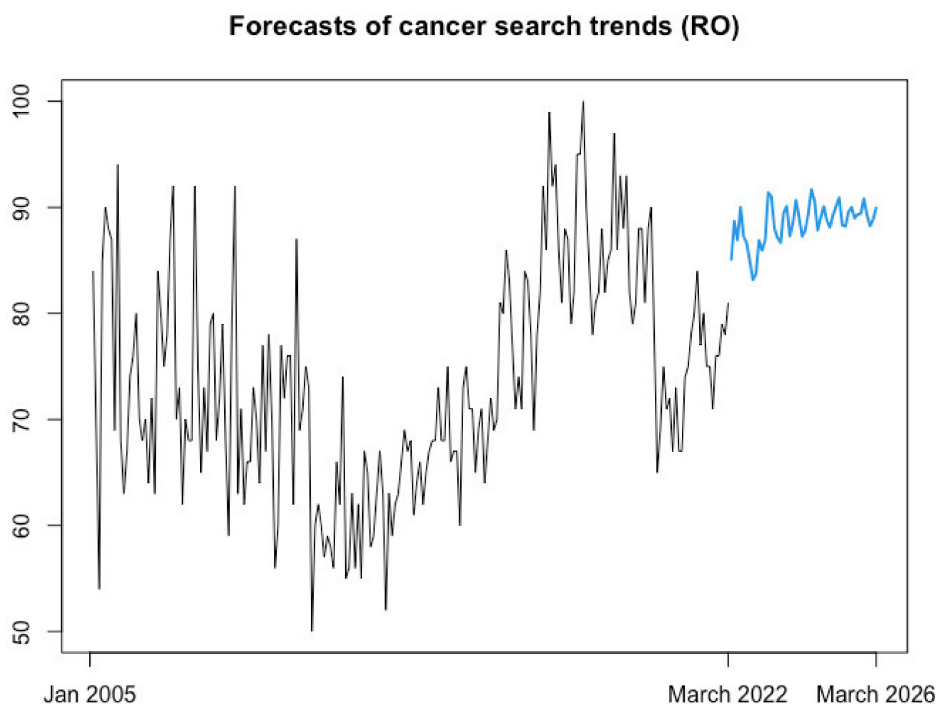


Figure 7. Forecasted trend over April 2022–March 2026 (48 months) for web queries for the term “cancer” in Romania issued with NNAR (12,6). Source: estimation results. Model information: average of 20 networks, each of which is a 12-6-1 network with 85 weight options.

3.3. Forecasts of Cancer Incidence and Cancer Mortality Rates in Romania over 2022–2026

Lastly, we put together the linear relationship model estimated before and the point estimates produced by the neural network autoregression model (NNAR) to estimate the cancer incidence and cancer mortality rate for the next four years. It is important to underline that all estimations are based on the status quo hypothesis, implying that these projections are expected if no changes in public health policy are implemented. Projected values for the two health indicators are centralized in Table 3.

Table 3. Forecasted values for cancer incidence and cancer mortality rates.

Year	Incidence Rate (Projected, Standardized)	Mortality Rate (Projected, Standardized)
2023	308.7	228.8
2024	313.0	233.0
2025	313.6	233.6
2026	313.8	233.8

We notice that estimations issued through the incidence and mortality models and based on NNAR projections of the related web-search index reflect a continuation of the increasing trends in cancer incidence and mortality in Romania, underlining the urgency to change the status quo in the Romanian public health system. Estimates thus indicate a standardized cancer incidence rate of 313.8 by 2026 and a standardized cancer mortality rate of 233.8 by the same horizon, increasing from levels of 307.7 and 227.1, respectively, registered in 2019.

4. Discussion

Approximately 10 million deaths have been attributed to cancer in 2020, or nearly one in six deaths. Concurrently, many cancer patients worldwide still lack access to timely,

high-quality diagnosis and treatment. Within the EU, CEE countries consistently report higher mortality rates, mainly as a result of delayed diagnosis and suboptimal treatment. For Romania, statistics are particularly worrisome, with the country reporting significantly lower survival rates for all main types of cancer and also a divergent trend in mortality rates relative to its EU counterparts that have managed to reverse the increasing trend. Consequently, “oncotourism” is especially characteristic of Romania, as diagnosed patients move away from the inefficient public healthcare system toward the private system or the healthcare systems of more developed EU countries.

Consequently, accurate predictions of cancer incidence and mortality rates are keys to informing policymakers and assisting in the policymaking process. The main goal of this study is to develop a robust model capable of capturing the evolution of cancer incidence and mortality rates and forecasting their evolution. Concurrently, we build on [58] and acknowledge that over the past decade, the use of internet data has become an important aspect of health informatics, with online sources becoming more accessible and offering data that can be used to analyze and forecast human behavior. As a result, we also agree with [59] that data from tracking online information seekers’ behavior is useful in public health surveillance and research.

Thus, in model development, the study made use of web-search data extracted from Google Trends, which was introduced as an independent variable in light of previous studies that acknowledge the population’s internet search habits as a reliable proxy for health problem occurrence. Moreover, this approach overcomes the current issue of the unreliability of official statistics, caused on one hand by the numerous undiagnosed cases after the COVID-19 outbreak and, on the other hand, by the unavailability or tardiness of treatment for diagnosed patients, directly affecting the rate of incidence and mortality. Concurrently, point estimates for the web-query index were issued through the best performing predictive model over the test-set (i.e., out-of-sample), after an assessment of the in-sample fit and out-of-sample forecasting accuracy of various statistical and machine-learning models (ARIMA, TBATS, and NNAR) had been performed via several accuracy metrics (MAE, RMSE, MAPE, and MASE). Estimations indicated that NNAR was the most capable of capturing the time series characteristics and of producing the most accurate estimates. Consequently, estimations for the web-query index were automatically produced with the NNAR model and sourced into the incidence/mortality models that have been previously developed. Ultimately, forecasts of cancer incidence and mortality rates in Romania by 2026 were issued, indicating a continuation of the increasing trend for both variables. Our results are in line with projections of [32] the Romanian Ministry of Health (2021), confirming the ascendant trend, although our point estimates fall below the public ministry’s predictions, indicating more conservative increasing rates.

Future predictions depend on multiple assumptions, most importantly on the status quo (i.e., no-change) hypothesis of the Romanian public health system. Similarly, forecasts do not consider the impact of relevant changes in impact factors, such as potential changes in smoking prevalence at the national level, changes in obesity, changes in alcohol consumption, changes in nutrition habits, increased funding of the public health system, increased screening, HPV vaccination, etc. As a result, we agree with [60] that it is critical to review predictions at regular intervals to incorporate the most recent trends in the data. However, similar to [42], we reason that current estimates do provide a useful baseline for the planning of cancer resources and for evaluating the impact of any changes produced in impact factors as a result of newly introduced public health policies and measures. Furthermore, as with most research, this study suffers from other limitations. Mostly, the use of Google Trends data does carry some vulnerabilities that should be acknowledged, including the construction of the search index itself [61]. As a consequence, the long-run stability of the time series is heavily dependent on the data’s time frame and frequency. As per [61], this study used monthly data that was best able to accurately capture the long-term trend. Moreover, random sampling is an inherent bias in Google Trends data [44]. However, this issue is particularly troublesome in forecasting when dealing with topics that

are less frequently searched for, which is not the case in the current study. Additionally, the performed resampling further mitigated the bias. Nonetheless, it should be acknowledged that this approach merely minimizes the bias and does not eliminate it. Finally, it should be mentioned that results should be interpreted with care, considering that data reflect the search habits of the population that has Internet access, which in turn depends on income level and other socioeconomic factors. Overall, I argue that the popularity of the search topic, together with the resampling strategy and previous studies that reinforce the usefulness of web search data as a powerful predictive instrument [44] does allow for confidence in the current research findings.

In addition, it should also be mentioned that the link between cancer variables and the web-query index has been assessed through the classical regression model. However, it has been increasingly acknowledged that the neutrosophic regression model [62–65], which issues the parameters in the indeterminacy interval range, can be more efficient in the uncertainty environment than the classical regression model [66]. Thus, the assessment of the historical link between cancer incidence/cancer mortality rates and the web search index through the neutrosophic regression model constitutes a good avenue for future research. The implementation of the proposed method on data specific to different gender and age groups, as well as to geographic regions of the country, could also reveal particularly vulnerable groups and/or areas and offer relevant information to policymakers.

From a policy perspective, the findings highlight that cancer will continue to be a significant burden for Romania, which should be carefully planned for. Complementarily, results indicate the need for better policies aimed at mitigating main risk factors such as smoking, alcohol consumption, obesity and overweight, unhealthy nutrition, lack of physical exercise, etc., and at increasing the financing and efficiency of the public health system by allocating future resources for cancer research, treatment, and prevention.

5. Conclusions

In conclusion, this study developed two novel cancer incidence/cancer mortality models based on population web-search habits and historical links with official health variables. The models were empirically estimated using data from one of the most vulnerable European Union (EU) members, Romania, and further used to forecast cancer incidence and mortality rates in the country by employing estimates for the web-search query index issued through the best performing out-of-sample forecasting method (NNAR). Research findings have important policy implications, and the novel framework, owing to its generalizability, can be applied to the same task in other countries. It provides the important advantage of overcoming a current issue related to the quality of official statistics in the aftermath of the COVID-19 pandemic that disrupted public health systems and caused a significant number of cancers to remain undiagnosed. Overall, the results indicate a continuation of the increasing trends in cancer incidence and mortality in Romania and thus underline the urgency to change the status quo in the Romanian public health system.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data employed in this study are publicly available from the Eurostat database, from Google Trends, and the Romanian Ministry of Health.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization (WHO). Cancer. 2022. Available online: <https://www.who.int/news-room/fact-sheets/detail/cancer> (accessed on 4 April 2022).
2. World Cancer Research Fund. Worldwide Cancer Data. 2022. Available online: <https://www.wcrf.org/dietandcancer/worldwide-cancer-data/> (accessed on 30 March 2022).

3. World Health Organization (WHO). Cancer. 2019. Available online: https://www.who.int/health-topics/cancer#tab=tab_1 (accessed on 4 April 2022).
4. Ma, X.; Yu, H. Cancer issue: Global burden of cancer. *Yale J. Biol. Med.* **2006**, *79*, 85.
5. Nagai, H.; Kim, Y.H. Cancer prevention from the perspective of global cancer burden patterns. *J. Thorac. Dis.* **2017**, *9*, 448. [CrossRef]
6. Zaorsky, N.G.; Churilla, T.M.; Egleston, B.L.; Fisher, S.G.; Ridge, J.A.; Horwitz, E.M.; Meyer, J.E. Causes of death among cancer patients. *Ann. Oncol.* **2017**, *28*, 400–407. [CrossRef]
7. Thun, M.J.; DeLancey, J.O.; Center, M.M.; Jemal, A.; Ward, E.M. The global burden of cancer: Priorities for prevention. *Carcinogenesis* **2010**, *31*, 100–110. [CrossRef]
8. World Cancer Day, Financial and Economic Impact of Cancer. 2022. Available online: <https://www.worldcancerday.org/financial-and-economic-impact-0> (accessed on 30 March 2022).
9. Mariotto, A.B.; Enewold, L.; Zhao, J.; Zeruto, C.A.; Yabroff, K.R. Medical care costs associated with cancer survivorship in the United States. *Cancer Epidemiol. Prev. Biomark.* **2020**, *29*, 1304–1312. [CrossRef]
10. United Nations (UN). New WHO Platform Promotes Global Cancer Prevention. 2022. Available online: <https://news.un.org/en/story/2022/02/1111312> (accessed on 7 April 2022).
11. White, M.C.; Peipins, L.A.; Watson, M.; Trivers, K.F.; Holman, D.M.; Rodriguez, J.L. Cancer prevention for the next generation. *J. Adolesc. Health* **2013**, *52*, S1–S7. [CrossRef]
12. Rapiti, E.; Guarnori, S.; Pastoors, B.; Miralbell, R.; Usel, M. Planning for the future: Cancer incidence projections in Switzerland up to 2019. *BMC Public Health* **2014**, *14*, 102. [CrossRef]
13. Petropoulos, F.; Spiliotis, E. The wisdom of the data: Getting the most out of univariate time series forecasting. *Forecasting* **2021**, *3*, 478–497. [CrossRef]
14. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*; Otexts. 2018. Available online: <https://otexts.com/fpp2/> (accessed on 30 March 2022).
15. Szilagyi, I.S.; Ullrich, T.; Lang-Illievich, K.; Klivinyi, C.; Schittek, G.A.; Simonis, H.; Bornemann-Cimenti, H. Google Trends for Pain Search Terms in the World’s Most Populated Regions Before and After the First Recorded COVID-19 Case: Infodemiological Study. *J. Med. Internet Res.* **2021**, *23*, e27214. [CrossRef]
16. Polgreen, P.M.; Chen, Y.; Pennock, D.M.; Nelson, F.D.; Weinstein, R.A. Using internet searches for influenza surveillance. *Clin. Infect. Dis.* **2008**, *47*, 1443–1448. [CrossRef]
17. Pew Research Center, Health Online. 2013. Available online: <http://www.pewinternet.org/2013/01/15/health-online-2013/> (accessed on 7 April 2022).
18. Nuti, S.V.; Wayda, B.; Ranasinghe, I.; Wang, S.; Dreyer, R.P.; Chen, S.I.; Murugiah, K. The use of google trends in health care research: A systematic review. *PLoS ONE* **2014**, *9*, e109583. [CrossRef] [PubMed]
19. Massicotte, P.; Eddebuettel, D. Gtrendsr: Perform and Display Google Trends Queries. R Package Version 1.4.4. 2019. Available online: <https://CRAN.R-project.org/package=gtrendsR> (accessed on 30 March 2022).
20. Forbes. Understanding What You’re Searching for in A Multilingual World. 2015. Available online: <https://www.forbes.com/sites/kalevleetaru/2015/10/18/understanding-what-youre-searching-for-in-a-multilingual-world/?sh=9e2b3f23e0f4> (accessed on 25 May 2022).
21. Tennekens, M. tmap: Thematic Maps in R. *J. Stat. Softw.* **2018**, *84*, 1–39. [CrossRef]
22. Jacob, L.; Loosen, S.H.; Kalder, M.; Luedde, T.; Roderburg, C.; Kostev, K. Impact of the COVID-19 pandemic on cancer diagnoses in general and specialized practices in Germany. *Cancers* **2021**, *13*, 408. [CrossRef] [PubMed]
23. Marques, N.P.; Silveira, D.M.M.; Marques, N.C.T.; Martelli, D.R.B.; Oliveira, E.A.; Martelli-Júnior, H. Cancer diagnosis in Brazil in the COVID-19 era. *Semin. Oncol.* **2021**, *48*, 156–159. [CrossRef]
24. Becker’s Hospital Review, As COVID-19 Dies Down, Undiagnosed Cancers Emerge. 2021. Available online: <https://www.beckershospitalreview.com/oncology/as-covid-19-dies-down-undiagnosed-cancers-emerge.html> (accessed on 6 April 2022).
25. Greiner, B.; Tipton, S.; Nelson, B.; Hartwell, M. Cancer screenings during the COVID-19 pandemic: An analysis of public interest trends. *Curr. Probl. Cancer* **2022**, *46*, 100766. [CrossRef]
26. Schootman, M.; Toor, A.; Cavazos-Rehg, P.; Jeffe, D.B.; McQueen, A.; Eberth, J.; Davidson, N.O. The utility of Google Trends data to examine interest in cancer screening. *BMJ Open* **2015**, *5*, e006678. [CrossRef]
27. Vrdoljak, E.; Wojtukiewicz, M.Z.; Pienkowski, T.; Bodoky, G.; Berzinec, P.; Finek, J.; Todorović, V.; Borojević, N.; Croitoru, A. Cancer epidemiology in Central and South Eastern European countries. *Croat. Med. J.* **2011**, *52*, 478–487. [CrossRef]
28. World Health Organization (WHO). Up to a quarter of Europeans Will Develop Cancer: From Prevention, Early Diagnosis, Screening and Treatment to Palliative Care, Countries Must Do More. 2020. Available online: <https://www.euro.who.int/en/health-topics/noncommunicable-diseases/cancer/news/news/2020/2/up-to-a-quarter-of-europeans-will-develop-cancer-from-prevention,-early-diagnosis,-screening-and-treatment-to-palliative-care,-countries-must-do-more> (accessed on 7 April 2022).
29. Tudor, C.; Sova, R. EU Net-Zero Policy Achievement Assessment in Selected Members through Automated Forecasting Algorithms. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 232. [CrossRef]
30. Furtunescu, F.; Bohiltea, R.E.; Voinea, S.; Georgescu, T.A.; Munteanu, O.; Neacsu, A.; Pop, C.S. Breast cancer mortality gaps in Romanian women compared to the EU after 10 years of accession: Is breast cancer screening a priority for action in Romania? (Review of the Statistics). *Exp. Ther. Med.* **2021**, *21*, 268. [CrossRef]

31. Azam, M.; Aslam, M.; Basharat, J.; Mughal, M.A.; Nadeem, M.S.; Anwar, F. An empirical study on quality of life and related factors of Pakistani breast cancer survivors. *Sci. Rep.* **2021**, *11*, 24391. [CrossRef]
32. Romanian Ministry of Health. Analysis of the Cancer Situation in 2021 (in Romanian). 2021. Available online: <https://www.ms.ro/2021/06/30/cancerul-este-un-risc-pentru-o-forma-severa-de-covid-19-nu-lasa-boala-sa-te-afecteze-si-tu-poti-lua-masuri-si-tu-poti-preveni-si-tu-poti-proteja/> (accessed on 7 April 2022).
33. Gillis, D.; Edwards, B.P. The utility of joinpoint regression for estimating population parameters given changes in population structure. *Heliyon* **2019**, *5*, e02515. [CrossRef]
34. Qiu, D.; Katanoda, K.; Marugame, T.; Sobue, T. A Joinpoint regression analysis of long-term trends in cancer mortality in Japan (1958–2004). *Int. J. Cancer* **2009**, *124*, 443–448. [CrossRef]
35. Crispo, A.; Barba, M.; Malvezzi, M.; Arpino, G.; Grimaldi, M.; Rosso, T.; Esposito, E.; Sergi, D.; Ciliberto, G.; Giordano, A.; et al. Cancer mortality trends between 1988 and 2009 in the metropolitan area of Naples and Caserta, Southern Italy: Results from a joinpoint regression analysis. *Cancer Biol. Ther.* **2013**, *14*, 1113–1122. [CrossRef]
36. Zahmatkesh, B.; Keramat, A.; Alavi, N.; Khosravi, A.; Kousha, A.; Motlagh, A.G.; Darman, M.; Partovipour, E.; Chaman, R. Breast cancer trend in Iran from 2000 to 2009 and prediction till 2020 using a trend analysis method. *Asian Pac. J. Cancer Prev.* **2016**, *17*, 1493–1498. [CrossRef]
37. Sarakarn, P.; Suwanrungruang, K.; Vatanasapt, P.; Wiangnon, S.; Promthet, S.; Jenwitheesuk, K.; Koonmee, S.; Tipsunthonsak, N.; Chen, S.L.S.; Yen, A.M.F.; et al. Joinpoint analysis trends in the incidence of colorectal cancer in Khon Kaen, Thailand (1989–2012). *Asian Pac. J. Cancer Prev. APJCP* **2017**, *18*, 1039.
38. Wilson, L.; Bhatnagar, P.; Townsend, N. Comparing trends in mortality from cardiovascular disease and cancer in the United Kingdom, 1983–2013: Joinpoint regression analysis. *Popul. Health Metr.* **2017**, *15*, 23. [CrossRef]
39. Dragomirescu, I.; Llorca, J.; Gómez-Acebo, I.; Dierssen-Sotos, T. A join point regression analysis of trends in mortality due to osteoporosis in Spain. *Sci. Rep.* **2019**, *9*, 4264. [CrossRef]
40. Atlatszo. 2019. Available online: <https://english.atlatszo.hu/2019/01/28/pay-or-die-onco-tourism-and-corruption-in-romania-and-hungary/> (accessed on 7 April 2022).
41. Investigative Journalism for Europe (IJ4EU). “Cancer Tourism” in Central and Eastern Europe. 2018. Available online: <http://www.investigativejournalismforeu.net/projects/cancer-tourism-in-central-and-eastern-europe/> (accessed on 7 April 2022).
42. Mistry, M.; Parkin, D.M.; Ahmad, A.S.; Sasieni, P. Cancer incidence in the United Kingdom: Projections to the year 2030. *Br. J. Cancer* **2011**, *105*, 1795–1803. [CrossRef]
43. Narita, M.F.; Yin, R. In Search of Information: Use of Google Trends’ Data to Narrow Information Gaps for Low-Income Developing Countries; International Monetary Fund: 2018. Available online: <https://www.elibrary.imf.org/view/journals/001/2018/286/article-A001-en.xml> (accessed on 10 May 2022).
44. Medeiros, M.C.; Pires, H.F. The Proper Use of Google Trends in Forecasting Models. *arXiv* **2021**, arXiv:2104.03065.
45. Stephens-Davidowitz, S.; Varian, H. A Hands-On Guide to Google Data. 2014. Available online: <https://people.ischool.berkeley.edu/~hal/Papers/2015/primer.pdf> (accessed on 20 May 2022).
46. Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **2001**, *16*, 199–231.
47. Charpentier, A.; Flachaire, E.; Ly, A. Econometrics and machine learning. *Econ. Stat.* **2018**, *505*, 147–169. [CrossRef]
48. Tudor, C.; Sova, R. Flexible decision support system for algorithmic trading: Empirical application on crude oil markets. *IEEE Access* **2022**, *10*, 9628–9644. [CrossRef]
49. Box, G.; Jenkins, G. *Time Series Analysis: Forecasting and Control*; Holden-Day: San Francisco, CA, USA, 1970.
50. De Livera, A.M.; Hyndman, R.J.; Snyder, R.D. Forecasting time series with complex seasonal patterns using exponential smoothing. *J. Am. Stat. Assoc.* **2011**, *106*, 1513–1527. [CrossRef]
51. Yu, C.; Xu, C.; Li, Y.; Yao, S.; Bai, Y.; Li, J.; Wang, L.; Wu, W.; Wang, Y. Time series analysis and forecasting of the hand-foot-mouth disease morbidity in China using an advanced exponential smoothing state space TBATS model. *Infect. Drug Resist.* **2021**, *14*, 2809. [CrossRef]
52. Pasini, A. Artificial neural networks for small dataset analysis. *J. Thorac. Dis.* **2015**, *7*, 953.
53. Munim, Z.H.; Shakil, M.H.; Alon, I. Next-day bitcoin price forecast. *J. Risk Financ. Manag.* **2019**, *12*, 103. [CrossRef]
54. Tudor, C.; Sova, R. Benchmarking GHG Emissions Forecasting Models for Global Climate Policy. *Electronics* **2021**, *10*, 3149. [CrossRef]
55. Hyndman, R.J.; Khandakar, Y. Automatic time series forecasting: The forecast package for R. *J. Stat. Softw.* **2008**, *27*, 1–22. [CrossRef]
56. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263.
57. Harvey, D.; Leybourne, S.; Newbold, P. Testing the equality of prediction mean squared errors. *Int. J. Forecast.* **1997**, *13*, 281–291. [CrossRef]
58. Mavragani, A.; Ochoa, G. Google Trends in infodemiology and infoveillance: Methodology framework. *JMIR Public Health Surveill.* **2019**, *5*, e13439. [CrossRef] [PubMed]
59. Dehkordy, S.F.; Carlos, R.C.; Hall, K.S.; Dalton, V.K. Novel data sources for women’s health research: Mapping breast screening online information seeking through Google trends. *Acad. Radiol.* **2014**, *21*, 1172–1176.
60. Smittenaar, C.R.; Petersen, K.A.; Stewart, K.; Moitt, N. Cancer incidence and mortality projections in the UK until 2035. *Br. J. Cancer* **2016**, *115*, 1147–1155. [CrossRef] [PubMed]

61. Eichenauer, V.Z.; Indergand, R.; Martínez, I.Z.; Sax, C. Obtaining consistent time series from Google Trends. *Econ. Inq.* **2022**, *60*, 694–705. [[CrossRef](#)]
62. Smarandache, F. Introduction to Neutrosophic Statistics. Infinite Study. 2014. Available online: <https://arxiv.org/pdf/1406.2000> (accessed on 24 May 2022).
63. Chen, J.; Ye, J.; Du, S.; Yong, R. Expressions of rock joint roughness coefficient using neutrosophic interval statistical numbers. *Symmetry* **2017**, *9*, 123. [[CrossRef](#)]
64. Aslam, M. A new sampling plan using neutrosophic process loss consideration. *Symmetry* **2018**, *10*, 132. [[CrossRef](#)]
65. Aslam, M. Design of sampling plan for exponential distribution under neutrosophic statistical interval method. *IEEE Access* **2018**, *6*, 64153–64158. [[CrossRef](#)]
66. Aslam, M.; Albassam, M. Application of neutrosophic logic to evaluate correlation between prostate cancer mortality and dietary fat assumption. *Symmetry* **2019**, *11*, 330. [[CrossRef](#)]