

Mining Google Trends data for nowcasting and forecasting colorectal cancer (CRC) prevalence

Cristiana Tudor and Robert Aurelian Sova

Bucharest University of Economic Studies, Bucharest, Romania

ABSTRACT

Background: Colorectal cancer (CRC) is the third most prevalent and second most lethal form of cancer in the world. Consequently, CRC cancer prevalence projections are essential for assessing the future burden of the disease, planning resource allocation, and developing service delivery strategies, as well as for grasping the shifting environment of cancer risk factors. However, unlike cancer incidence and mortality rates, national and international agencies do not routinely issue projections for cancer prevalence. Moreover, the limited or even nonexistent cancer statistics for large portions of the world, along with the high heterogeneity among world nations, further complicate the task of producing timely and accurate CRC prevalence projections. In this situation, population interest, as shown by Internet searches, can be very important for improving cancer statistics and, in the long run, for helping cancer research.

Methods: This study aims to model, nowcast and forecast the CRC prevalence at the global level using a three-step framework that incorporates three well-established univariate statistical and machine-learning models. First, data mining is performed to evaluate the relevancy of Google Trends (GT) data as a surrogate for the number of CRC survivors. The results demonstrate that population web-search interest in the term “colonoscopy” is the most reliable indicator to nowcast CRC disease prevalence. Then, various statistical and machine-learning models, including ARIMA, ETS, and FNNAR, are trained and tested using relevant GT time series. Finally, the updated monthly query series spanning 2004–2022 and the best forecasting model in terms of out-of-sample forecasting ability (*i.e.*, the neural network autoregression) are utilized to generate point forecasts up to 2025.

Results: Results show that the number of people with colorectal cancer will continue to rise over the next 24 months. This in turn emphasizes the urgency for public policies aimed at reducing the population's exposure to the principal modifiable risk factors, such as lifestyle and nutrition. In addition, given the major drop in population interest in CRC during the first wave of the COVID-19 pandemic, the findings suggest that public health authorities should implement measures to increase cancer screening rates during pandemics. This in turn would deliver positive externalities, including the mitigation of the global burden and the enhancement of the quality of official statistics.

Submitted 13 March 2023

Accepted 14 July 2023

Published 4 October 2023

Corresponding author

Cristiana Tudor,
cristiana.tudor@net.ase.ro

Academic editor

Ka-Chun Wong

Additional Information and
Declarations can be found on
page 27

DOI 10.7717/peerj-cs.1518

© Copyright

2023 Tudor and Sova

Distributed under

Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Bioinformatics, Computational Biology, Data Mining and Machine Learning, Data Science, Neural Networks

Keywords Automated algorithms, Nowcast, Forecast, Cancer, Google Trends, ARIMA, State space, Neural network autoregression, Public health, Resource allocation

INTRODUCTION

As of 2020, the *World Health Organization (WHO) (2022)* estimated that cancer remained the top cause of death globally, accounting for nearly 10 million fatalities (*Tudor, 2022a*). Yet, cancer is not just a major public health concern but also an economic issue whose burden is growing (*Cancer Atlas, 2022*). This is especially true for low- and middle-income nations, which accounted for 70% of cancer deaths recorded at the global level in 2020 (*American Cancer Society, 2022b*). Unsurprisingly, the UN Sustainable Development Goals (SDGs) (*United Nations, 2022*) include SDG target 3.4, which aims to “by 2030,” reduce by one-third premature mortality from non-communicable diseases (NCDs) through prevention and treatment and promote mental health and well-being (*Kocarnik et al., 2022; Tran et al., 2022*) including the non-communicable disease of cancer.

Colorectal cancer (CRC) is the third most prevalent and second deadliest form of cancer worldwide (*Cervantes et al., 2022; Kadakuntla et al., 2021; Keum & Giovannucci, 2019; Mazidimoradi, Tiznobaik & Salehiniya, 2022; Uhlig et al., 2021; Xie, Chen & Fang, 2020; Xi & Xu, 2021*). Specifically, with 1,9 million new cases in 2020, colorectal cancer ranks third among all diagnosed cancers and accounts for 10.7% of all new cases, as shown in [Table 1](#).

In 2020, colorectal cancer was the third most common cancer in the world for men, accounting for 11.4% of all new cases (*World Cancer Research Fund International (WCRF), 2022*). In contrast, CRC cancer was the second most prevalent cancer worldwide for women in 2020, accounting for 9.9% of new cases (WCRF, 2022). In addition, prior research has shown that CRC is more prevalent in advanced economies, while its incidence is increasing in middle- and low-income nations due to westernization, and early-onset CRC is also becoming more prevalent (*Xi & Xu, 2021*). In addition, new figures reveal that CRC is the second-leading cause of cancer-related mortality worldwide, accounting for nearly 1 million fatalities annually (*International Agency for Research on Cancer (IARC), 2022*).

Effectively addressing this increasing burden globally requires both a thorough understanding and accurate forecasting of its trends. Different from cancer incidence, cancer prevalence is the proportion of individuals within a population who have been diagnosed with cancer at some point in their lives, regardless of the date of diagnosis (*Maddams et al., 2009*). Compared to the average population, this population category, often dubbed ‘cancer survivors’, places a higher strain on the healthcare system (*Capocaccia et al., 2002*). In this context, cancer prevalence projections are key to assessing the future burden of the disease, efficiently planning resource allocation and creating service delivery strategies, and comprehending the changing environment of cancer risk factors (*Bray et al., 2013; Maddams, Utley & Møller, 2012; Smittenaar et al., 2016*). These predictions offer valuable intelligence to both the planners of health service resource allotting bodies and to providers of dedicated care and treatment, but, different from cancer incidence and mortality rates, cancer prevalence projections are not routinely issued by national and international agencies, and thus are not widely available (*Maddams, Utley & Møller, 2012*). The limited or non-existent cancer statistics for large parts of the world, together with the high heterogeneity encountered among world countries (*Thun*

Table 1 Cancer incidence (worldwide, 2020).

Rank	Cancer	New cases (both sexes)
1	Breast	2,261,419
2	Lung	2,206,771
3	Colorectal	1,931,590
Rank	Cancer	New cases (Men)
1	Lung	1,435,943
2	Prostate	1,414,259
3	Colorectal	1,065,960
Rank	Cancer	New cases (Women)
1	Breast	2,261,419
2	Colorectal	865,630
3	Lung	770,828

Note:

Data source: (World Cancer Research Fund International (WCRF), 2022: <https://www.wcrf.org/cancer-trends/worldwide-cancer-data/>).

et al., 2010), further complicate this task. All these factors further contribute to highlighting the importance and need for more research offering accurate cancer prevalence projections at various levels and form the motivators for the current study.

Time series forecasting is a critical and rapidly expanding topic (*Aras & Kocakoç, 2016*), but it remains a challenging task (*Petropoulos & Spiliotis, 2021*) that is further hindered by the aforementioned issues when it comes to producing cancer-related projections. However, to overcome the main pitfalls in cancer prevalence forecasting, over the last decade, the usefulness of Internet-search data for health informatics has been increasingly documented, with Internet resources becoming more accessible and providing data that can be used to assess and forecast human behavior (*Polgreen et al., 2008; Mavragani & Ochoa, 2019; Szilagyi et al., 2021*). Consequently, big data provided by the Google Trends platform has proven valuable in health and medical research (*Nuti et al., 2014; Szilagyi et al., 2021; Tudor & Sova, 2022*), first in infodemiology studies (*Eysenbach, 2011; Bernardo et al., 2013; Mavragani & Ochoa, 2019; Mavragani, 2020; Kamiński, Łoniewski & Marlicz, 2020*), but lately being increasingly employed in cancer-related research (*Schootman et al., 2015; Greiner et al., 2022; Tudor, 2022a*). Consequently, as per *Salathé et al. (2012)* and *Sulyok, Ferenci & Walker (2021)*, among others, online digital data sources have been proven to improve disease surveillance, monitoring, modeling, and forecasting.

Moreover, the COVID-19 pandemic and the related measures imposed by governments worldwide to tackle it have had an impact on all aspects of life, including medical care (*Eftimov et al., 2020; Trinh et al., 2022*). As such, a non-trivial consequence of the pandemic-related restrictions consists of delays in cancer screenings and early diagnostics (*Sharpless, 2020; Bakouny et al., 2021; Greiner et al., 2022; Trinh et al., 2022*), halting of clinical trials, and delays in treatments (*Bakouny et al., 2020; Richards et al., 2020*), thus decreasing the relevancy of official statistics and having negative consequences for cancer projections and overall for cancer research (*Saini et al., 2020*). In turn, this further

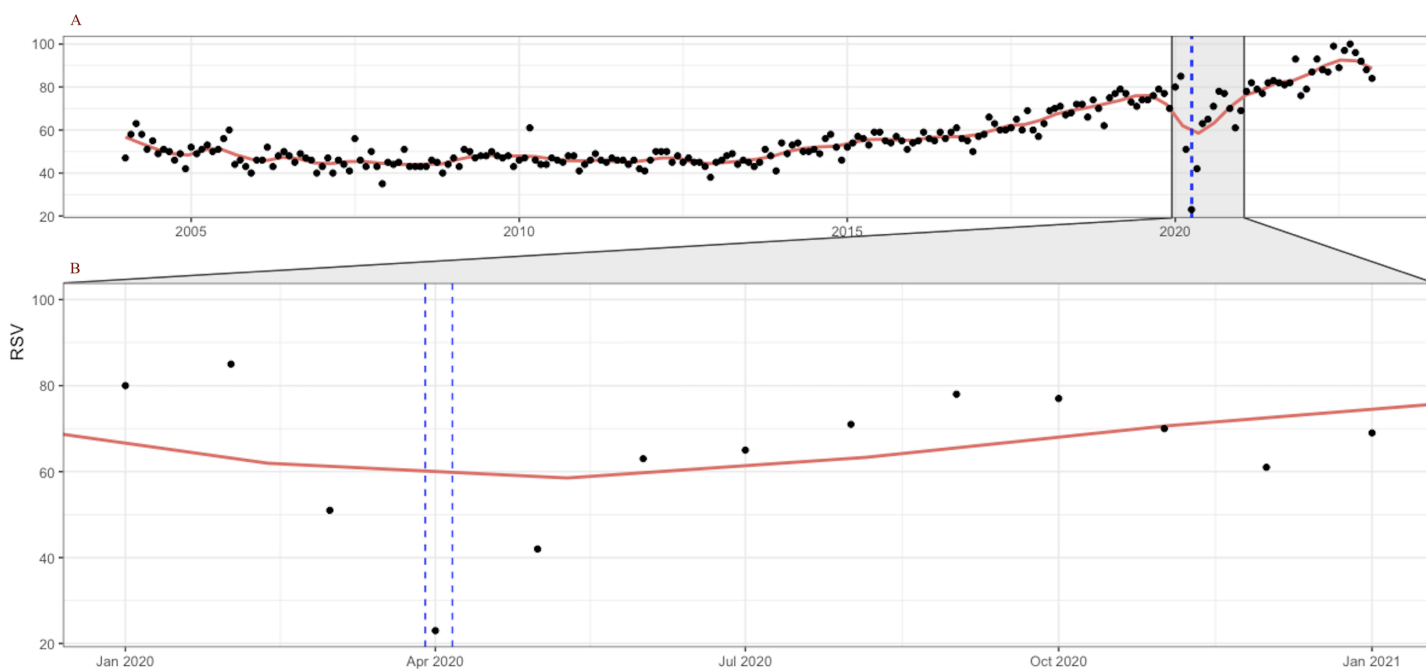


Figure 1 Google Trends search index for the keyword “colonoscopy” (January 2004–December 2022, category = “Health”) (A); Zoom in on January–December 2020 evolution (B). LOESS smoothing shown in red. Source: authors’ representation in R software (Wickham et al., 2023). Full-size [DOI: 10.7717/peerj-cs.1518/fig-1](https://doi.org/10.7717/peerj-cs.1518/fig-1)

increases the utility of public interest in cancer as reflected in web-search data to substitute for cancer indicators that have been affected by the coronavirus pandemic.

For example, Fig. 1 shows the all-time evolution of population interest for the keyword “coloscopy” at the global level and identifies without a doubt the plummet occurring during the month of April 2020, when the lowest popularity of the term has registered. In the following months, the trend reversed and continued on an ascending trend thereafter, mirroring the evolution of screening rates, as documented in the literature. The lower part of the chart contains a zoomed view over the entire year 2020, underlining the sharp decrease in interest starting in March 2020 and reaching an all-time low level in April 2020. The chart also draws a locally weighted polynomial regression line (LOESS) (Cleveland, 1979; Cleveland & Devlin, 1988) to increase trend readability.

Furthermore, the seasonal plot depicted in Fig. 2 contributes to clearly identify the particular pattern present for the year 2020 and the disruption that occurred in the relative search volume index (RSI) series in April 2020 during the first wave of the COVID-19 pandemic.

The data thus confirm previous findings that showed significant decreases in the number of screenings for a variety of cancers, including colorectal, immediately after the onset of the COVID-19 pandemic, which was no longer present by June. This further attests to the relevancy of the relative search index as a proxy for cancer screenings, with important implications for research, policy, and industry. The chart also helps to reveal other underlying seasonal patterns, such as a decrease in web-search interest for CRC in December each year, which should be considered for accurate forecasting.

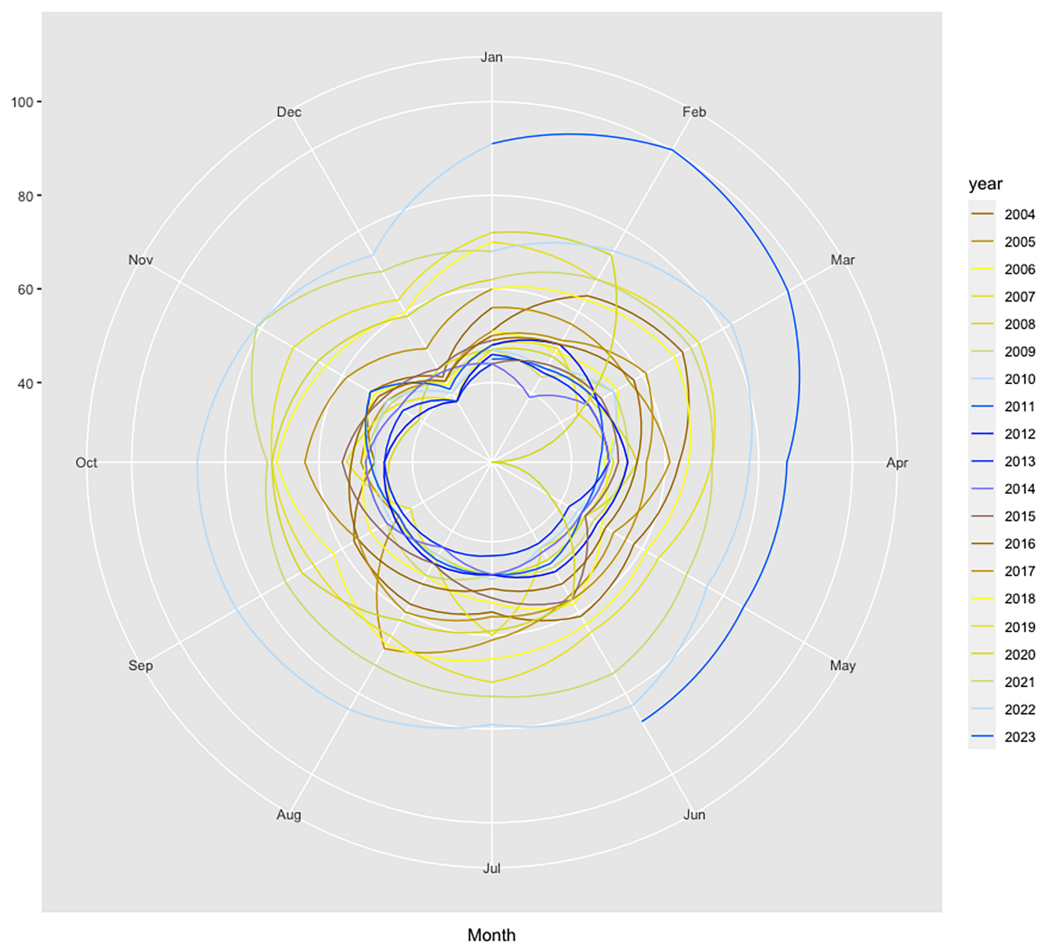



Figure 2 Polar seasonal plot for GT queries submitted for “colonoscopy” worldwide during 2004–2022. Source: authors’ representation in the “colorblindr” and ggplot2 R packages (Wickham *et al.*, 2023). Full-size  DOI: 10.7717/peerj-cs.1518/fig-2

Additionally, GT data has been shown to present non-linear characteristics that must be accounted for in modeling and forecasting endeavors. Very recently, *Borup & Schütte (2022)* demonstrate that taking them into account in models that employ GT data greatly increases performance over linear models. Amongst state-of-the-art models, artificial neural networks (ANNs) have been particularly useful for modeling complex systems due to their adaptability, parallel processing, and learning capabilities (*Sarangapani, 2018*), while properly accounting for data non-linearity (*Pasini, 2015; Tudor, 2022b*). There are many types of ANN architectures, amongst which the most widely used are feed-forward neural networks (FNNs) or multilayer perceptrons (MPs) (*Hsieh, 2004*).

The primary objective of the current study is to provide accurate forecasts for the CRC prevalence rate through a three-step forecasting framework that embeds an FNN autoregressive model (FNNAR). To reach its goal, it sources Google Trends data to assess public interest in CRC, and calibration is performed using the FNNAR model that captures non-linear patterns and traditional univariate forecasting techniques, including the exponential smoothing state space model (ETS) (*Ord, Koehler & Snyder, 1997; Hyndman*

et al., 2002) and the auto-regressive integrated moving average model (ARIMA) developed by *Box & Jenkins (1970)*, and then the best performing forecasting model in an out-of-sample setting is identified and used to issue forecasts for the CRC prevalence rate by the year 2025. To the best of the authors' knowledge, this is the first attempt to model and forecast CRC prevalence rates by using GT web-query data. Two secondary goals are also proposed and followed: (i) to document trends and heterogeneity in the prevalence of colorectal cancer by regions and country income level; (ii) to assess the efficacy of using public interest, reflected by Google Trends data, as an indicator of actual cancer prevalence and develop a CRC prevalence model based on web-search data. Different from the thin body of literature preoccupied with modeling and forecasting cancer indicators, the current research makes the following contributions: (i) it focuses on prevalence rates, as opposed to screening or incidence rates; (ii) it documents the capability of GT data to proxy for cancer statistics through multiple tools and assesses the relationship between CRC-related web searches and CRC prevalence statistics to identify the best keyword capable of proxying or even substituting prevalence rates. (iii) it performs preliminary testing and comparatively explores the in-sample and out-of-sample capabilities of alternative models, both statistical and machine-learning (*i.e.*, as per the categories proposed by *Breiman (2001)*) to confirm the over-performant model and its optimum parameters.

The manuscript is structured as follows: The second section describes the data employed in the investigations and introduces the various elements embedded in the forecasting framework. The next section presents the findings, whereas the fourth section contains a discussion of the main results, and the fifth section concludes the study.

MATERIALS AND METHODS

Data

CRC prevalence data were extracted from the publicly available database of the Our World in Data (OWID) platform *via* the `owid()` function embedded in the dedicated R software package (<https://ourworldindata.org/>). At its turn, OWID sources CRC prevalence data from the Global Burden of Disease (GBD) database of the Institute for Health Metrics and Evaluation (IHME) (<https://www.healthdata.org/gbd>), which is thus the original cancer statistics source. The CRC prevalence data covers the 2004–2019 period (*i.e.*, $T = 16$) for a total of 202 individual countries and country panels. First, heterogeneity across world countries is assessed through “plotmeans” style plots with the R’s “gplots” package of *Gregory, Warnes & Lodewijk (2016)*. This method carries the non-trivial advantage of plotting group means along with confidence intervals (CI), which, together with descriptive statistics reported in [Table 3](#), offers an insightful view of the extent of heterogeneity present in the data. Of note, the t distribution is used by default to compute confidence intervals within the `plotmeans()` function. However, as *Rovetta (2023)* explains, confidence (or credible) intervals concern the mean distribution and generally are not the best indicators for the dataset variability.

Table 2 Income and geographical panels employed in the study.

Income panels	Geographical panels	Abbreviation
Higher income	Sub-Saharan Africa	SSA
Upper middle income	South Asia	SA
Lower middle income	North America	NAm
Low income	Middle East & North Africa	MENA
	Latin America & Caribbean	LAC
	Europe & Central Asia	ECA
	East Asia & Pacific	EAP

Moreover, the heterogeneity in worldwide CRC prevalence is further explored by employing both income-based and geography-based panels, as per the World Bank (WB) classification.

This approach is particularly helpful for informing global equitable health policies. Consequently, four income-based panels and their corresponding CRC prevalence rates are extracted. Subsequently, seven geography-based panels are also delineated, and the relevant observations are sourced. Table 2 reflects the income and geographical panels extracted from OWID, along with the abbreviations employed for the latter category.

The Google Trends platform (<https://www.google.com/trends/>) reports the relative search volume index (RSI) of a keyword in a region, over a specified time period, with an RSI of 100 reflecting the point of peak popularity (Silva et al., 2019). As such, the RSI is estimated as per Eq. (1) and subsequently scaled from 0 to 100 (Tudor & Sova, 2022).

$$RSI = \frac{\text{Total search volume for "keyword" submitted in the specified geography over the specified period}}{\text{Total searches submitted in the specified geography over the specified period}} \quad (1)$$

Google Trends data is sourced in this study by making use of R's "gtrendsR" package (Massicotte & Eddelbuettel, 2022), calling the gtrends() function, specifying alternatively "colorectal cancer" and "colonoscopy" as the search terms of interest, the default geography of "all" corresponding to global queries, and specifying the "health" category (i.e. category no. 45) to assure reliable results. Moreover, the string specifying the time span of the query is set to "all," which implies that monthly RSI data starting on January 2004 (i.e., the beginning of Google Trends) and spanning to December 2022 is extracted. Consequently, a time series of 229 monthly RSI observations is available and will be employed for modeling and forecasting purposes in this study. Tudor (2022a) offers further relevant details on the methodology behind the RSI index. The "gtrendsR" package is not only a great tool for extracting GT data, but allows the specification of multiple relevant elements, is capable of performing an interest-by-region analysis and is useful for making searches reproducible (however, as explained later, sample instability is an issue that must be accounted for). It should be mentioned that the ISO language code only influences the data returned by related topics, as per the package description, and thus does not impact the sourced data.

Method

This study develops a three-step framework to forecast monthly CRC prevalence evolution based on Google Trends relative search index.

First, a monthly model of CRC prevalence is estimated based on RSI for the world panel (as per World Bank), such that:

$$CRC = f(RSI) \quad (2)$$

where the CRC prevalence in year i is modelled as a linear function f of the same-year RSI, plus some white noise.

To perform the estimation, the monthly RSI index is first converted to annual averages to match the frequency of the CRC prevalence series imported from the OWID database. Moreover, the parameters are estimated *via* the ordinary least squares (OLS) estimator. Of note, the main research interest does not lie on fine-tuning the CRC-RSI relationship, but rather on the second step of the framework, *i.e.*, accurate forecasting of RSI. Additionally, particular care is given to the choice of the estimation period, juggling the need to include enough observations and the fact that the Google platform has taken off as a main search engine only in the most recent decade. As per Britannica (<https://www.britannica.com/topic/Google-Inc>), by 2004, Google was processing daily 200 million searches, increasing to over 3 billion by the end of 2011. As of March 2023, Google's market share of the global search market across all devices is nearly 84%, making it the dominant search engine worldwide (*Statista, 2023*), which translates into 9 billion daily searches (*EarthWeb, 2023*). In light of these figures and aiming to increase the relevancy of current findings, the estimations performed within the first step of the framework (*i.e.*, to assess the link between distinct RSI pertaining to alternative relevant keywords and official CRC prevalence statistics), have been limited to the most recent decade of available data. However, this in turn reduces the number of observations in the data sample employed in the first step within the proposed framework.

Second, the RSI time series is modeled in-sample, tested out-of-sample, and ultimately forecasted and validated against official statistics and/or results of other studies, where available. To this end, multiple tools are employed.

As such, the nonlinearity of monthly RSI data is examined by two alternative tests: the Tsay test for quadratic nonlinearity (*Tsay, 1986*), as well as the likelihood ratio test for threshold autoregression (LR) proposed by *Chan (1991)*. In both cases, the null hypothesis is that the time series includes an AR process (*Munim, 2022*). Next, for modeling and forecasting purposes, the RSI index is split into a training set (made up of 80% of observations) used for in-sample fitting of three alternative models (*i.e.*, ETS, ARIMA, and FNNAR) and a testing set (containing the remaining 20% of the observations) on which the forecasting ability of the best-fit model in-sample in each category (*i.e.*, best ETS, best ARIMA, best FNNAR) is comparatively assessed. For this purpose, several forecasting accuracy statistics (*Hyndman & Athanasopoulos, 2018*) are calculated (shown below). Moreover, the Diebold and Mariano (DM) test for superior forecasting ability (*Diebold & Mariano, 1995*) is estimated to confirm the superior forecasting ability of the best forecasting model relative to the second ranked candidate.

Root mean squared error:

$$RMSE = \sqrt{\text{mean}(e_t^2)} \quad (3)$$

Mean absolute error:

$$MAE = \text{mean}(|e_t|) \quad (4)$$

Mean absolute percentage error:

$$MAPE = \text{mean}(|p_t|) \quad (5)$$

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

Mean absolute scaled error:

$$MASE = \text{mean}(|q_j|), \quad (6)$$

where $q_j = \frac{e_t}{\frac{1}{N-1} \sum_{i=2}^N |y_t - y_{t-1}|}$ in the case of non-seasonal series and

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

In all cases, the forecast error is defined as:

$$e_{T+h} = Y_{T+h} - \hat{Y}_{T+h|T} \quad (7)$$

To sum up, the data splitting rule for the training-testing RSI modeling and forecasting is given in the script snippet in [Box 1](#).

Box 1. Script for implementing the splitting rule

```
RSI <- trends$interest_over_time
x<- ts(RSI$hits)
test_x <- window(x, start=c(185,1),end=c(229,1))#testing window
x <- window(x, end=c(184,1)) #training window
```

Simple statistical methods have dominated the field of time series forecasting for many decades, both in academia and industry, due to their relatively accurate, fast-to-compute, and interpretable forecasts, with neural networks emerging as especially accurate in univariate time series forecasting (*Semenoglou, Spiliotis & Assimakopoulos, 2023*). However, due to their capability of revealing the influence of many parameters, multivariate time series models have uncontested advantages, but their goodness of fit is intrinsically related to the quality and availability of data related to these parameters (*Atchadé & Sokadjo, 2022*). Thus, given that the more complex the model, the more the need for data, and the lesser the availability and quality of cancer prevalence statistics, especially in less-developed economies and during the COVID-19 pandemic, we rely on the best-identified univariate model in the current research. Our choice is further backed by the findings of *Ziel & Weron (2018)* that conclude that, for the

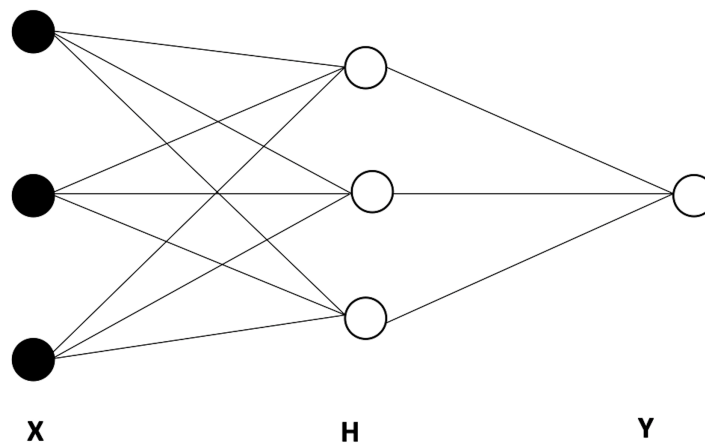


Figure 3 The basic architecture of an artificial neural network (ANN), including one hidden layer with three nodes. [Full-size !\[\]\(1679558f37f6db0dd8360a2a7e913e90_img.jpg\) DOI: 10.7717/peerj-cs.1518/fig-3](https://doi.org/10.7717/peerj-cs.1518/fig-3)

task of electricity price forecasting, the multivariate modeling approach does not uniformly outperform the univariate one across 12 distinct electricity spot price datasets and is even often outperformed by the latter. Furthermore, we rely on the findings of *Jaidka et al. (2021)* that study 208 Designated Market Areas (DMAs) throughout the United States and find that information-seeking behavior from Google Trends data is 19% more accurate than sociodemographic and regional controls at predicting the prevalence of noncommunicable diseases. Consequently, as per *Ziel, Steinert & Husmann (2015)*, we argue that our approach can surpass the complexity of cancer prevalence forecasting by offering several advantages relative to multivariate settings, such as being capable to model the search volume index and to provide accurate forecasts without the need for any data manipulation nor for any additional information, and, not in the least, by employing efficient and rapid state of the art estimation techniques.

As mentioned earlier, artificial neural networks (ANNs) have emerged as particularly useful for modeling and forecasting complex systems while accounting for non-linearities. The FNNs, are the most extensively utilized neural network (NN) models (*Hsieh, 2004*). A neural network is made up of several small computational units or nodes that are stacked in layers and run in parallel, the connections among nodes, and an activation function (*Chen & Billings, 1992; Allende, Moraga & Salas, 2002*). *Figure 3* provides the basic architecture of an ANN, which is made up of an input layer X, linked to at least one hidden H, which is, in turn, linked to the output layer Y (*Chen & Billings, 1992*).

The feedforward network is a type of neural network in which the input feeds forward through the network layers to the output. In the case of a feedforward neural network autoregression (FNNAR) model, the input layer consists of the past observations of a series (*Munim, Shakil & Alon, 2019*). Consequently, the vector of predicted values Y is found as follows:

$$Y = f(H) \quad (8)$$

where

$$H = \{\text{weight matrix } [(p * k)] * X + \text{bias vector} \quad (9)$$

and f is the activation function, with p standing for the nonseasonal inputs for the linear AR process, and k for the number of nodes in the hidden layer.

FNNAR models are fitted through automated algorithms in the “forecast” package in R software. In particular, the `nnetar()` function makes 25 repetitions and finds parameters through AIC minimization. By default, the number of nodes in the hidden layer is calculated as $k = (p + P + 1)/2$.

Additionally, to assure the robustness of findings, ETS and ARIMA models are automatically fitted in the R environment.

The exponential smoothing state space model (ETS) is based on the work of [Holt \(1957\)](#), [Winters \(1960\)](#), and [Brown \(1959\)](#). Predictions given by ETS are weighted averages of prior data, with greater weights assigned to more recent observations and weights decreasing exponentially ([Hyndman & Athanasopoulos, 2018](#); [Silva et al., 2019](#)). ETS estimates comprise three components: the trend (T), seasonal (S), and error (E) components. In addition, the trend includes both a level term (l) and a growth term (b). According to [Yang et al. \(2015\)](#), the trend and seasonal components might be none (N), additive (A), additive damped (Ad), multiplicative (M), or multiplicative damped (Md). Hence, an ETS model is represented by a three-character string (Z,Z,Z) ([Perone, 2021](#)), with the first Z representing the error assumption of the state-space model and the second and third Zs representing the trend type and the season.

The use of the automated ETS technique *via* the `ets()` function inside the forecast package in R ([Hyndman & Khandakar, 2008](#); [Hyndman et al., 2022](#)) assesses 30 ETS equations during the modeling procedure *i.e.*, all 30 formulae are provided by [Hyndman & Khandakar \(2008\)](#).

The approach optimizes the smoothing parameters and the starting state variable and employs a penalized likelihood, *i.e.*, the corrected Akaike’s Information Criterion (AICc), to choose the best model on the training set, which is then used to generate point predictions.

Autoregressive integrated moving average (ARIMA) models were created by [Box & Jenkins \(1970\)](#) and are among the most widely used parametric time series analysis and forecasting methods ([Silva et al., 2019](#)).

In equation form, according to [Hyndman & Athanasopoulos \(2018\)](#), an ARIMA (p,q,d) (P,Q,D) seasonal model is given as:

$$\begin{aligned} (1 - \varphi_1 B - \dots - \varphi_p B^p)(1 - \Phi_1 B^s - \dots - \Phi_P B^{sP})(1 - B)^d(1 - B^s)^D Y_t \\ = (1 - \theta_1 B - \dots - \theta_q B^q)(1 - \Theta_1 B^s - \dots - \Theta_P B^{sQ}) \varepsilon_t \end{aligned} \quad (10)$$

where s is the seasonal period, the lowercase and the capital letters represent nonseasonal and seasonal parameters, whereas ε_t is a zero-mean random variable with the standard deviation σ . The “auto.arima” function from the “forecast” package in R software conducts ARIMA modeling using an automated and optimized method that is capable of efficiently traversing the space of models in order to identify the ideal model. The function determines if seasonal differencing is required for the training data, calculates unit root

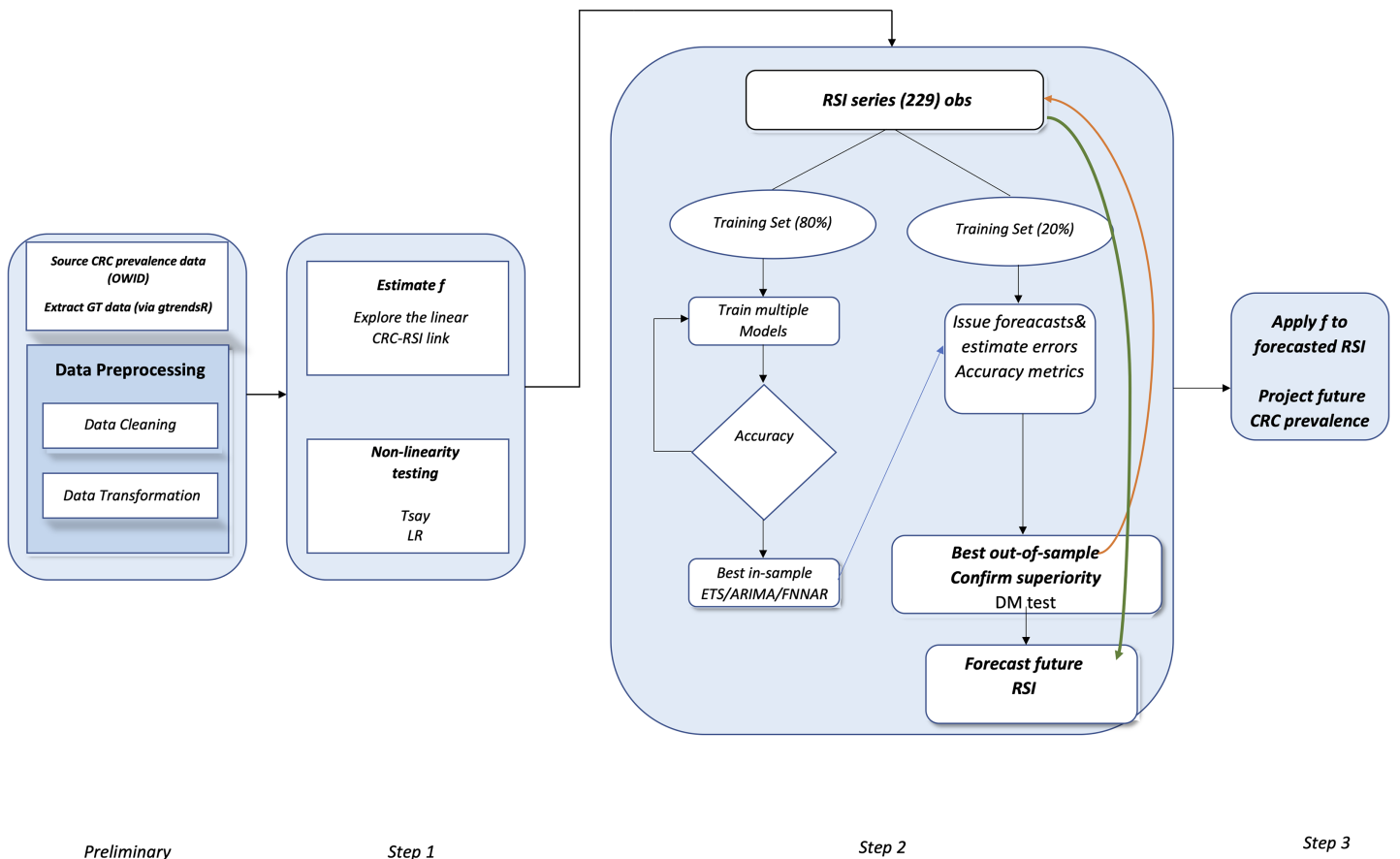


Figure 4 Flowchart of the proposed method.

Full-size DOI: 10.7717/peerj-cs.1518/fig-4

tests, and selects model parameters using a step-by-step AICc reduction procedure (Tudor, 2022a, 2022b).

The best forecasting model on the out-of-sample set is used to issue point forecast for the RSI index setting a forecasting horizon h of 24 months, thus reaching the end of 2024. Of note, h is chosen by complying with two important criteria in time series forecasting. First, ideally the forecasting horizon must be shorter than the testing window (Hyndman & Athanasopoulos, 2018). Second, longer-term forecasts have been shown to fail do provide superior information relative to the mean of the explained variable (Breitung & Knüppel, 2021). Consequently, setting the length of h equal to 24 assures that both criteria are met.

Third, the function f estimated at step one is applied to the RSI series forecasted at step two, under the no-change hypothesis, in order to produce forecasts for the CRC prevalence rate up to the end of 2024, as follows:

$$\widehat{CRC} = \hat{f}(\widehat{RSI}) \quad (11)$$

To sum-up, the proposed framework embeds the sequential main steps depicted in Fig. 4.

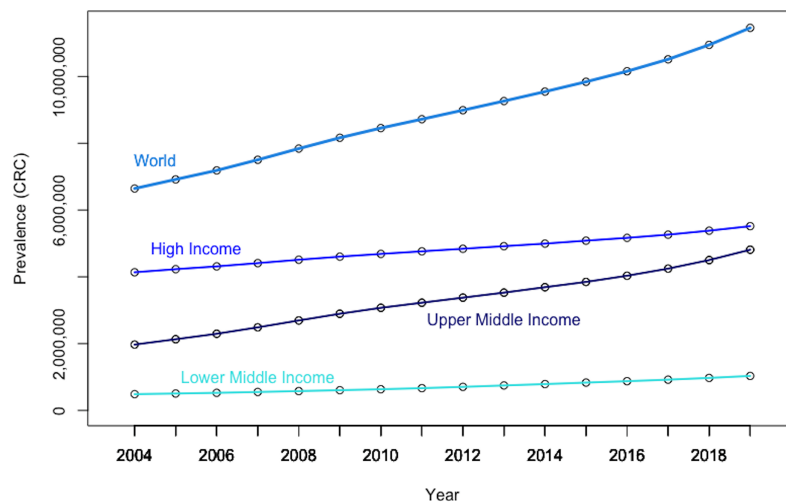


Figure 5 Evolution in CRC prevalence rates at the world level and by income-based country panels. Authors' representation in R software. Our World in Data (OWID) <https://ourworldindata.org/>. The low-income panel, showing a flat line at zero on the current scale, is eliminated from the chart to improve overall readability. Authors' representation in R software. Source of data, OWID.

Full-size DOI: 10.7717/peerj-cs.1518/fig-5

RESULTS

Worldwide trends and heterogeneity in CRC prevalence

Figure 5 reveals that the global panel presents a rapidly increasing CRC prevalence rate over the last decades, which is mostly driven by the upper-middle income panel, confirming that the rising number of CRC cases is creating an increasing global public health concern (Xi & Xu, 2021).

The graphical depiction further shows that the prevalence rate of CRC in high-income countries (*i.e.*, as per the World Bank classification) has consistently been about 10 times that of low- and middle-income economies. Moreover, it should be noted that the growth rate of CRC prevalence is significantly steeper in upper-middle income countries than in both rich and poor countries, suggesting increased exposure to CRC risk factors, albeit at a high basal level. Nevertheless, it is increasing rapidly in less developed countries due to increased exposure to CRC risk factors.

Furthermore, there is also high geographical heterogeneity in CRC prevalence at the world level (Fig. 6). Thus, data indicates that the East-Asia& Pacific region is registering both a high basal CRC prevalence level, as well as a high growth rate, confirming that immediate measures are paramount to offer some relief against this burden.

Additionally, the heterogeneity amongst individuals is also apparent in Fig. 7, which denotes the mean CRC prevalence level over the sample period for each of the 202 countries included in the analysis. The plot is issued through the `plotmeans()` function in the “gplots” package, which uses the *t* distribution to compute confidence intervals, with a 0.95 confidence level for error bars.

The high country heterogeneity also emerges from the descriptive statistics for the entire sample, which are reported in Table 3. Estimations show a very high standard deviation

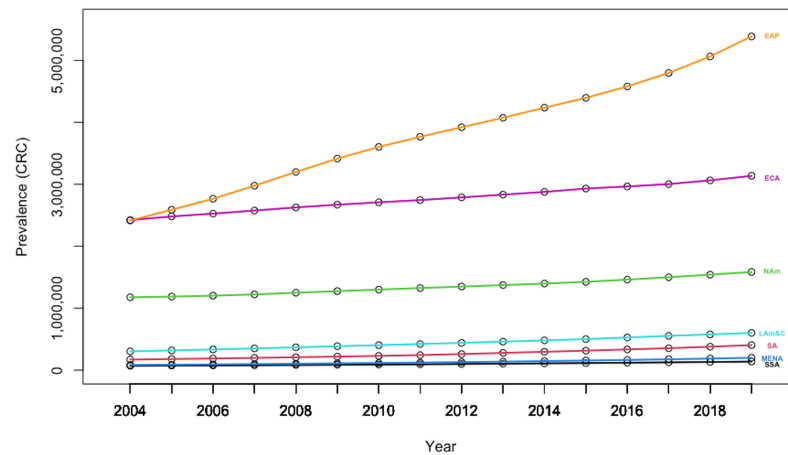


Figure 6 Evolution in CRC prevalence rates by geography-based country panels. Authors' representation in R software. Source of data: Our World in Data (OWID) <https://ourworldindata.org/>. Full-size DOI: 10.7717/peerj-cs.1518/fig-6

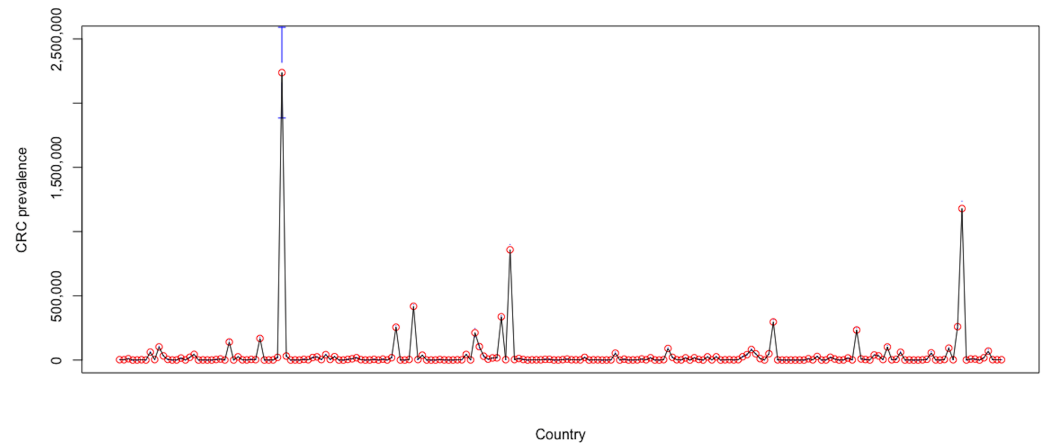


Figure 7 Heterogeneity amongst countries over 2004–2019. Authors' representation in R software. Source of data, Our World in Data (OWID) <https://ourworldindata.org/>. Full-size DOI: 10.7717/peerj-cs.1518/fig-7

Table 3 Descriptive statistics for CRC prevalence during 2004–2019.

Statistic	CRC prevalence
Mean	86,722.88
Standard deviation	657,663.06
Min	0.00
Max	11,457,627

(657,663.06) and huge range (11,457,627) for the levels of CRC prevalence for world countries during 2004–2019, whereas its mean level was 86,722.88.

Relevancy of alternative RSIs for CRC prevalence

Figure 8, which plots the trends of the CRC prevalence and of the RSI “colonoscopy” in panel a, and the scatterplot between the two variables in panel b reveals a disaggregation

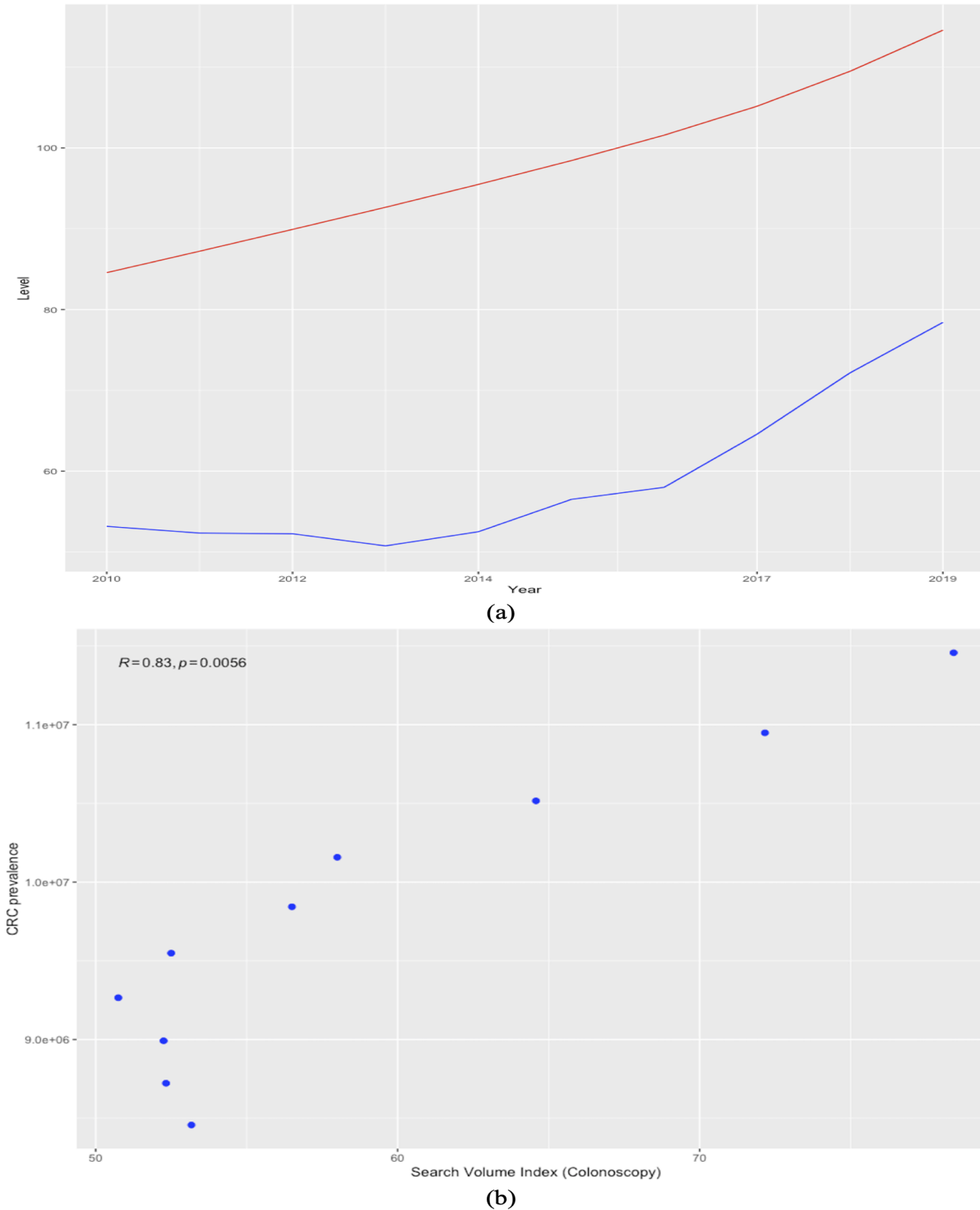


Figure 8 Trends in worldwide cancer prevalence ($\times 100,000$, red line) and in the relative search index (RSI) for keyword “colonoscopy” (blue line) (2010–2019) (A). the relationship between the two variables over 2010-2019 (scatter plot) (B). The Spearman correlation coefficient is automatically reported on the plot issued with the `ggscatter()` function in the “ggpubr” package. [Full-size !\[\]\(b345a1c4255362eec3746050dd71ccac_img.jpg\) DOI: 10.7717/peerj-cs.1518/fig-8](https://doi.org/10.7717/peerj-cs.1518/fig-8)

between the two series during the first years in the sample, when the RSI series presents a decreasing trend, whereas the CRC prevalence series shows an uninterrupted increasing trend throughout the analysis period. This in turn leads next to the implementation of a joinpoint (or change-point) regression analysis (Gillis & Edwards, 2019) to confirm and assess the changing trends that emerge from the visual inspection. Of note, this kind of analysis has been routinely applied to detect changing trends in cancer mortality and/or incidence series (see, among others, Qiu et al., 2009; Crispo et al., 2013; Sarakarn et al., 2017; Wilson, Bhatnagar & Townsend, 2017). Results of the joint-point regression analysis, graphically reflected in Fig. 9, confirm that the RSI series presents one join point in 2014, leading to two periods with distinct trends, as follows: a negative trend until 2014, and an increasing trend during 2014–2019 (Fig. 9). On the other hand, the joinpoint regression analysis also found one joinpoint in CRC prevalence series around 2014, but in this case the joinpoint delineates two positive trends. All estimations have been performed R's software "segmented" package. Hence, given the disaggregation between the series during the first segment, to comply with the linearity assumption, we subset the secondary segment to implement the linear models and estimate the statistics of interest.

Estimation results for the two linear models are summarized in Table 4. Of note, as the RSI indexes are scaled from zero to 100 by construction, to make the data comparable and obtain more consistent estimates, the CRC prevalence rates are divided by 100,000 prior to estimating the model. First, it should be noted that effect sizes are the paramount outcome of empirical statistical research (Lakens, 2013). As explained by Lakens (2013), which in turn cites Rosenthal (1994), effect sizes are usually classified into two families: the *d* family (*i.e.*, standardized mean differences) and the *r* family (*i.e.*, measures of strength of association), whereas the *d* family effect sizes are conceptually determined by the difference between observations divided by the standard deviation of these observations, and the *r* family effect sizes characterize the proportion of variance that can be attributed to group membership. Moreover, as Lee (2016) explains, *p*-values can only inform of the statistical significance, but not the magnitude of the effect, whereas CIs mitigate, but do not solve, this issue by providing a range of possibility. Furthermore, it should be mentioned that the Shapiro-Wilk test (Shapiro & Wilk, 1965) estimated for the three series could not reject the null hypothesis that all samples stem from a normal distribution.

In light of these considerations, we estimate and report the Pearson correlation coefficient and the value of Cohen's *d* statistics (Cohen, 1988) together with its correction for sample bias (*i.e.*, known as Hedges' *g*) between the CRC prevalence series and each RSI series, including the 95% CIs for all estimates. Both Cohen's *d* and Hedges' *g* statistics are estimated by calling the `cohen.d()` function within the Efficient Effect Size Computation or "effsize" package (Torchiano, 2020) in R software. Results agree that the RSI for the keyword "colonoscopy" is the best indicator for cancer prevalence, showing a 95% CI for the Pearson correlation coefficient of (0.9324; 0.9992), and a goodness-of-fit (R-squared) of 98.56%, indicating that it can explain most of the variability in CRC prevalence at the world level during the most recent years of available data. In turn, the RSI for the keyword "colorectal cancer" can also be a reliable indicator for cancer prevalence, although not a substitute, contributing to explain close to 73% and showing lower correlation with the

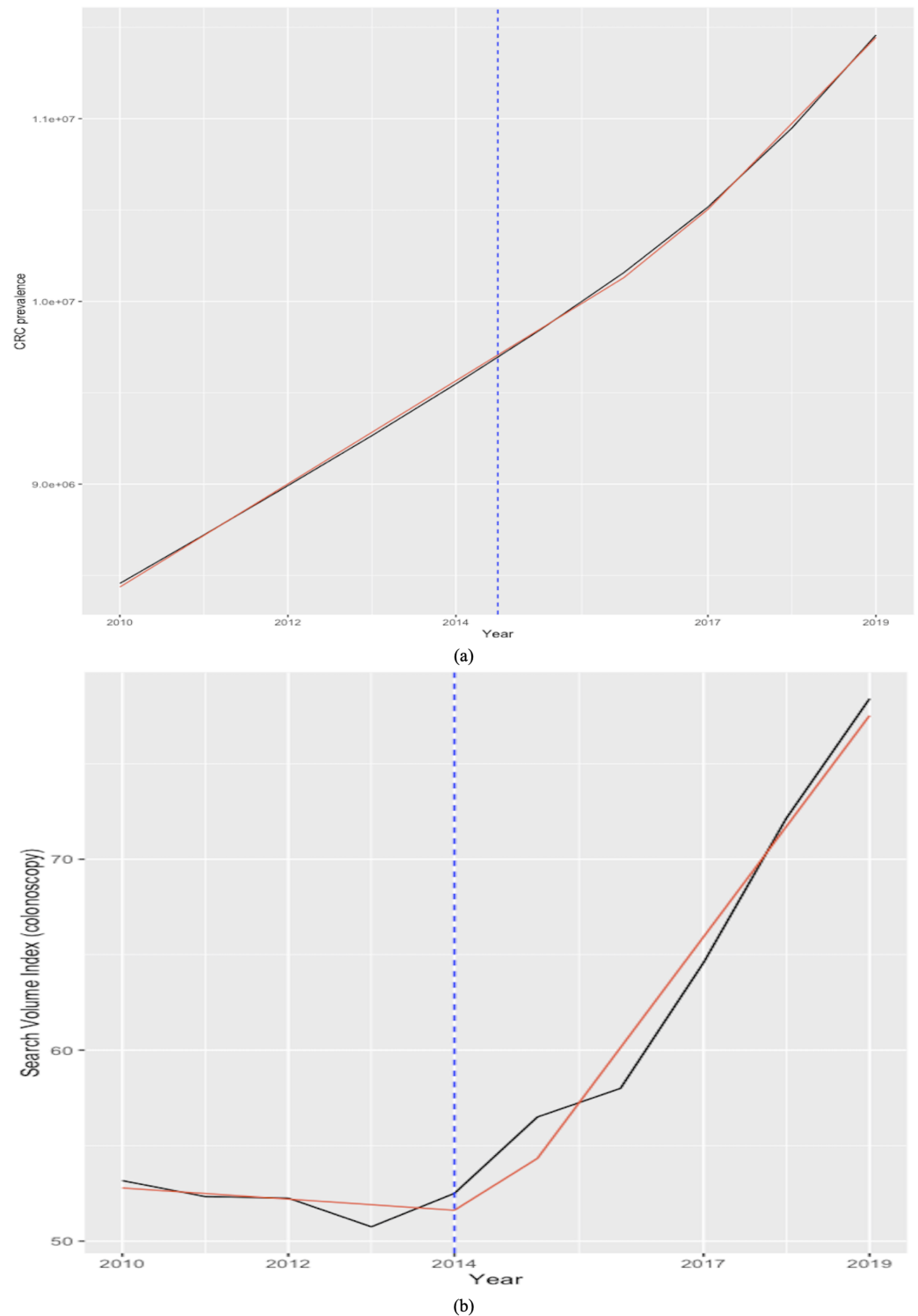


Figure 9 Jointpoint regression analysis (CRC prevalence rate in A and RSI for keyword “colonoscopy” in B). Source of data: CRC prevalence data is sourced from Our World in Data (OWID) <https://ourworldindata.org/>, which in turn sources data from IMHE, and RSI data is sourced from Google Trends. The “segmented” package within R software is used for the implementation of the jointpoint regression. [Full-size](#) DOI: 10.7717/peerj-cs.1518/fig-9

Table 4 Estimates of the link between CRC prevalence rate and alternative RSI indexes.

	RSI “colonoscopy”	RSI “colorectal cancer”
Pearson corr (Lower 95%; Upper 95%)	0.9927 (0.9324; 0.9992)	0.8567 (0.1484; 0.9840)
R squared	0.9856	0.7341
Slope (<i>p</i> -value) (Lower 95%; Upper 95%)	0.70 (0.00) (0.5876; 0.8251)	1.12 (0.029) (0.5323; 2.2997)
Cohen’s d (Lower 95%; Upper 95%)	2.8626 (1.0323; 4.6929)	5.3896 (2.6213; 8.1580)
Hedges’ g (Lower 95%; Upper 95%)	2.6424 (1.0173; 4.2674)	4.9750 (2.5724; 7.3777)

Note:

Elements of the linear models: Dependent variable: CRC prevalence rate; Independent variable: corresponding RSI index; Cohen’s d is estimated by calling the `cohen.d()` function within the Efficient Effect Size Computation or “`effsize`” package in R software; Hedges’ g statistics is computed by feeding the instruction (`hedges.correction==TRUE`) when calling `cohen.d()`.

Table 5 Results of the nonlinearity tests.

Test	Test statistic	<i>p</i> -value
Tsay	15.32	0.00
LR	110.11	0.00

CRC prevalence. Moreover, although the effect size magnitude is large for both relationships according to the thresholds provided in [Cohen \(1992\)](#), it is nonetheless significantly smaller for the RSI corresponding to the keyword “coloscopy” than for the RSI sourced for the keyword “colorectal cancer”. Consequently, estimations, albeit based on a small sample, agree with previous findings, indicating that online information-seeking for the keyword “colonoscopy” is the best proxy for the prevalence of CRC at the world level.

Forecasting CRC prevalence from Google search interest

The results of the nonlinearity tests performed on the RSI time series are reported in [Table 5](#). Results agree that the web-query index is nonlinear, as both tests strongly reject the null hypothesis.

Consequently, accounting for the proven existence of data nonlinearity, this study proceeds to estimate a univariate neural network forecasting models that can properly handle this characteristic, *i.e.*, FNNAR. For comparative purposes and to further increase the reliability of results, the widely used ETS and ARIMA models are also estimated, and their respective forecasting capability assessed. Results enclosed in [Table 6](#) reveal that the ANN model, NNAR, is acknowledged as best performing in terms of out-of-sample forecasting accuracy by all scale- and scale-free metrics. It should also be mentioned that the estimated DM test further confirmed the superiority of the neural network model when compared to the second ranked forecasting model, ARIMA.

Table 6 Accuracy measures over the testing set (out-of-sample) for competing predictive models.

	RMSE	MAE	MAPE	MASE
FNNAR	14.42	9.39	17.12	2.55
ETS	15.38	10.77	19.65	2.92
ARIMA	14.69	10.86	17.77	2.94

Table 7 Robustness check: accuracy measures over the testing set (out-of-sample) for competing predictive models for the consolidated sample.

	RMSE	MAE	MAPE	MASE
FNNAR	12.56	8.32	15.20	1.93
ETS	20.27	17.27	28.55	4.01
ARIMA	12.89	8.87	15.79	2.06

To assure the robustness of current findings, we first mitigate the sample instability that characterizes GT-sourced RSI series by following the technique proposed by [Medeiros & Pires \(2021\)](#), which solves the sampling instability of RSIs by sourcing and averaging across multiple samples with the same specifications. The aforementioned study shows that this approach improves both model selection and forecast accuracy. Here, we source and average across five different RSI samples constructed by indicating the same keyword, category, and geography. Results, reported in [Table 7](#), again indicate the autoregressive neural network model as best performing in terms of out-of-sample forecasting accuracy and also reveal that employing the consolidated sample does mitigate forecast errors over the testing set for both FNNAR and ETS, offering some support for the findings of [Medeiros & Pires \(2021\)](#).

Moreover, we also perform a secondary robustness check by repartitioning the original RSI sample through the implementation of a 174-55 data splitting rule, as per [Box 2](#).

Box 2. Script for implementing the alternative data partitioning rule

```
RSI <- trends$interest_over_time
x<- ts(RSI$hits)
test_x <- window(x, start=c(175,1),end=c(229,1))#testing window
x <- window(x, end=c(174,1)) #training window
```

[Table 8](#) contains the accuracy measures estimated over the newly defined test-set that were issued by the three competing models, now trained over the first 174 observations in the sample. The nonlinear FNNAR has again been most able to capture variations in data and thus to issue superior out-of-sample forecasts for the RSI.

In light of the consistent superior performance over alternative test sets, the autoregressive NN model is subsequently reestimated over the entire series spanning January 2004 to December 2022 to further issue the expected RSI for CRC over the

Table 8 Robustness check: accuracy measures over the alternative testing set (*i.e.*, for alternative data partitioning rule) for competing predictive models.

	RMSE	MAE	MAPE	MASE
FNNAR	13.14	10.31	16.62	2.45
ETS	13.77	11.01	17.06	2.62
ARIMA	13.99	11.27	17.35	2.68

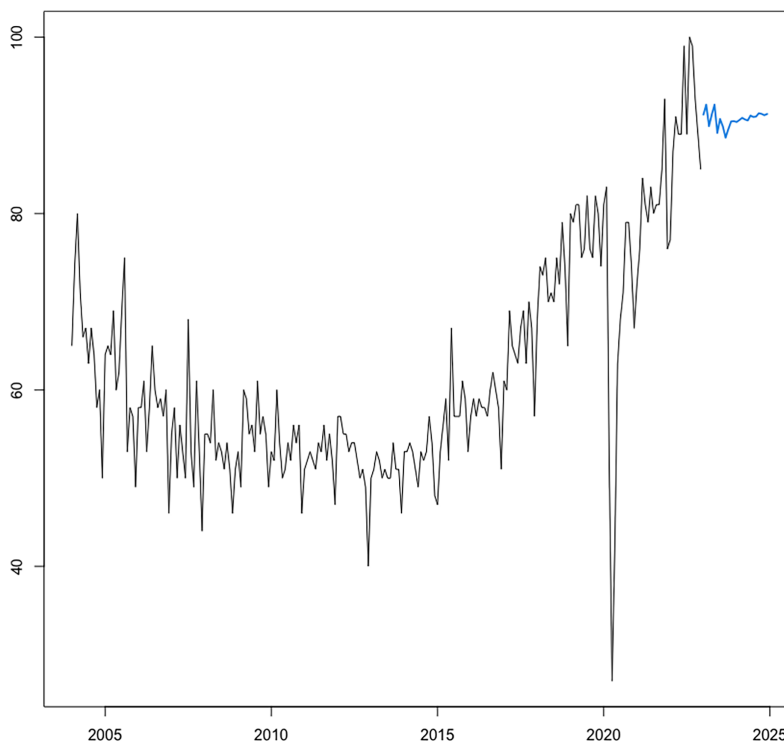


Figure 10 Forecasting of CRC search interest (RSI) at $h = 24$. Source: estimation results. The optimal neural network forecasting models is of the form NNAR (7,4,1), including seven autoregressive lags and four nodes in the hidden layer. [Full-size](#) DOI: [10.7717/peerj-cs.1518/fig-10](https://doi.org/10.7717/peerj-cs.1518/fig-10)

forecasting horizon of 24 months corresponding to the January 2023–December 2024 interval. Projections (visually represented in Fig. 10) show a continuation of population web-search interest in CRC, with an overall growth rate of 16.42 percent relative to the year 2019 (*i.e.*, the point corresponding to the most recent observation for registered CRC prevalence rates at the world level), which in absolute terms would indicate a total burden of disease of 12,361,225 cases, or an additional number of 903,598 CRC “survivors” by the end of 2024. Of note, current projections reinforce the estimations of the Global Cancer Observatory (<https://gco.iarc.fr/>), reported by *Xi & Xu (2021)* and *Morgan et al. (2023)*, among others, which expect a surge in colorectal cancer incidence and consequently report an expected count of 3.2 million new CRC cases by 2040, or an increase of 63% relative to 2020 levels. Furthermore, current projections also agree with the estimations of the *American Cancer Society (2022a)* cited by the *National Cancer Institute (2022)*, which

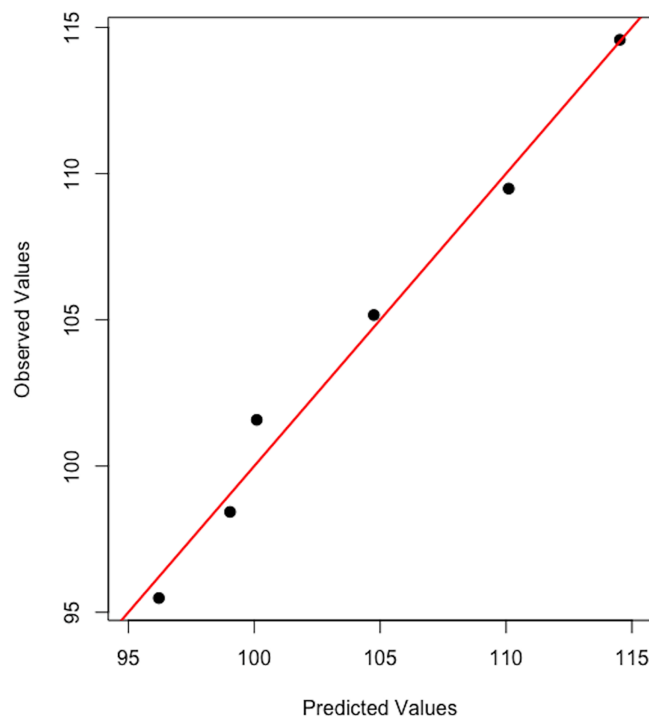


Figure 11 Observed vs predicted global CRC prevalence. Full-size  DOI: 10.7717/peerj-cs.1518/fig-11

expect the number of US all-cancers survivors to increase by 24.4%, thus reaching 22.5 million, by 2032.

As a final step, we attempt to validate the capability of the proposed framework by employing it for the task of nowcasting and forecasting the level of worldwide CRC prevalence and compare its projections with the last year of available statistics at the world level (*i.e.*, 2019) provided by the Global Burden of Disease (GBD) database of the Institute for Health Metrics and Evaluation (IHME) (<https://www.healthdata.org/gbd>).

Figure 11 attests that the linear model constructed within the step 1 of the framework manages to accurately nowcast the level of global CRC prevalence for 2019, indicating a global burden of 11,452,157 compared to 11,457,627 reported by the GBD database of IHME (*i.e.*, 5,470 residual cases at the world level for year 2019).

For forecasting purposes, we subset the first 192 observations in the sample (spanning January 2004–December 2019) and employ an alternative data partitioning strategy (*i.e.*, 180–12). Thus, the FNNAR is trained over the first 180 observations (training window spanning from December 2018) and used to make predictions for the relative search index for “coloscopy” through December 2019. Figure 12 shows the graphical representation of the FNNAR predictions over January to December of 2019, indicating that the model has been able to detect trends fairly well, although it has slightly underestimated the relative search interest over the testing set, issuing an average forecast for RSI of 71.5 over 2019 relative to the actual average RSI level of 78.33. Then, projections issued by FNNAR for the end of 2019 are fed into the linear model constructed during the first step of the framework to finally estimate the global CRC prevalence level corresponding to the year 2019, which is

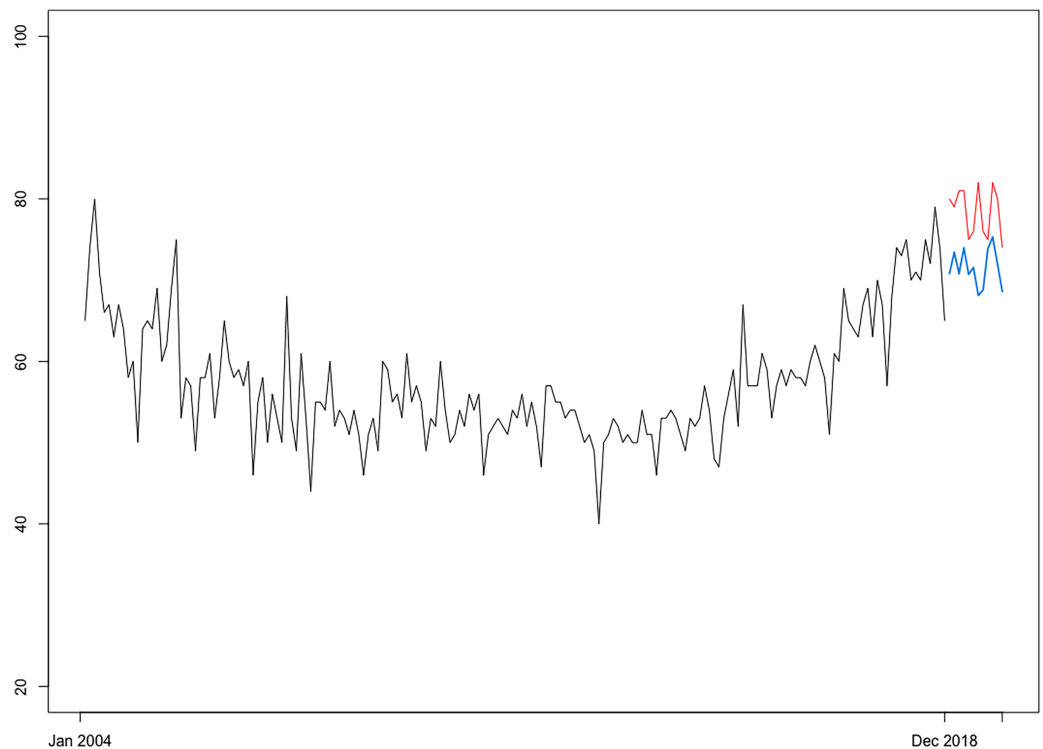


Figure 12 Projections of CRC search interest (RSI) over January 2019–December 2019 ($h = 12$) issued by FNNAR (Average of 20 networks, each of which is a 13-7-1 network with 106 weights). Blue line, NNAR projections; Red line, real RSI index. [Full-size !\[\]\(1679558f37f6db0dd8360a2a7e913e90_img.jpg\) DOI: 10.7717/peerj-cs.1518/fig-12](https://doi.org/10.7717/peerj-cs.1518/fig-12)

then compared to statistics published by IHME. Results show that the proposed framework has been able to capture the global CRC prevalence level for 2019, indicating an interval of (10, 114, 105–11, 812, 163) for the total number of cases globally (*i.e.*, 10,963,134-point forecast), whereas the GBD database of IHME reports a global burden for colorectal cancer of 11,457,627 for 2019. However, to obtain a clearer picture of the framework's ability to project the global burden of the disease, additional validation is required; this will be accomplished with the release of new official statistics corresponding to more recent years.

DISCUSSION

With the help of big data provided by the Google Trends platform and publicly available data on CRC prevalence rates for 202 countries and several distinct income and geographic panels according to the World Bank classification, this study performs the following tasks: (i) it explores geographical and income-based disparities in CRC prevalence at the world level; (ii) it assesses associations between cancer prevalence and population web-search behavior, thus documenting the relevancy of GT data to proxy for health problem occurrences, and (iii) it comparatively investigates the modeling and forecasting abilities of well-established univariate techniques (including FNNAR, ETS and ARIMA) for RSI time series to finally project CRC prevalence rates for a 24-month forecasting horizon (*i.e.*, up to 2025).

Particularly, based on Google Trends data spanning 2004–2022 and the estimated RSI through the end of 2024, as well as CRC prevalence data for the period 2004–2020, future prevalence rates were projected through the feedforward autoregressive neural network forecasting model (FNNAR), which, due to its ability to learn from the data and deal with non-linearity, emerged as superior in terms of out-of-sample forecasting performance within a pool of three traditional competing univariate predictive models.

CRC prevalence projections point to an increase of 19% relative to the 2019 level, which translates into a total of over 13.64 million people living with colorectal cancer by 2025, or close to 2.2 million over the level registered at the end of 2019, resonating with the findings of previous studies (among others, *Smittenaar et al., 2016*; *Xi & Xu, 2021*) and indicating that increased demands upon the health and insurance services should be expected and carefully planned for. However, similar to *Jakobsen et al. (2021)*, it should be acknowledged that cancer-related indicators should be regularly revised and updated for efficient resource allocation.

Other important results indicate high income and geographical heterogeneity in CRC prevalence at the world level, with high-income countries registering significantly higher prevalence rates than the countries in other income categories, and upper-middle-income countries showing the highest growth rates in CRC prevalence. These findings are in line with *Ades et al. (2013)*, which confirmed that higher economic development and increased expenditures on health at a national level were linked with increased cancer incidence and decreased cancer mortality, which in turn explains the increased prevalence rates encountered in high- and middle-upper-income economies. Current results also agree with previous studies that indicate a rapidly increasing CRC prevalence in less developed countries due to increased exposure to CRC risk factors. Consequently, the study fully supports *Arnold et al. (2017)* in that, to lower the number of CRC patients in the next decades, targeted resource-dependent interventions are required, such as primary prevention in low-income areas and early identification in high-income settings.

The increasing CRC prevalence rates detected, especially in industrialized countries, can be explained by the progress and consequent increase in screening rates, which are in turn reflected in higher cancer incidences. Further, the increase in screening rates has a direct impact on prevalence and contributes, along with lower mortality rates, to the steep increase in prevalence. Indeed, screening seeks to improve patient prognosis by improving early identification and treatment and lowering colorectal cancer (CRC) incidence and death (*Siegel, Miller & Jemal, 2019*; *Gaur & Jagtap, 2022*). Over recent decades, the importance of CRC cancer screening as a main preventive measure has grown (*Colditz & Dart, 2013*), whereas colonoscopy has been recognized as the gold standard for CRC screening (*Food and Drug Administration (FDA), 2021*). For example, at the United States (US) level, the state with the highest rate of screening has been identified as Massachusetts (*Colditz & Dart, 2013*).

Additionally, the relevancy of web-query data as a proxy for awareness and interest in cancer screening has been previously documented (*Schootman et al., 2015*). As such, the strong connection between RSI for “colonoscopy” and the CRC prevalence rate detected in the current study can be explained through the screening-incidence-prevalence

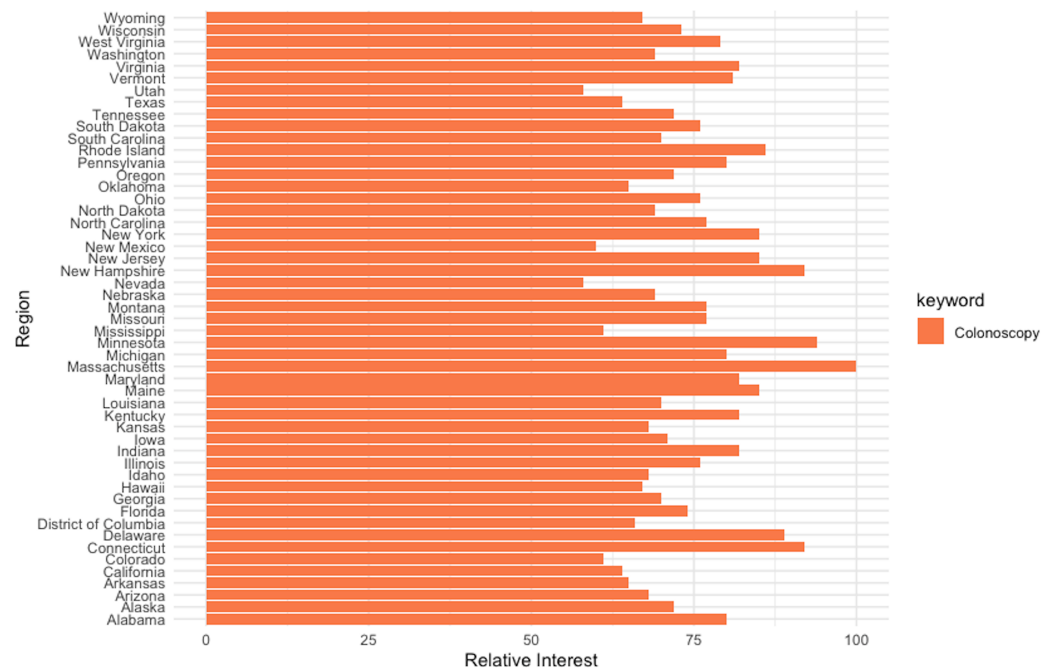


Figure 13 Interest-by-region analysis for the keyword “colonoscopy” at the US level. Source, Authors’ representation in R software. [Full-size !\[\]\(1663bb69f307a960345edb0e712f8c02_img.jpg\) DOI: 10.7717/peerj-cs.1518/fig-13](https://doi.org/10.7717/peerj-cs.1518/fig-13)

transmission channel. Furthermore, the breakdown by region of the web-search interest for the keyword “colonoscopy” performed at the “US” geography level (enclosed in Fig. 13) identifies Massachusetts as the US state with the highest relative search interest, which further supports on one hand the aforementioned argumentation and on the other hand the previous literature on the use of web query data in health research.

Overall, this research resonates with *Jun, Yoo & Choi (2018)* that the use of big data, such as the GT relative search volume indexes, has shifted recently from modeling to forecasting. Additionally, it contends that, for cancer research, web-search data might present several advantages over official statistics, including: (i) first, official data are reported with a lag, and thus timely accurate research is hampered by this unavoidable delay; (ii) second, outlier events that disrupt health care systems, such as the COVID-19 pandemic, cause significant underdiagnosis (*Jacob et al., 2021; Marques et al., 2021*), which invalidates official data that no longer accurately capture the variation in incidence, hence also leading to faulty prevalence indicators. In line with *Salathé et al. (2012)* and *Sulyok, Ferenci & Walker (2021)*, local and timely information on disease and health patterns is extracted from population web-search interests. Moreover, future research focused on the most vulnerable countries, with undeveloped healthcare systems and often unreliable or nonexistent statistics, can draw significant benefits from extracting reliable health information from web-queries.

Limitations

Research that employs Google Trends data, as is the case with the current study, is based on the central premise that people turn to Google to find subject-related information when

they need it, and, due to its visibility, this information demand can subsequently be used as a reliable predictor in a variety of settings (Bleher & Dimpfl, 2022). Moreover, compared to official statistics, GT data are available in real-time and cover a wide sample of countries, regions, and even some large cities, offering valuable insights especially when other data are lacking or are only available with important time lags (Eichenauer et al., 2022), as is the case with cancer prevalence data that makes the subject of the current research.

Consequently, Google Trends data have proven particularly helpful in shedding light on population health behaviors (Arora, McKee & Stuckler, 2019) and hold the potential to improve the capabilities of public health surveillance systems (Althouse et al., 2015; Tkachenko et al., 2017). However, there are also some non-trivial limitations to the GT data that should be properly acknowledged (Tudor, 2022a) and, where possible, overcome.

First, Google Trends RSI is aggregated from a sample of search queries (Narita & Yin, 2018), thus suffering from sample bias (Medeiros & Pires, 2021) which, together with sample instability that arises from its construction (Eichenauer et al., 2022), can have serious consequences for the reliability of research findings that use GT data. In fact, this makes any study employing GT data inherently irreproducible (Rovetta, 2021). To mitigate, albeit not overcome, these issues, we construct a consistent sample by additionally employing a robust sampling procedure that involves sourcing multiple samples at various intervals and averaging across to construct a “consolidated” RSI. Moreover, it should also be mentioned that the aforementioned weaknesses related to GT data are prominent for less searched topics or keywords (Medeiros & Pires, 2021), which was not the case for the main keyword of interest in this study (*i.e.*, “colonoscopy”) over the analyzed period. Moreover, previous research has shown that “worldwide” RSIs, as is the case in the current study, are the most reliable GT samples (Rovetta, 2021).

Second, Google Trends data is susceptible to being influenced by external events, such as news events or media coverage (Sato et al., 2021; Satpathy, Kumar & Prasad, 2023). Thus, given that the period under consideration in the current study contains the COVID-19 pandemic, it should be noted that the media coverage may have influenced the search volume in pandemic-related terms during the analyzed period (Rovetta & Castaldo, 2021), and, given the construction of the RSI as explained in the method section, this in turn may have direct consequences for the relative interest in unrelated terms, such as the one used in this research (*i.e.*, “colonoscopy”). Nonetheless, it should also be mentioned that even allowing for the possibility that media coverage may affect population web searches, GT data still offers a way to measure web interest in a particular subject more effectively than any other approaches previously employed (Rovetta, 2021). Furthermore, as the COVID-19 pandemic significantly affected CRC screening rates, with direct consequences for the accuracy of CRC prevalence data and their usefulness for cancer research, we align with previous studies and argue that digital traces are the best available tools for assessing disease prevalence.

CONCLUSIONS

Significant advancements in cancer screening and treatment have increased both cancer incidence statistics and patients' life expectancy. As a result, a larger portion of the

population is becoming increasingly reliant on social services, particularly health care, and this trend is especially visible in high- and upper-middle-income countries. In this context, the prevalence of cancer has emerged as a vital indicator for effective policy and proper resource allocation, both of which require reliable estimates. Consequently, accurate CRC prevalence projections carry important implications, informing the policymakers, the insurance companies, and the public authorities on the estimated future burden, the size of programs aimed at releasing this burden, and the related research financing needs.

However, unlike cancer incidence and mortality rates, cancer prevalence projections are not routinely issued by national and international agencies, and the limited or nonexistent cancer statistics for large portions of the world, along with the high heterogeneity among world nations, further complicate the task of producing timely and accurate CRC projections. In this context, population interest in the form of internet-submitted queries are crucial for enhancing cancer statistics and ultimately assisting cancer research.

This article proposes a three-step framework that constructs a CRC prevalence model from Google Trends search volume series to forecast CRC prevalence at a monthly frequency with a 24-month forecasting horizon. Multiple ETS, ARIMA, and FNN models are fitted to the monthly Google Trends series, and all best-fit models in each category are employed for the task of issuing 45-point forecasts on the test data window and subsequently ranked by multiple accuracy metrics. Results consistently indicate that the machine-learning neural network outperforms in the out-of-sample setting, due to its capability to capture non-linearity, which has been identified in the data by two alternative preliminary tests.

Other important findings document that population web-search interest is a reliable indicator of disease prevalence and a linear model that links RSI and CRC prevalence at the world level can accurately nowcast the global burden of disease.

Estimates issued through the proposed framework for a 24-month forecasting horizon indicate that the global burden of colorectal cancer (CRC) can increase by over 10.6%, from about 11.46 million people living with CRC at the end of 2019 to more than 12.67 million cases by 2025.

Of note, the significant geographical and income-based disparity in CRC prevalence is documented on the global scene, with countries that belong to the East Asia and Pacific region and the upper-middle-income panel being most vulnerable to this burden. This finding in turn carries important implications and should be acknowledged in the process of issuing equitable global health policies.

Moreover, as the costs connected with CRC are extremely high for both individuals and society, current findings further emphasize the stringent need to mitigate this global burden through a focus on prevention and early detection. In this context, public policies aimed at increasing health expenditures and subsequent CRC screening, corroborated with public campaigns directed at educating the population to decrease exposure to the main modifiable (*i.e.*, non-genetic) risk factors such as alcohol consumption, tobacco use, overweight and obesity, poor dietary habits, a lack of regular physical activity, *etc.*, are paramount. In addition, the findings suggest that public health authorities should take

measures to increase cancer screening rates during pandemics, which would have positive externalities such as reducing the global burden and enhancing official statistics.

However, it is important to underline that projections should be revised regularly for healthcare planning and resource allocation purposes. Additionally, another limitation of the proposed framework should be acknowledged, as it does not include any further progress in treatment and/or improvement in screening rates over the forecasting horizon (*i.e.*, the framework only operates under the “*status quo*” hypothesis).

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Cristiana Tudor conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Robert Aurelian Sova conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Code and raw data are available in the [Supplemental Files](#).

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj-cs.1518#supplemental-information>.

REFERENCES

- Ades F, Senterre C, De Azambuja E, Sullivan R, Popescu R, Parent F, Piccart M. 2013. Discrepancies in cancer incidence and mortality and its relationship to health expenditure in the 27 European Union member states. *Annals of Oncology* 24(11):2897–2902 DOI 10.1093/annonc/mdt352.
- Allende H, Moraga C, Salas R. 2002. Artificial neural networks in time series forecasting: a comparative analysis. *Kybernetika* 38(6):685–707.
- Althouse BM, Scarpino SV, Meyers LA, Ayers JW, Bargsten M, Baumbach J, Brownstein JS, Castro L, Clapham H, Cummings DAT, Del Valle S, Eubank S, Fairchild G, Finelli L, Generous N, George D, Harper DR, Hébert-Dufresne L, Johansson MA, Konty K, Lipsitch M, Milinovich G, Miller JD, Nsoesie EO, Olson DR, Paul M, Polgreen PM, Priedhorsky R, Read JM, Rodriguez-Barraquer I, Smith DJ, Stefansen C, Swerdlow DL, Thompson D, Vespignani A, Wesolowski A. 2015. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Science* 4(1):1–8 DOI 10.1140/epjds/s13688-015-0054-0.

- American Cancer Society. 2022a.** *Cancer Treatment & survivorship facts & figures 2022-2024*. Atlanta: American Cancer Society.
- American Cancer Society. 2022b.** The global cancer burden. Available at <https://www.cancer.org/about-us/our-global-health-work/global-cancer-burden.html> (accessed 5 January 2023).
- Aras S, Kocakoç ID. 2016.** A new model selection strategy in time series forecasting with artificial neural networks: IHTS. *Neurocomputing* **174**(2):974–987 DOI 10.1016/j.neucom.2015.10.036.
- Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. 2017.** Global patterns and trends in colorectal cancer incidence and mortality. *Gut* **66**(4):683–691 DOI 10.1136/gutjnl-2015-310912.
- Arora VS, McKee M, Stuckler D. 2019.** Google trends: opportunities and limitations in health and health policy research. *Health Policy* **123**(3):338–341 DOI 10.1016/j.healthpol.2019.01.001.
- Atchadé MN, Sokadjo YM. 2022.** Overview and cross-validation of COVID-19 forecasting univariate models. *Alexandria Engineering Journal* **61**(4):3021–3036 DOI 10.1016/j.aej.2021.08.028.
- Bakouny Z, Hawley JE, Choueiri TK, Peters S, Rini BI, Warner JL, Painter CA. 2020.** COVID-19 and cancer: current challenges and perspectives. *Cancer Cell* **38**(5):629–646 DOI 10.1016/j.ccell.2020.09.018.
- Bakouny Z, Paciotti M, Schmidt AL, Lipsitz SR, Choueiri TK, Trinh QD. 2021.** Cancer screening tests and cancer diagnoses during the COVID-19 pandemic. *JAMA Oncology* **7**(3):458–460 DOI 10.1001/jamaoncol.2020.7600.
- Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. 2013.** Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *Journal of Medical Internet research* **15**(7):e2740 DOI 10.2196/jmir.2740.
- Bleher J, Dimpfl T. 2022.** Knitting multi-annual high-frequency google trends to predict inflation and consumption. *Econometrics and Statistics* **24**(3):1–26 DOI 10.1016/j.ecosta.2021.10.006.
- Borup D, Schütte ECM. 2022.** In search of a job: forecasting employment growth using Google Trends. *Journal of Business & Economic Statistics* **40**(1):186–200 DOI 10.1080/07350015.2020.1791133.
- Box G, Jenkins G. 1970.** *Time series analysis: forecasting and control*. San Francisco, CA, USA: Holden-Day.
- Bray F, Ren JS, Masuyer E, Ferlay J. 2013.** Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *International Journal of Cancer* **132**(5):1133–1145 DOI 10.1002/ijc.27711.
- Breiman L. 2001.** Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science* **16**(3):199–231 DOI 10.1214/ss/1009213725.
- Breitung J, Knüppel M. 2021.** How far can we forecast? Statistical tests of the predictive content. *Journal of Applied Econometrics* **36**(4):369–392 DOI 10.1002/jae.2817.
- Brown RG. 1959.** *Statistical forecasting for inventory control*. New York, NY, USA: McGraw Hill.
- Cancer Atlas. 2022.** The burden of cancer. Available at <https://canceratlas.cancer.org/the-burden/the-burden-of-cancer/> (accessed 5 January 2023).
- Capocaccia R, Colonna M, Corazziari I, De Angelis R, Francisci S, Micheli A, Mugno E. 2002.** Measuring cancer prevalence in Europe: the EUROPREVAL project. *Annals of Oncology* **13**(6):831–839 DOI 10.1093/annonc/mdf152.
- Cervantes A, Adam R, Roselló S, Arnold D, Normanno N, Taïeb J, Seligmann J, De Baere T, Osterlund P, Yoshino T, Martinelli E, ESMO Guidelines Committee. 2022.** Metastatic

- colorectal cancer: ESMO clinical practice guideline for diagnosis, treatment and follow-up. *Annals of Oncology* **34**(1):10–32 DOI 10.1016/j.annonc.2022.10.003.
- Chan KS. 1991.** Percentage points of likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)* **53**(3):691–696 DOI 10.1111/j.2517-6161.1991.tb01858.x.
- Chen SABS, Billings SA. 1992.** Neural networks for nonlinear dynamic system modelling and identification. *International Journal of Control* **56**(2):319–346 DOI 10.1080/00207179208934317.
- Cleveland WS. 1979.** Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **1979**(74):829–836 DOI 10.1080/01621459.1979.10481038.
- Cleveland WS, Devlin SJ. 1988.** Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**(403):596–610 DOI 10.1080/01621459.1988.10478639.
- Cohen J. 1988.** *Statistical power analysis for the behavioral sciences*. Second Edition. New York: Academic Press.
- Cohen J. 1992.** A power primer. *Psychological Bulletin* **112**(1):155–159 DOI 10.1037/0033-2909.112.1.155.
- Colditz GA, Dart H. 2013.** Massachusetts leads the nation in colorectal cancer screening: what lessons can we learn from Implementing prevention-translating epidemiology to practice. *Epidemiology* **3**:e111 DOI 10.4172/2161-1165.1000e111.
- Crispo A, Barba M, Malvezzi M, Arpino G, Grimaldi M, Rosso T, Esposito E, Sergi D, Ciliberto G, Giordano A, Montella M. 2013.** Cancer mortality trends between 1988 and 2009 in the metropolitan area of Naples and Caserta, Southern Italy: results from a joinpoint regression analysis. *Cancer Biology & Therapy* **14**(12):1113–1122 DOI 10.4161/cbt.26425.
- Diebold FX, Mariano RS. 1995.** Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**:253–263 DOI 10.1080/07350015.1995.10524599.
- EarthWeb. 2023.** Google searches per day in 2023, Available at: how many google searches per day in 2023? (Full Statistics). (accessed 22 May 2023). Available at <https://earthweb.com/how-many-google-searches-per-day/>.
- Eftimov T, Popovski G, Petković M, Seljak BK, Kocev D. 2020.** COVID-19 pandemic changes the food consumption patterns. *Trends in Food Science & Technology* **104**(3):268–272 DOI 10.1016/j.tifs.2020.08.017.
- Eichenauer VZ, Indergand R, Martínez IZ, Sax C. 2022.** Obtaining consistent time series from Google Trends. *Economic Inquiry* **60**(2):694–705 DOI 10.1111/ecin.13049.
- Eysenbach G. 2011.** Infodemiology and infoveillance: tracking online health information and cyberbehavior for public health. *American Journal of Preventive Medicine* **40**(5):S154–S158 DOI 10.1016/j.amepre.2011.02.006.
- Food and Drug Administration (FDA). 2021.** Colorectal cancer: what you should know about screening. Available at <https://www.fda.gov/consumers/consumer-updates/colorectal-cancer-what-you-should-know-about-screening>.
- Gaur K, Jagtap MM. 2022.** Role of artificial intelligence and machine learning in prediction, diagnosis, and prognosis of cancer. *Cureus* **14**(11):e31008 DOI 10.7759/cureus.31008.
- Gillis D, Edwards BP. 2019.** The utility of joinpoint regression for estimating population parameters given changes in population structure. *Heliyon* **5**(11):e02515 DOI 10.1016/j.heliyon.2019.e02515.

- Gregory RWBB, Warnes BB, Lodewijk B. 2016.** gplots: various R programming tools for plotting data. R package version 3(1). Available at <https://cran.r-project.org/web/packages/gplots/index.html> (accessed 4 January 2023).
- Greiner B, Tipton S, Nelson B, Hartwell M. 2022.** Cancer screenings during the COVID-19 pandemic: an analysis of public interest trends. *Current Problems in Cancer* **46(1)**:100766 DOI [10.1016/j.currproblcancer.2021.100766](https://doi.org/10.1016/j.currproblcancer.2021.100766).
- Holt CC. 1957.** *Forecasting seasonals and trends by exponentially weighted averages (O.N.R. Memorandum No. 52)*. Pittsburgh, PA, USA: Carnegie Institute of Technology.
- Hsieh WW. 2004.** Nonlinear multivariate and time series analysis by neural network methods. *Reviews of Geophysics* **42(1)**:17,437 DOI [10.1029/2002RG000112](https://doi.org/10.1029/2002RG000112).
- Hyndman RJ, Athanasopoulos G. 2018.** Evaluating forecast accuracy. In: *Forecasting: Principles and Practice*. Second Edition. Melbourne, Australia: OTexts.
- Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeeen F. 2022.** Forecast: forecasting functions for time series and linear models. R package Version 8.16. 2022. Available at <https://pkg.robjhyndman.com/forecast/> (accessed 10 January 2023).
- Hyndman RJ, Khandakar Y. 2008.** Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software* **26(3)**:1–22 DOI [10.18637/jss.v027.i03](https://doi.org/10.18637/jss.v027.i03).
- Hyndman RJ, Koehler AB, Snyder RD, Grose S. 2002.** A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting* **18(3)**:439–454 DOI [10.1016/S0169-2070\(01\)00110-8](https://doi.org/10.1016/S0169-2070(01)00110-8).
- International Agency for Research on Cancer (IARC). 2022.** Colorectal cancer awareness month 2022. Available at <https://www.iarc.who.int/featured-news/colorectal-cancer-awareness-month-2022/> (accessed 4 January 2023).
- Jacob L, Loosen SH, Kalder M, Luedde T, Roderburg C, Kostev K. 2021.** Impact of the COVID-19 pandemic on cancer diagnoses in general and specialized practices in Germany. *Cancers* **13(3)**:408 DOI [10.3390/cancers13030408](https://doi.org/10.3390/cancers13030408).
- Jaidka K, Eichstaedt J, Giorgi S, Schwartz HA, Ungar LH. 2021.** Information-seeking vs. sharing: which explains regional health? An analysis of Google Search and Twitter trends. *Telematics and Informatics* **59(4)**:101540 DOI [10.1016/j.tele.2020.101540](https://doi.org/10.1016/j.tele.2020.101540).
- Jakobsen E, Olsen KE, Bliddal M, Hornbak M, Persson GF, Green A. 2021.** Forecasting lung cancer incidence, mortality, and prevalence to year 2030. *BMC Cancer* **21(1)**:1–9 DOI [10.1186/s12885-021-08696-6](https://doi.org/10.1186/s12885-021-08696-6).
- Jun SP, Yoo HS, Choi S. 2018.** Ten years of research change using Google Trends: from the perspective of big data utilizations and applications. *Technological Forecasting and Social Change* **2018(130)**:69–87 DOI [10.1016/j.techfore.2017.11.009](https://doi.org/10.1016/j.techfore.2017.11.009).
- Kadakuntla A, Wang T, Medgyesy K, Rrapi E, Litynski J, Adynski G, Tadros M. 2021.** Colorectal cancer screening in the COVID-19 era. *World Journal of Gastrointestinal Oncology* **13(4)**:238–251 DOI [10.4251/wjgo.v13.i4.238](https://doi.org/10.4251/wjgo.v13.i4.238).
- Kamiński M, Łoniewski I, Marlicz W. 2020.** “Dr. Google, I am in Pain”—Global Internet Searches Associated with Pain: a retrospective analysis of Google trends data. *International Journal of Environmental Research and Public Health* **17(3)**:954 DOI [10.3390/ijerph17030954](https://doi.org/10.3390/ijerph17030954).
- Keum N, Giovannucci E. 2019.** Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature Reviews Gastroenterology & Hepatology* **16(12)**:713–732 DOI [10.1038/s41575-019-0189-8](https://doi.org/10.1038/s41575-019-0189-8).
- Kocarnik JM, Compton K, Dean FE, Fu W, Gaw BL, Harvey JD, Henrikson HJ, Lu D, Pennini A, Xu R, Ababneh E, Abbasi-Kangevari M, Abbastabar H, Abd-Elsalam SM, Abdoli A, Abedi**

- A, Abidi H, Abolhassani H, Adedeji IA, Adnani QES, Advani SM, Afzal MS, Aghaali M, Ahinkorah BO, Ahmad S, Ahmad T, Ahmadi A, Ahmadi S, Ahmed Rashid T, Ahmed Salih Y, Akalu GT, Aklilu A, Akram T, Akunna CJ, Al Hamad H, Alahdab F, Al-Aly Z, Ali S, Alimohamadi Y, Alipour V, Aljunid SM, Alkhayyat M, Almasi-Hashiani A, Almasri NA, Al-Maweri SAA, Almustanyir S, Alonso N, Alvis-Guzman N, Amu H, Anbesu EW, Ancuceanu R, Ansari F, Ansari-Moghaddam A, Antwi MH, Anvari D, Anyasodor AE, Aqeel M, Arabloo J, Arab-Zozani M, Aremu O, Ariffin H, Aripov T, Arshad M, Artaman A, Arulappan J, Asemi Z, Asghari Jafarabadi M, Ashraf T, Atorkey P, Aujayeb A, Ausloos M, Awedew AF, Ayala Quintanilla BP, Ayenew T, Azab MA, Azadnajafabad S, Azari Jafari A, Azarian G, Azzam AY, Badiye AD, Bahadory S, Baig AA, Baker JL, Balakrishnan S, Banach M, Bärnighausen TW, Barone-Adesi F, Barra F, Barrow A, Behzadifar M, Belgaumi UI, Bezabhe WMM, Bezabih YM, Bhagat DS, Bhagavathula AS, Bhardwaj N, Bhardwaj P, Bhaskar S, Bhattacharyya K, Bhojaraja VS, Bibi S, Bijani A, Biondi A, Bisignano C, Bjørge T, Bleyer A, Blyuss O, Bolarinwa OA, Bolla SR, Braithwaite D, Brar A, Brenner H, Bustamante-Teixeira MT, Butt NS, Butt ZA, Caetano dos Santos FL, Cao Y, Carreras G, Catalá-López Fán, Cembranel F, Cerin E, Cernigliaro A, Chakinala RC, Chattu SK, Chattu VK, Chaturvedi P, Chimed-Ochir O, Cho DY, Christopher DJ, Chu D-T, Chung MT, Conde J, Cortés S, Cortesi PA, Costa VM, Cunha AR, Dadras O, Dagnew AB, Dahlawi SMA, Dai X, Dandona L, Dandona R, Darwesh AM, das Neves Jé, De la Hoz FP, Demis AB, Denova-Gutiérrez E, Dhamnetiya D, Dhimal ML, Dhimal M, Dianatinasab M, Diaz D, Djalalinia S, Do HP, Doaei S, Dorostkar F, dos Santos Figueiredo FW, Driscoll TR, Ebrahimi H, Eftekharzadeh S, El Tantawi M, El-Abid H, Elbarazi I, Elhabashy HR, Elhadi M, El-Jaafary SI, Eshrati B, Eskandarieh S, Esmailzadeh F, Etemadi A, Ezzikouri S, Faisaluddin M, Faraon EJA, Fares J, Farzadfar F, Feroze AH, Ferrero S, Ferro Desideri L, Filip I, Fischer F, Fisher JL, Foroutan M, Fukumoto T, Gaal PA, Gad MM, Gadanya MA, Gallus S, Gaspar Fonseca M, Getachew Obsa A, Ghafourifard M, Ghashghae A, Ghith N, Gholamalizadeh M, Gilani SA, Ginindza TG, Gizaw ATT, Glasbey JC, Golechha M, Goleij P, Gomez RS. 2022. Cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life years for 29 cancer groups from 2010 to 2019: a systematic analysis for the Global Burden of Disease Study 2019. *JAMA oncology* 8(3):420–444 DOI 10.1001/jamaoncol.2021.6987.
- Lakens D. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology* 4:863 DOI 10.3389/fpsyg.2013.00863.
- Lee DK. 2016. Alternatives to P value: confidence interval and effect size. *Korean Journal of Anesthesiology* 69(6):555–562 DOI 10.4097/kjae.2016.69.6.555.
- Maddams J, Brewster D, Gavin A, Steward J, Elliott J, Utley M, Møller H. 2009. Cancer prevalence in the United Kingdom: estimates for 2008. *British Journal of Cancer* 101(3):541–547 DOI 10.1038/sj.bjc.6605148.
- Maddams J, Utley M, Møller H. 2012. Projections of cancer prevalence in the United Kingdom, 2010–2040. *British Journal of Cancer* 107(7):1195–1202 DOI 10.1038/bjc.2012.366.
- Marques NP, Silveira DMM, Marques NCT, Martelli DRB, Oliveira EA, Martelli-Júnior H. 2021. Cancer diagnosis in Brazil in the COVID-19 era. *Seminars in Oncology* 48:156–159 DOI 10.1053/j.seminoncol.2020.12.002.
- Massicotte P, Eddelbuettel D. 2022. gtrendsR: perform and display google trends queries. R package version 15.1. Available at <https://cran.r-project.org/web/packages/gtrendsR/gtrendsR.pdf>.
- Mavragani A. 2020. Infodemiology and infoveillance: scoping review. *Journal of Medical Internet Research* 22(4):e16206 DOI 10.2196/16206.
- Mavragani A, Ochoa G. 2019. Google Trends in infodemiology and infoveillance: methodology framework. *JMIR Public Health and Surveillance* 5(2):e13439 DOI 10.2196/13439.

- Mazidimoradi A, Tiznobaik A, Salehiniya H. 2022.** Impact of the COVID-19 pandemic on colorectal cancer screening: a systematic review. *Journal of Gastrointestinal Cancer* **53**(3):730–744 DOI [10.1007/s12029-021-00679-x](https://doi.org/10.1007/s12029-021-00679-x).
- Medeiros MC, Pires HF. 2021.** The proper use of google trends in forecasting models. *ArXiv preprint*. DOI [10.48550/arXiv.2104.03065](https://doi.org/10.48550/arXiv.2104.03065).
- Morgan E, Arnold M, Gini A, Lorenzoni V, Cabasag CJ, Laversanne M, Vignat J, Ferlay J, Murphy N, Bray F. 2023.** Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut* **72**(2):338–344 DOI [10.1136/gutjnl-2022-327736](https://doi.org/10.1136/gutjnl-2022-327736).
- Munim ZH. 2022.** State-space TBATS model for container freight rate forecasting with improved accuracy. *Maritime Transport Research* **3**:100057 DOI [10.1016/j.martra.2022.100057](https://doi.org/10.1016/j.martra.2022.100057).
- Munim ZH, Shakil MH, Alon I. 2019.** Next-day bitcoin price forecast. *Journal of Risk and Financial Management* **12**(2):103 DOI [10.3390/jrfm12020103](https://doi.org/10.3390/jrfm12020103).
- Narita MF, Yin R. 2018.** In search of information: use of google trends' data to narrow information gaps for low-income developing countries. *International Monetary Fund* **2018**(286) DOI [10.5089/9781484390177.001](https://doi.org/10.5089/9781484390177.001).
- National Cancer Institute. 2022.** Statistics and Graphs. Available at <https://cancercontrol.cancer.gov/ocs/statistics#statistics-footnote1> (accessed 22 May 2023).
- Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, Murugiah K. 2014.** The use of google trends in health care research: a systematic review. *PLOS ONE* **9**(10):e109583 DOI [10.1371/journal.pone.0109583](https://doi.org/10.1371/journal.pone.0109583).
- Ord JK, Koehler AB, Snyder RD. 1997.** Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association* **92**(440):1621–1629 DOI [10.1080/01621459.1997.10473684](https://doi.org/10.1080/01621459.1997.10473684).
- Pasini A. 2015.** Artificial neural networks for small dataset analysis. *Journal of Thoracic Disease* **7**(5):953 DOI [10.3978/j.issn.2072-1439.2015.04.61](https://doi.org/10.3978/j.issn.2072-1439.2015.04.61).
- Perone G. 2021.** Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy. *The European Journal of Health Economics* **123**(6):917–940 DOI [10.1007/s10198-021-01347-4](https://doi.org/10.1007/s10198-021-01347-4).
- Petropoulos F, Spiliotis E. 2021.** The wisdom of the data: getting the most out of univariate time series forecasting. *Forecasting* **3**(3):478–497 DOI [10.3390/forecast3030029](https://doi.org/10.3390/forecast3030029).
- Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. 2008.** Using internet searches for influenza surveillance. *Clinical Infectious Diseases* **47**(11):1443–1448 DOI [10.1086/593098](https://doi.org/10.1086/593098).
- Qiu D, Katanoda K, Marugame T, Sobue T. 2009.** A Joinpoint regression analysis of long-term trends in cancer mortality in Japan (1958-2004). *International Journal of Cancer* **124**(2):443–448 DOI [10.1002/ijc.23911](https://doi.org/10.1002/ijc.23911).
- Richards M, Anderson M, Carter P, Ebert BL, Mossialos E. 2020.** The impact of the COVID-19 pandemic on cancer care. *Nature Cancer* **1**(6):565–567 DOI [10.1038/s43018-020-0074-y](https://doi.org/10.1038/s43018-020-0074-y).
- Rosenthal R. 1994.** Parametric measures of effect size. In: Cooper H, Hedges LV, eds. *The Handbook of Research Synthesis*. New York, NY: Sage, 231–244.
- Rovetta A. 2021.** Reliability of Google Trends: analysis of the limits and potential of web infoveillance during COVID-19 pandemic and for future research. *Frontiers in Research Metrics and Analytics* **6**:670226 DOI [10.3389/frma.2021.670226](https://doi.org/10.3389/frma.2021.670226).
- Rovetta A. 2023.** Common statistical errors in scientific investigations: a simple guide to avoid unfounded decisions. *Cureus* **15**(1):e33351 DOI [10.7759/cureus.33351](https://doi.org/10.7759/cureus.33351).
- Rovetta A, Castaldo L. 2021.** Influence of mass media on Italian web users during the COVID-19 pandemic: infodemiological analysis. *JMIRx Med* **2**(4):e32233 DOI [10.2196/32233](https://doi.org/10.2196/32233).

- Saini KS, de las Heras Bña, de Castro J, Venkitaraman R, Poelman M, Srinivasan G, Saini ML, Verma S, Leone M, Aftimos P, Curigliano G. 2020. Effect of the COVID-19 pandemic on cancer treatment and research. *The Lancet Haematology* 7(6):e432–e435 DOI 10.1016/S2352-3026(20)30123-X.
- Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, Campbell EM, Cattuto C, Khandelwal S, Mabry PL, Vespignani A. 2012. Digital epidemiology. *PLOS Computational Biology* 8(7):e1002616 DOI 10.1371/journal.pcbi.1002616.
- Sarakarn P, Suwanrungruang K, Vatanasapt P, Wiangnon S, Promthet S, Jenwitheesuk K, Chen THH. 2017. Joinpoint analysis trends in the incidence of colorectal cancer in Khon Kaen, Thailand (1989-2012). *Asian Pacific Journal of Cancer Prevention: APJCP* 18(4):1039 DOI 10.22034/APJCP.2017.18.4.1039.
- Sarangapani J. 2018. *Neural network control of nonlinear discrete-time systems*. Boca Raton: CRC press.
- Sato K, Mano T, Iwata A, Toda T. 2021. Need of care in interpreting Google Trends-based COVID-19 infodemiological study results: potential risk of false-positivity. *BMC Medical Research Methodology* 21(1):1–10 DOI 10.1186/s12874-021-01338-2.
- Satpathy P, Kumar S, Prasad P. 2023. Suitability of Google Trends™ for Digital surveillance during ongoing COVID-19 epidemic: a case study from India. *Disaster Medicine and Public Health Preparedness* 17:e28 DOI 10.1017/dmp.2021.249.
- Schootman M, Toor A, Cavazos-Rehg P, Jeffe DB, McQueen A, Eberth J, Davidson NO. 2015. The utility of Google Trends data to examine interest in cancer screening. *BMJ Open* 5(6):e006678 DOI 10.1136/bmjopen-2014-006678.
- Semenoglou AA, Spiliotis E, Assimakopoulos V. 2023. Data augmentation for univariate time series forecasting with neural networks. *Pattern Recognition* 134(3):109132 DOI 10.1016/j.patcog.2022.109132.
- Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 3(4):591–611 DOI 10.2307/2333709.
- Sharpless NE. 2020. COVID-19 and cancer. *Science* 368(6497):1290 DOI 10.1126/science.abd3377.
- Siegel RL, Miller KD, Jemal A. 2019. Cancer statistics. 2019 CA: a Cancer Journal for Clinicians 69(1):7–34 DOI 10.3322/caac.21551.
- Silva ES, Hassani H, Madsen DØ, Gee L. 2019. Googling fashion: forecasting fashion consumer behaviour using google trends. *Social Sciences* 8(4):111 DOI 10.3390/socsci8040111.
- Smittenaar CR, Petersen KA, Stewart K, Moitt N. 2016. Cancer incidence and mortality projections in the UK until 2035. *British Journal of Cancer* 115(9):1147–1155 DOI 10.1038/bjc.2016.304.
- Statista. 2023. Market share of leading search engines worldwide from January 2015 to April 2023. Available at <https://www.statista.com/statistics/1381664/worldwide-all-devices-market-share-of-search-engines/> (accessed 22 May 2023).
- Sulyok M, Ferenci T, Walker M. 2021. Google Trends Data and COVID-19 in Europe: correlations and model enhancement are European wide. *Transboundary and Emerging Diseases* 68(4):2610–2615 DOI 10.1111/tbed.13887.
- Szilagyi IS, Ullrich T, Lang-Illievich K, Klivinyi C, Schittek GA, Simonis H, Bornemann-Cimenti H. 2021. Google Trends for pain search terms in the world's most populated regions before and after the first recorded COVID-19 case: infodemiological study. *Journal of Medical Internet Research* 23(4):e27214 DOI 10.2196/27214.

- Thun MJ, DeLancey JO, Center MM, Jemal A, Ward EM. 2010. The global burden of cancer: priorities for prevention. *Carcinogenesis* 31(1):100–110 DOI 10.1093/carcin/bgp263.
- Tkachenko N, Chotvijit S, Gupta N, Bradley E, Gilks C, Guo W, Crosby H, Shore E, Thiarai M, Procter R, Jarvis S. 2017. Google Trends can improve surveillance of Type 2 diabetes. *Scientific Reports* 7(1):4993 DOI 10.1038/s41598-017-05091-9.
- Torchiano M. 2020. *effsize: efficient effect size computation*. R package version 0.8.1. Available at <https://CRAN.R-project.org/package=effsize>.
- Tran KB, Lang JJ, Compton K, Xu R, Acheson AR, Henrikson HJ, Kocarnik JM, Penberthy L, Aali A, Abbas Q, Abbasi B, Abbasi-Kangevari M, Abbasi-Kangevari Z, Abastabar H, Abdelmasseh M, Abd-Elsalam S, Abdelwahab AA, Abdoli G, Abdulkadir HA, Abedi A, Abegaz KH, Abidi H, Aboagye RG, Abolhassani H, Absalan A, Abtew YD, Abubaker Ali H, Abu-Gharbieh E, Achappa B, Acuna JM, Addison D, Addo IY, Adegboye OA, Adesina MA, Adnan M, Adnani QES, Advani SM, Afrin S, Afzal MS, Aggarwal M, Ahinkorah BO, Ahmad AR, Ahmad R, Ahmad S, Ahmad S, Ahmadi S, Ahmed H, Ahmed LA, Ahmed MB, Ahmed Rashid T, Aiman W, Ajami M, Akalu GT, Akbarzadeh-Khiavi M, Aklilu A, Akonde M, Akunna CJ, Al Hamad H, Alahdab F, Alanezi FM, Alanzi TM, Alessy SA, Algammal AM, Al-Hanawi MK, Alhassan RK, Ali BA, Ali L, Ali SS, Alimohamadi Y, Alipour V, Aljunid SM, Alkhayyat M, Al-Maweri SAA, Almustanyir S, Alonso N, Alqalyoobi S, Al-Raddadi RM, Al-Rifai RHH, Al-Sabah SK, Al-Tammemi AB, Altawalah H, Alvis-Guzman N, Amare F, Ameyaw EK, Aminian Dehkordi JJ, Amirzade-Iranaq MH, Amu H, Amusa GA, Ancuceanu R, Anderson JA, Animut YA, Anoushiravani A, Anoushirvani AA, Ansari-Moghaddam A, Ansha MG, Antony B, Antwi MH, Anwar SL, Anwer R, Anyasodor AE, Arabloo J, Arab-Zozani M, Aremu O, Argaw AM, Ariffin H, Aripov T, Arshad M, Artaman A, Arulappan J, Aruleba RT, Aryannejad A, Asaad M, Asemahagn MA, Asemi Z, Asghari-Jafarabadi M, Ashraf T, Assadi R, Athar M, Athari SS, Atout MMDW, Attia S, Aujayeb A, Ausloos M, Avila-Burgos L, Awedew AF, Awoke MA, Awoke T, Ayala Quintanilla BP, Ayana TM, Ayen SS, Azadi D, Azadnajafabad S, Azami-Aghdash S, Azanaw MM, Azangou-Khyavy M, Azari Jafari A, Azizi H, Azzam AYY, Babajani A, Badar M, Badiye AD, Baghcheghi N, Bagheri N, Bagherieh S, Bahadory S, Baig AA, Baker JL, Bakhtiari A, Bakshi RK, Banach M, Banerjee I, Bardhan M, Barone-Adesi F, Barra F, Barrow A, Bashir NZ, Bashiri A, Basu S, Batiha A-MM, Begum A, Bekele AB, Belay AS, Belete MA, Belgaumi UI, Bell AW, Belo L, Benzian H, Berhie AY, Bermudez ANC, Bernabe E, Bhagavathula AS, Bhala N, Bhandari BB, Bhardwaj N, Bhardwaj P, Bhattacharyya K, Bhojaraja VS, Bhuyan SS, Bibi S, Bilchut AH, Bintoro BS, Biondi A, Birega MGB, Birhan HE, Bjørge T, Blyuss O, Bodicha BBA, Bolla SR, Bolor A, Bosetti C, Braithwaite D, Brauer M, Brenner H, Briko AN, Briko NI, Buchanan CM, Bulamu NB, Bustamante-Teixeira MT, Butt MH, Butt NS. 2022. The global burden of cancer attributable to risk factors, 2010–19: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* 400(10352):563–591 DOI 10.1016/S0140-6736(22)01438-6.
- Trinh TTK, Lee YY, Suh M, Jun JK, Choi KS. 2022. Changes in cancer screening before and during COVID-19: findings from the Korean national cancer screening survey 2019 and 2020. *Epidemiology and Health* 44:e2022051 DOI 10.4178/epih.e2022051.
- Tsay RS. 1986. Nonlinearity tests for time series. *Biometrika* 73(2):461–466 DOI 10.1093/biomet/73.2.461.
- Tudor C. 2022a. A novel approach to modeling and forecasting cancer incidence and mortality rates through web queries and automated forecasting algorithms: evidence from romania. *Biology* 11(6):857 DOI 10.3390/biology11060857.
- Tudor C. 2022b. The impact of the COVID-19 pandemic on the global web and video conferencing SaaS market. *Electronics* 11(16):2633 DOI 10.3390/electronics11162633.

- Tudor C, Sova R. 2022.** Infodemiological study on the impact of the COVID-19 pandemic on increased headache incidences at the world level. *Scientific Reports* **12**(1):1–15 DOI [10.1038/s41598-022-13663-7](https://doi.org/10.1038/s41598-022-13663-7).
- Uhlig J, Cecchini M, Sheth A, Stein S, Lacy J, Kim HS. 2021.** Microsatellite instability and KRAS mutation in stage IV colorectal cancer: prevalence, geographic discrepancies, and outcomes from the national cancer database. *Journal of the National Comprehensive Cancer Network* **19**(3):307–318 DOI [10.6004/jnccn.2020.7619](https://doi.org/10.6004/jnccn.2020.7619).
- United Nations. 2022.** Sustainable development goals. Available at <https://sdgs.un.org> (accessed 2 December 2022).
- Wilson L, Bhatnagar P, Townsend N. 2017.** Comparing trends in mortality from cardiovascular disease and cancer in the United Kingdom, 1983-2013: joinpoint regression analysis. *Population Health Metrics* **15**(1):1–9 DOI [10.1186/s12963-017-0141-5](https://doi.org/10.1186/s12963-017-0141-5).
- Winters PR. 1960.** Forecasting sales by exponentially weighted moving averages. *Management Science* **6**:324 DOI [10.1007/978-3-642-51565-1](https://doi.org/10.1007/978-3-642-51565-1).
- World Health Organization (WHO). 2022.** Cancer. Available at <https://www.who.int/news-room/fact-sheets/detail/cancer> (accessed 2 December 2022).
- Xi Y, Xu P. 2021.** Global colorectal cancer burden in 2020 and projections to 2040. *Translational Oncology* **14**(10):101174 DOI [10.1016/j.tranon.2021.101174](https://doi.org/10.1016/j.tranon.2021.101174).
- Xie YH, Chen YX, Fang JY. 2020.** Comprehensive review of targeted therapy for colorectal cancer. *Signal Transduction and Targeted Therapy* **5**(1):1–30 DOI [10.1038/s41392-020-0116-z](https://doi.org/10.1038/s41392-020-0116-z).
- Yang D, Sharma V, Ye Z, Lim LI, Zhao L, Aryaputera AW. 2015.** Forecasting of global horizontal irradiance by exponential smoothing, using decompositions. *Energy* **2015**(81):111–119 DOI [10.1016/j.energy.2014.11.082](https://doi.org/10.1016/j.energy.2014.11.082).
- Ziel F, Steinert R, Husmann S. 2015.** Efficient modeling and forecasting of electricity spot prices. *Energy Economics* **47**(1):98–111 DOI [10.1016/j.eneco.2014.10.012](https://doi.org/10.1016/j.eneco.2014.10.012).
- Ziel F, Weron R. 2018.** Day-ahead electricity price forecasting with high-dimensional structures: univariate vs. multivariate modeling frameworks. *Energy Economics* **70**(8):396–420 DOI [10.1016/j.eneco.2017.12.016](https://doi.org/10.1016/j.eneco.2017.12.016).
- World Cancer Research Fund International (WCRF). 2022.** Worldwide cancer data. Available at <https://www.wcrf.org/cancer-trends/worldwide-cancer-data/> (accessed 22 May 2023).
- Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K, Yutani H, Dunnington D. 2023.** ggplot2: create elegant data visualizations using the grammar of graphics version 3.4.3. Available at <https://CRAN.R-project.org/package=ggplot2>.