scientific reports

Check for updates

OPEN

Future world cancer death rate prediction

Oleg Gaidai[1], Ping Yan[1] & Yihan Xing[2]✉

Cancer is a worldwide illness that causes significant morbidity and death and imposes an immense cost on global public health. Modelling such a phenomenon is complex because of the non-stationarity and complexity of cancer waves. Apply modern novel statistical methods directly to raw clinical data. To estimate extreme cancer death rate likelihood at any period in any location of interest. Traditional statistical methodologies that deal with temporal observations of multi-regional processes cannot adequately deal with substantial regional dimensionality and cross-correlation of various regional variables. Setting: multicenter, population-based, medical survey data-based biostatistical approach. Due to the non-stationarity and complicated nature of cancer, it is challenging to model such a phenomenon. This paper offers a unique bio-system dependability technique suited for multi-regional environmental and health systems. When monitored over a significant period, it yields a reliable long-term projection of the chance of an exceptional cancer mortality rate. Traditional statistical approaches dealing with temporal observations of multi-regional processes cannot effectively deal with large regional dimensionality and cross-correlation between multiple regional data. The provided approach may be employed in numerous public health applications, depending on their clinical survey data.

The National Cancer Institute defines cancer as a group of disorders in which aberrant cells may proliferate and invade neighbouring tissue. Cancer may develop in most regions of the body, resulting in various cancer forms, as indicated below, and can sometimes spread via the blood and lymph systems.

Cancer's statistical characteristics received much attention from the current scientific community [1–8]. Using current theoretical statistical methods [9–15], it is often rather challenging to compute realistic biological system dependability factors and outbreak probability under actual cancer settings. Typically, this results from many degrees of system freedom and random variables driving vastly dispersed dynamic biological systems. In theory, the dependability of a complex biological system may be precisely evaluated using sufficient observations or direct Monte Carlo simulations. Beginning in 1990, however, a portion of the available cancer observation numbers are limited[16–21]. Motivated by the latter point, the authors have developed a unique dependability technique for biological and health systems to forecast and control cancer epidemics more precisely. The whole globe was selected because of the enormous internet health observations and associated research[1].

In health and engineering fields, statistical modelling of lifetime data and extreme value theory (EVT) are widespread. For example, Gumbel utilised EVT to predict the demography of distinct communities in[20–23]. Recent papers arguing for and against the upper bounds distribution of life expectancy were done by[24]. Often, papers in these fields presume a parametric bivariate lifetime distribution obtained from the exponential distribution to get statistically relevant data[24]. In[25], the author proposes a new approach that uses Power Variance Function copulas (e.g., Clayton, Gumbel and Inverse Gaussian copulas), conditional sampling, and numerical approximation used in survival analysis. While in a paper by[26], the authors explain that EVT has been used to predict mutation in evolutionary genetics and further develop a likelihood framework from EVT that was used to determine the fitness effects of the mutation.

Similarly, in[27], The author applies a Beta-Burr distribution to this EVT hypothesis to calculate the fitness impact. While in[28], the author presents a bivariate logistic regression model, which was afterwards used to access multiple MS fatalities with walking difficulties and in a cognitive experiment for visual identification. Finally[3], is a relevant work utilising EVT to evaluate the chance of a global cancer breakout. In[22,23], similarly, researchers employed EVT to predict and identify cancer abnormalities.

In this research, a cancer outbreak is seen as an unanticipated occurrence that may occur in any location of a nation at any moment; hence, the spatial spread is considered. Moreover, a specific non-dimensional factor λ is introduced to forecast the cancer risk at any given time and location. Environmental impacts on biological systems are ergodic. The second possibility is to see the process as reliant on specific external characteristics

[1]Shanghai Ocean University, Shanghai, China. [2]University of Stavanger, Stavanger, Norway. ✉email: yihan.xing@uis.no

whose time-dependent change may be modelled as an ergodic process on its own. The incidence data of cancer in one hundred ninety-five world countries during the years 1990–2019 were retrieved from the public website[1], considered a multi-degree-of-freedom (MDOF) spatio-temporal dynamic bio-system with highly inter-correlated regional components/dimensions.

This research tries to reduce the danger of future cancer outbreaks by forecasting them. However, it focuses simply on the yearly number of documented patient deaths and not on the symptoms themselves. Figure 1 presents the map of the world's countries.

Further research should incorporate one of the common complexity measures, such as fractal, attractor/embedding dimension, and entropy.

Methods

Consider an MDOF (multi-degree of freedom) structure subjected to random ergodic environmental factors (stationary in time). The second possibility is to see the process as reliant on certain external characteristics whose time-dependent change may be modelled as an ergodic process on its own. The MDOF biomedical response vector process $\boldsymbol{R}(t) = (X(t), Y(t), Z(t), \dots)$ is measured and/or simulated over a sufficiently long time interval $(0, T)$. Unidimensional global maxima over the duration of time $(0, T)$ are denoted as $X_T^{\max} = \max_{0\le t\le T} X(t)$, $Y_T^{\max} = \max_{0\le t\le T} Y(t), Z_T^{\max} = \max_{0\le t\le T} Z(t), \dots$..By sufficiently long time T one primarily means a large value of T with respect to the dynamic system auto-correlation time[33–40].

Let $X_1, \dots, X_{N_X}$ be consequent in time local maxima of the process $X(t)$ at monotonously increasing discrete time instants $t_1^X < \cdots < t_{N_X}^X$ in $(0, T)$. The analogous definition follows for other MDOF response components $Y(t), Z(t), \dots$ with $Y_1, \dots, Y_{N_Y}; Z_1, \dots, Z_{N_Z}$ and so on. For simplicity, all $\boldsymbol{R}(t)$ components, and therefore its maxima are assumed to be non-negative. The aim is to estimate the system failure probability

$$1 - P = \mathrm{Prob}(X_T^{\max} > \eta_X \cup Y_T^{\max} > \eta_Y \cup Z_T^{\max} > \eta_Z \cup \dots) \tag{1}$$

with

$$P = \int_{(0,0,0,\dots)}^{(\eta_X,\eta_Y,\eta_Z,\dots)} \iiint p_{X_T^{\max},Y_T^{\max},Z_T^{\max},\dots}\left(X_T^{\max}, Y_T^{\max}, Z_T^{\max}, \dots\right) dX_T^{\max} dY_{N_Y}^{\max} dZ_{N_z}^{\max} \dots \tag{2}$$

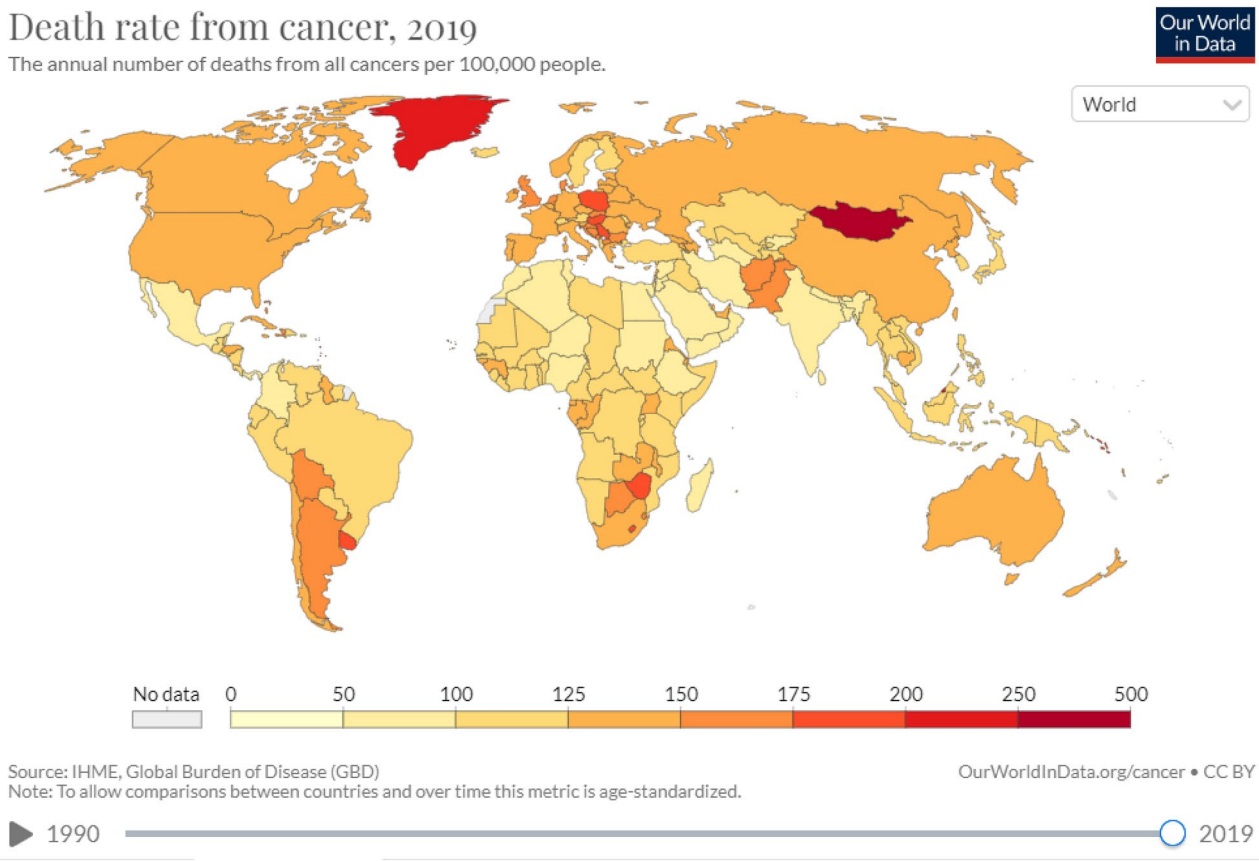

Figure 1. Map of the world with countries and cancer deaths. All world countries were studied in this paper[1].

being the probability of non-exceedance for response components $\eta_X, \eta_Y, \eta_Z, \ldots$ critical values; $\cup$ denotes logical unity operation; and $p_{X_T^{\max}, Y_T^{\max}, Z_T^{\max}, \ldots}$ being joint probability density of the global maxima over the entire time span $(0, T)$.

In practice, it is not possible to accurately estimate the latter joint probability distribution $p_{X_T^{\max}, Y_T^{\max}, Z_T^{\max}, \ldots}$ due to its high dimensionality and available data set limitations. In other words, the time instant when either $X(t)$ exceeds η_X, or $Y(t)$ exceeds η_Y, or $Z(t)$ exceeds η_Z, and so on, the system being regarded as immediately failed. Fixed failure levels η_X, η_Y, η_Z,...are of course individual for each unidimensional response component of $\boldsymbol{R}(t)$. $X_{N_X}^{\max} = \max\{X_j; j = 1, \ldots, N_X\} = X_T^{\max}$, $Y_{N_Y}^{\max} = \max\{Y_j; j = 1, \ldots, N_Y\} = Y_T^{\max}$,$Z_{N_z}^{\max} = \max\{Z_j; j = 1, \ldots, N_Z\} = Z_T^{\max}$, and so on.

Next, the local maxima time instants $\left[t_1^X < \cdots < t_{N_X}^X; t_1^Y < \cdots < t_{N_Y}^Y; t_1^Z < \cdots < t_{N_Z}^Z\right]$ in monotonously non-decreasing order are sorted into one single merged time vector $t_1 \leq \cdots \leq t_N$.Note that $t_N = \max\{t_{N_X}^X, t_{N_Y}^Y, t_{N_Z}^Z, \ldots\}, N = N_X + N_Y + N_Z + \ldots$. In this case t_j represents local maxima of one of MDOF bio-system response components either $X(t)$ or $Y(t)$, or $Z(t)$ and so on. That means that having $\boldsymbol{R}(t)$ time record, one just has to continually and concurrently screen for local maximums of unidimensional response components and record their exceeding the MDOF limit vector $(\eta_X, \eta_Y, \eta_Z, ...)$ in any of its components $X, Y, Z, \ldots$. The maxima of local unidimensional response components are blended into a non-decreasing temporal vector $\overrightarrow{R} = (R_1, R_2, \ldots, R_N)$ in accordance with the merged time vector $t_1 \leq \cdots \leq t_N$. That is to say, each local maxima R_j is the actual encountered local maxima corresponding to either $X(t)$ or $Y(t)$, or $Z(t)$ and so on. Finally, the unified limit vector $(\eta_1, \ldots, \eta_N)$ is introduced with each component η_j is either η_X, η_Y or η_Z and so on, depending on which of $X(t)$ or $Y(t)$, or $Z(t)$ etc., corresponding to the current local maxima with the running index j.

Next, a scaling parameter $0 < \lambda \leq 1$ is implemented to artificially lower limit values for all response components concurrently, namely the new MDOF limit vector $\left(\eta_X^\lambda, \eta_Y^\lambda, \eta_z^\lambda, ...\right)$ with $\eta_X^\lambda \equiv \lambda \cdot \eta_X, \equiv \lambda \cdot \eta_Y, \eta_z^\lambda \equiv \lambda \cdot \eta_Z$, … is introduced. The unified limit vector $\left(\eta_1^\lambda, \ldots, \eta_N^\lambda\right)$ is introduced with each component η_j^λ is either η_X^λ, η_Y^λ or η_z^λ and so on. The latter automatically defines probability $P(\lambda)$ as a function of λ, note that $P \equiv P(1)$ from Eq. (1). Non-exceedance probability $P(\lambda)$ can be now estimated as follows

$$\begin{aligned} P(\lambda) &= \text{Prob}\left\{R_N \leq \eta_N^\lambda, \ldots, R_1 \leq \eta_1^\lambda\right\} \\ &= \text{Prob}\{R_N \leq \eta_N^\lambda | R_{N-1} \leq \eta_{N-1}^\lambda, \ldots, R_1 \leq \eta_1^\lambda\} \cdot \text{Prob}\left\{R_{N-1} \leq \eta_{N-1}^\lambda, \ldots, R_1 \leq \eta_1^\lambda\right\} \\ &= \left(\prod_{j=2}^{N} \text{Prob}\{R_j \leq \eta_j^\lambda | R_{j-1} \leq \eta_{1j-}^\lambda, \ldots, R_1 \leq \eta_1^\lambda\}\right) \cdot \text{Prob}\left(R_1 \leq \eta_1^\lambda\right) \end{aligned} \tag{3}$$

In practice, a dependency between neighbouring R_j is not always negligible; thus, the following one-step (called here conditioning level $k = 1$) memory approximation is introduced

$$\text{Prob}\{R_j \leq \eta_j^\lambda | R_{j-1} \leq \eta_{j-1}^\lambda, \ldots, R_1 \leq \eta_1^\lambda\} \approx \text{Prob}\{R_j \leq \eta_j^\lambda | R_{j-1} \leq \eta_{j-1}^\lambda\} \tag{4}$$

for $2 \leq j \leq N$ (called here conditioning level $k = 2$). The approximation introduced by Eq. (4) can be further expressed as

$$\text{Prob}\{R_j \leq \eta_j^\lambda | R_{j-1} \leq \eta_{j-1}^\lambda, \ldots, R_1 \leq \eta_1^\lambda\} \approx \text{Prob}\{R_j \leq \eta_j^\lambda | R_{j-1} \leq \eta_{j-1}^\lambda, R_{j-2} \leq \eta_{j-2}^\lambda\} \tag{5}$$

where $3 \leq j \leq N$ (will be called conditioning level $k = 3$), and so on. The goal is to monitor each isolated failure that occurs locally first in time, thereby preventing cascade local inter-correlated exceedances.

Equation (5) presents subsequent refinements of the statistical independence assumption. The latter type of approximation enables capturing the statistical dependence effect between neighbouring maxima with increased accuracy. Since the original MDOF bio-process $\boldsymbol{R}(t)$ was assumed ergodic and therefore stationary, the probability $p_k(\lambda)$: = Prob $\{R_j > \eta_j^\lambda \mid R_{j-1} \leq \eta_{j-1}^\lambda,\ R_{j-k+1} \leq \eta_{j-k+1}^\lambda\}$ for $j \geq k$ will be independent of j but only dependent on conditioning level k. Thus non-exceedance probability can be approximated as in the Naess-Gaidai method[29,30], where

$$P_k(\lambda) \approx \exp(-N \cdot p_k(\lambda)), k \geq 1 \tag{6}$$

Note that Eq. (6) follows from Eq. (1) by neglecting $\text{Prob}(R_1 \leq \eta_1^\lambda) \approx 1$, as the design failure probability is usually very small. Further, it is assumed N"k. Note that Eq. (5) is similar to the well-known mean up-crossing rate equation for the probability of exceedance[32]. There is obvious convergence with respect to the conditioning parameter k

$$P = \lim_{k \to \infty} P_k(1); p(\lambda) = \lim_{k \to \infty} p_k(\lambda) \tag{7}$$

Note that Eq. (6) for $k = 1$ turns into the quite well-known non-exceedance probability relationship with the mean up-crossing rate function

$$P(\lambda) \approx \exp(-\nu^+(\lambda)\, T); \nu^+(\lambda) = \int_0^\infty \zeta p_{R\dot{R}}(\lambda, \zeta) d\zeta \tag{8}$$

where $\nu^+(\lambda)$ is the mean up-crossing rate of the response level λ for the above assembled non-dimensional vector $R(t)$ assembled from scaled MDOF bio-system response $\left(\frac{X}{\eta_X}, \frac{Y}{\eta_Y}, \frac{Z}{\eta_Z}, \ldots\right)$. Note that constructed $\overrightarrow{R}$-vector has no data loss at all; see Fig. 2.

In the preceding, the assumption of stationarity has been employed. The proposed methodology can also treat the non-stationary case. An illustration of how the methodology can be used to treat non-stationary cases is provided. Consider a scattered diagram of $m = 1, .., M$ environmental states, each short-term bio-environmental state having a probability q_m, so that $\sum_{m=1}^{M} q_m = 1$. The corresponding long-term equation is then

$$p_k(\lambda) \equiv \sum_{m=1}^{M} p_k(\lambda, m) q_m \tag{9}$$

with $p_k(\lambda, m)$ being the same function as in Eq. (7) but corresponding to a specific short-term environmental state with the number m. The above introduced $p_k(\lambda)$ as functions are often regular in the tail, specifically for values of λ approaching and exceeding 1. More precisely, for $\lambda \geq \lambda_0$, the distribution tail behaves similarly to $\exp\{-(a\lambda + b)^c + d\}$ with a, b, c, d being suitably fitted constants for suitable tail cut-on λ_0 value. Therefore, one can write

$$p_k(\lambda) \approx \exp\{-(a_k\lambda + b_k)^{c_k} + d_k\}, \lambda \geq \lambda_0 \tag{10}$$

Next, by plotting $\ln\{\ln(p_k(\lambda)) - d_k\}$ versus $\ln(a_k\lambda + b_k)$, often nearly perfectly linear tail behaviour is observed. Optimal values of the parameters a_k, b_k, c_k, p_k, q_k may also be determined using a sequential quadratic programming (SQP) method incorporated in the NAG Numerical Library[31].

For levels of λ approaching 1, the approximate limits of a p-% confidence interval (CI) of $p_k(\lambda)$ can be given as follows[41–46]

$$\mathrm{CI}^{\pm}(\lambda) = p_k(\lambda)(1 \pm \frac{f(p)}{\sqrt{(N-k+1)p_k(\lambda)}}). \tag{11}$$

with $f(p)$ being estimated from the inverse normal distribution, for example, $f(90\%) = 1.65$, $f(95\%) = 1.96$. with N being the total number of local maxima assembled in the analysed vector $\overrightarrow{R}$.

Results

Predictions of cancer-related mortality have been the focus of epidemiology and mathematical biology for a long time. It is common knowledge that the dynamics of public health are a highly non-linear, multidimensional, spatially cross-correlated dynamic system that is always difficult to analyse. Previous studies have used a variety of approaches to model cancer cases. This section presents the application of the above-described methodology to the real-life cancer data sets, presented as a new annual recorded time series for all world countries. The statistical information presented in this section was obtained from the official World website[1]. The website provides cancer death rates per country from 1990 to 2019. Patient death numbers from one hundred ninety-five different world countries were chosen as components $X, Y, Z, ...$, thus constituting an example of a one hundred ninety-five dimensional (195D) dynamic biological system. To unify all 195 measured time series $X, Y, Z, \ldots$ the following scaling was performed

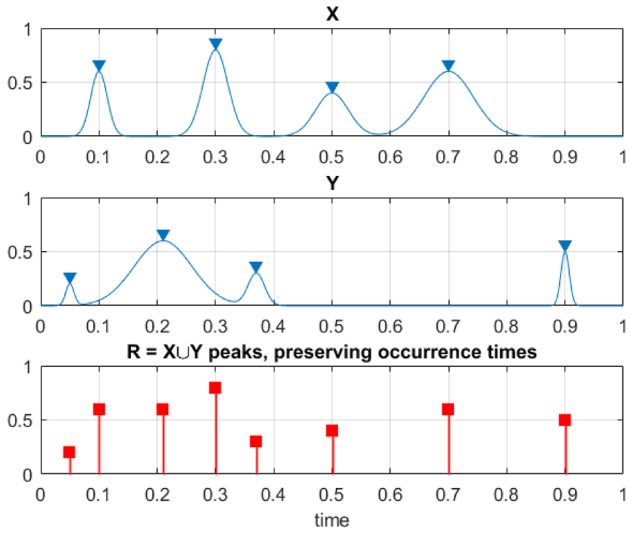

Figure 2. Example of how two example processes, X and Y, are merged to create a new synthetic vector $\overrightarrow{R}$.

$$X \to \frac{X}{\eta_X}, Y \to \frac{Y}{\eta_Y}, Z \to \frac{Z}{\eta_Z}, \ldots \tag{12}$$

making all 195 responses non-dimensional and having the same failure limit equal to 1. Failure limits $\eta_X, \eta_Y, \eta_Z, \ldots$, or in other words, cancer thresholds, are not an obvious choice. The most straightforward choice would be for different countries to set failure limits equal to the corresponding country population in per cent to local population, basically making $X, Y, Z, \ldots$ equal to the annual death rate per country. Next, all local maxima from 195 measured time series were merged into one single time series by keeping them in time non-decreasing order: $\overrightarrow{R} = (\max\{X_1, Y_1, Z_1, \ldots\}, \ldots, \max\{X_N, Y_N, Z_N, \ldots\})$ with the whole vector $\overrightarrow{R}$ being sorted according to non-decreasing times of occurrence of these local maxima.

Figure 3 presents the number of new annual recorded deaths as a 195D vector $\overrightarrow{R}$, consisting of assembled regional new annual death rate for each corresponding country. Greenland, Mongolia, Monaco and Hungary data were excluded from analysis, since were regarded as outliers. Note that vector $\overrightarrow{R}$ is assembled of different regional components with different cancer backgrounds. Index j is just a running index of local maxima encountered in a non-decreasing time sequence.

Figure 4 presents the annual death rate (percentage of deaths from cancer to the population of a given country) prediction, 100 years return level extrapolation according to Eq. (10) towards cancer outbreak with a 100-year return period, indicated by the horizontal dotted line. Somewhat beyond, $\lambda = 0.18\%$ cut-on value was used, percentage of the local population on the horizontal axis. The dotted lines indicate extrapolated 95%

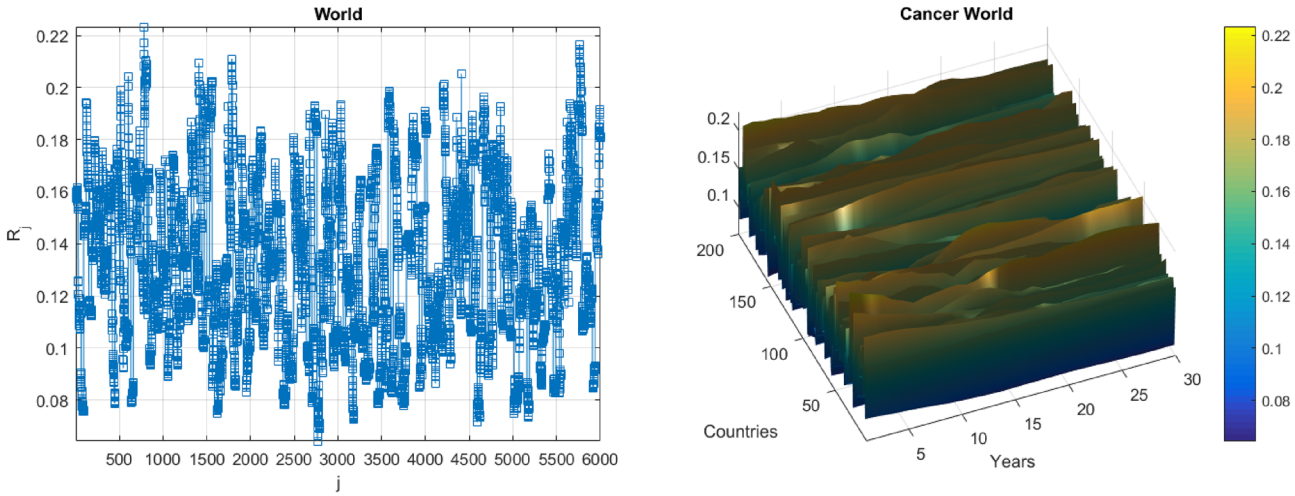

Figure 3. Annual cancer annual death cases. Left: as % of local population per country and year. Right: in per cent as 195D vector $\overrightarrow{R}$. Scaled by Eq. (9) in per cent of the corresponding country population.

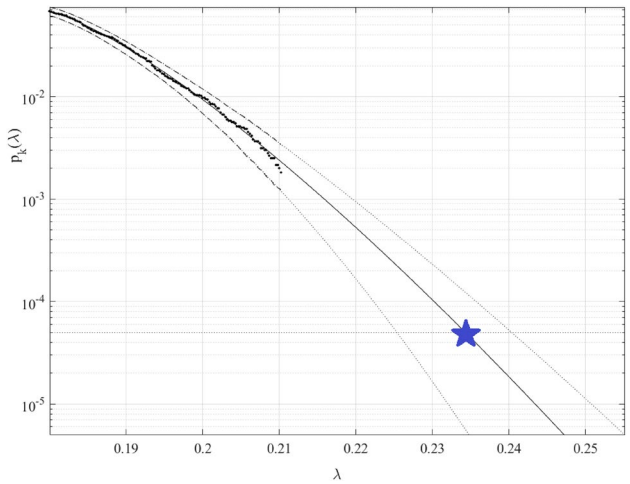

Figure 4. Death rate prediction. 100 years return level extrapolation of $p_k(\lambda)$ towards critical level (indicated by a star) in per cent of the local population. Extrapolated 95% CI indicated by dotted lines. Percentage of the local population on the horizontal axis.

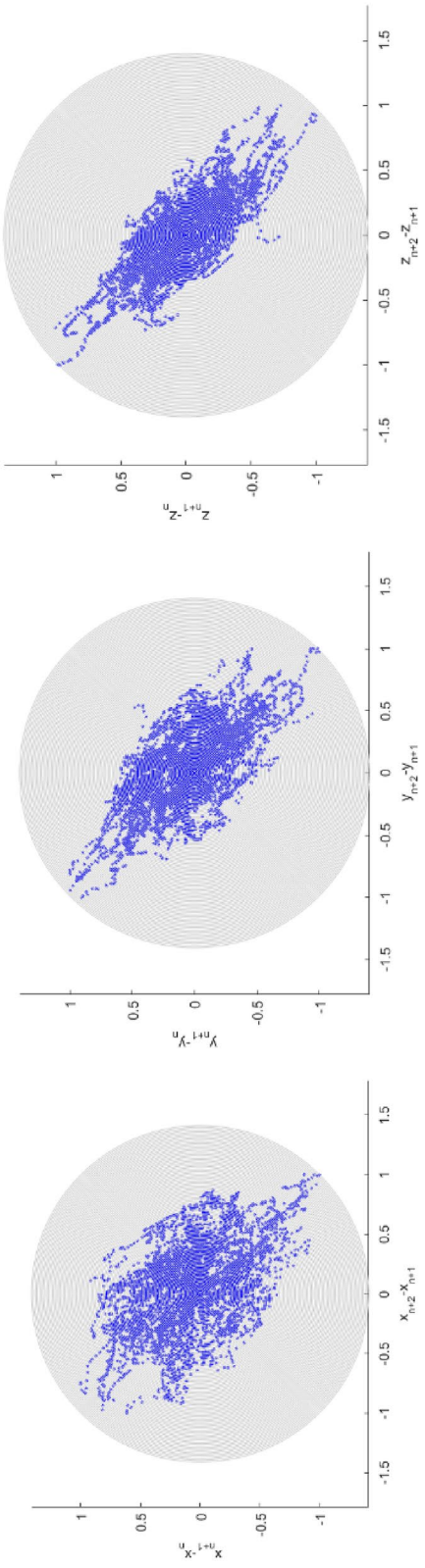

Figure 5. Cancer global statistics. Left: SODP plot. Middle: TODP, Right: FODP.

confidence interval according to Eq. (11). According to Eq. (5) $p(\lambda)$ is directly related to the target failure probability $1 - P$ from Eq. (1). Therefore, in agreement with Eq. (5), system failure probability $1 - P \approx 1 - P_k(1)$ can be estimated. Note that in Eq. (6), N corresponds to the total number of local maxima in the unified response vector $\overrightarrow{R}$. Conditioning parameter $k = 3$ was found to be sufficient due to occurrence of convergence with respect to k, see Eq. (6). Figure 4 exhibits reasonably narrow 95% CI. The latter is an advantage of the proposed method.

The predicted cancer death rate in any world country in any year to come for the next 100 years was found to be about 0.24%.

Note that, although being unique, the above-described technique has the distinct benefit of using existing measured data sets very effectively owing to its capacity to deal with the multidimensionality of the health system and to execute correct extrapolation using relatively small data sets.Note that the predicted non-dimensional λ level, indicated by the star in Fig. 4, represents the probability of cancer outbreak in any world country in the years to come.

In order to validate the suggested methodology, a twice smaller data set was used to obtain predictions for the same probability levels of interest as in Fig. 4. The twice smaller data set was obtained from the original data set by sampling every second consecutive data point. Predicted λ, based on reduced data set, was found within 95% CI based on the entire data set, indicated in Fig. 4.

The second-order difference plot (SODP) originated from the Poincare plot. SODP provides observing the statistical situation of consecutive differences in time series data.

Figure 5 presents SODP along with a third-order difference plot TODP and a fourth-order difference plot FODP. These kinds of plots can be used for data pattern recognition and comparison with other data sets, for example, for the entropy artificial intelligence (AI) recognition approach[32]. Note that EVT is asymptotic and 1DOF, while this study introduces MDOF and sub-asymptotic approaches. To summarise, the predicted non-dimensional λ level, indicated by the star in Fig. 4, represents the probability of world cancer deaths in the years to come. The methodology's limitation lies in its assumption of the underlying bio-environmental process quasi-stationarity.

Discussion

Traditional health systems reliability methods dealing with observed time series do not have the advantage of dealing efficiently with systems possessing high dimensionality and cross-correlation between different system responses. The essential advantage of the introduced methodology is its ability to study the reliability of high dimensional non-linear dynamic systems.

Despite the simplicity, the present study successfully offers a novel multidimensional modelling strategy and a methodological avenue to implement forecasting of the cancer death rate. Proper setting of health system alarm limits (failure limits) per country has been discussed.

This paper studied recorded cancer death rates from all world countries, constituting an example of a one hundred ninety-five dimensional (195D) observed from 1990 to 2019. In real-time, the novel reliability method was applied to cancer annual death rate numbers as a multidimensional system. The theoretical reasoning behind the proposed method is given in detail. Note that the use of direct either measurement or Monte Carlo simulation for dynamic biological system reliability analysis is attractive; however, dynamic system complexity and its high dimensionality require the development of novel robust and accurate techniques that can deal with a limited data set at hand, utilising available data as efficient as possible.

The main conclusion is that the public health system under local environmental and epidemiologic conditions is well managed. This study predicted an annual death rate 100-year return period risk level equal to about 0.24%. Therefore, under current national health management conditions, cancer still represents a future threat to world health.

This study further aimed to develop a general-purpose, robust, and straightforward multidimensional reliability method. The method introduced in this paper has been previously validated by application to a wide range of simulation models, but for only one-dimensional system responses and, in general, very accurate predictions were obtained. Both measured and numerically simulated time series responses can be analysed. It is shown that the proposed method produced a reasonable confidence interval. Thus, the suggested methodology may become appropriate for various non-linear dynamic biological systems reliability studies. Finally, the suggested methodology can be used in many public health applications. The presented cancer example does not limit areas of new method applicability (Supplementary file).

The suggested method can work well with non-stationary data sets (for example, seasonal variations) as soon as they represent the proof of interest. If, however, there is an underlying trend in the process of interest or the data was manipulated, those effects have to be identified. In that case, trend analysis should be performed, a topic for future studies. In any case, authors assume that within 3 years, horizon quasi-stationarity may be assumed. Therefore, the limitation of this study lies within the assumption of bio-system quasi-stationarity, which is, of course, not valid for many years to come.

Data availability

The datasets analysed during the current study are available online[1] https://ourworldindata.org/causes-of-death. The authors confirm that all methods were performed following the relevant guidelines and regulations according to the Declarations of Helsinki.

Code availability

For software used to extrapolate probability tails in this study, see https://github.com/cran/acer.

Received: 22 May 2022; Accepted: 4 January 2023
Published online: 06 January 2023

References

1. Ritchie, H., Spooner, F. & Roser, M. Causes of death. In *Our World in DataOur World in Data*, https://ourworldindata.org/causes-of-death.
2. Siegel, R., Miller, K., Fuchs, H. & Jemal, A. Cancer statistics. *CA Cancer J. Clin.* https://doi.org/10.3322/caac.21708 (2022).
3. Yabroff, K. R. *et al.* Association of the COVID-19 pandemic with patterns of statewide cancer services. *J. Natl. Cancer Inst.* **2021**, 28 (2021).
4. Surveillance, Epidemiology, and End Results (SEER) Program. *SEER*Stat Database: Incidence- SEER 9 Registries Research Data with Delay- Adjustment, Malignant Only, November 2020 Submission (1975- 2018) <Katrina/Rita Population Adjustment>- Linked to County Attributes- Total US, 1969- 2018 Counties. National Cancer Institute, Division of Cancer Control and Population Sciences, Surveillance Research Program, Surveillance Systems Branch* (2021).
5. Surveillance, Epidemiology, and End Results (SEER) Program. *SEER*Stat Database: Incidence- SEER 18 Registries Research Data + Hurricane Katrina Impacted Louisiana Cases, November 2020 Submission (2000- 2018) <Katrina/Rita Population Adjustment>- Linked to County Attributes- Total US, 1969–2018 Counties. National Cancer Institute, Division of Cancer Control and Population Sciences, Surveillance Research Program, Surveillance Systems Branch* (2021).
6. Surveillance Research Program. SEER*Explorer: an interactive website for SEER cancer statistics. National Cancer Institute 2021 (Accessed15 Apr 2021); seer.cancer.gov/explorer/.
7. Surveillance, Epidemiology, and End Results (SEER) Program. *SEER*Stat Database: Incidence- SEER Research Limited- Field Data With Delay- Adjustment, 21 Registries, Malignant Only, November 2020 Submission (2000- 2018)- Linked To County Attributes- Time Dependent (1990- 2018) Income/Rurality, 1969- 2019 Counties. National Cancer Institute, Division of Cancer Control and Population Sciences, Surveillance Research Program* (2021).
8. Surveillance Research Program, Statistic Methodology and Applications. *DevCan: Probability of Developing or Dying of Cancer Software. Version 6.7.9. National Cancer Institute* (2021).
9. Surveillance, Epidemiology, and End Results (SEER) Program. *SEER*Stat Database: North American Association of Central Cancer Registries (NAACCR) Incidence Data- Cancer in North America Analytic File, 1995- 2018, With Race/Ethnicity, Custom File With County, American Cancer Society Facts and Figures Projection Project (which includes data from the Center for Disease Control and Prevention's National Program of Cancer Registries, the Canadian Council of Cancer Registries' Provincial and Territorial Registries, and the National Cancer Institute's SEER Registries, certified by the NAACCR as meeting high- quality incidence data standards for the specified time periods). National Cancer Institute, Division of Cancer Control and Population Sciences, Surveillance Research Program* (2021).
10. Sherman, R., Firth, R. & Charlton, M. *et al. Cancer in North America: 2014- 2018. Volume One: Combined Cancer Incidence for the United States, Canada and North America. North American Association of Central Cancer Registries, Inc* (2021).
11. Sherman, R., Firth, R. & Charlton, M. *et al. Cancer in North America: 2014- 2018. Volume Two: Registry- Specific Cancer Incidence in the United States and Canada. North American Association of Central Cancer Registries, Inc* (2021).
12. Surveillance, Epidemiology, and End Results (SEER) Program. *SEER*Stat Database: Mortality- All Causes of Death, Total US (1969-2019) <Katrina/Rita Population Adjustment>- Linked To County Attributes- Total US, 1969- 2019 Counties (underlying mortality data provided by the National Center for Health Statistics). National Cancer Institute, Division of Cancer Control and Population Sciences, Surveillance Research Program* (2021).
13. Wingo, P. A. *et al.* Long- term trends in cancer mortality in the United States, 1930–1998. *Cancer* **97**(12 suppl), 3133–3275 (2003).
14. Murphy, S. L., Kochanek, K. D., Xu, J. & Heron, M. *Deaths: Final Data for 2012. National Vital Statistics Reports. Vol 63, No. 9. National Center for Health Statistics* (2015).
15. Steliarova-Foucher, E., Stiller, C., Lacour, B. & Kaatsch, P. International classification of childhood cancer. *Cancer* **103**, 1457–1467 (2005).
16. Fritz, A. *et al. International Classification of Diseases for Oncology* (World Health Organization, 2000).
17. World Health Organization (WHO). In *International Statistical Classification of Diseases and Related Health Problems, 10th revision. Vol I- III. WHO* (2011).
18. Surveillance Research Program. In *SEER*Stat software, version 8.3.8. National Cancer Institute* (2020).
19. Surveillance Research Program. In *Joinpoint Regression Program version 4.9.0.1. National Cancer Institute, Statistical Research and Applications Branch* (2021).
20. Mariotto, A. B. *et al.* Geographical, racial and socio- economic variation in life expectancy in the US and their impact on cancer relative survival. *PLoS ONE* **13**, e0201034 (2018).
21. Clegg, L. X., Fever, E. J., Mistune, D. N., Fay, M. P. & Hankey, B. F. Impact of reporting delay and reporting error on cancer incidence rates and trends. *J. Natl. Cancer Inst.* **94**, 1537–1545 (2002).
22. Gumbel, E. *Statistics of Extremes* (Columbia University Press, 1958).
23. Sarkar, S. K. A continuous bivariate exponential distribution. *J. Am. Stat. Assoc.* **82**(398), 667–675 (1987).
24. Gupta, R. D. & Kundu, D. Theory & methods: Generalised exponential distributions. *Aust. N. Z. J. Stat.* **41**(2), 173–188 (1999).
25. Romeo, J. S., Meyer, R. & Gallardo, D. I. Bayesian bivariate survival analysis using the power variance function copula. *Lifetime Data Anal.* **24**, 355–383. https://doi.org/10.1007/s10985-017-9396-1 (2018).
26. Beisel, C. J., Rokyta, D. R., Wichman, H. A. & Joyce, P. Testing the extreme value domain of attraction for distributions of beneficial fitness effects. *Genetics* **176**(4), 2441–2449 (2007).
27. Joyce, P. & Abdo, Z. Determining the distribution of fitness effects using a generalised Beta-Burr distribution. *Theor. Popul. Biol.* **122**, 88–96 (2018).
28. Kristensen, S. B. & Bibby, B. M. A bivariate logistic regression model based on latent variables. *Stat. Med.* **39**(22), 2962–2979 (2020).
29. Naess, A. & Gaidai, O. Estimation of extreme values from sampled time series. *Struct. Saf.* **31**(4), 325–334 (2009).
30. Naess, A. & Moan, T. *Stochastic Dynamics of Marine Structures* (Cambridge University Press, 2013).
31. Numerical Algorithms Group. *NAG Toolbox for Matlab* (World NAG Ltd, 2010).
32. Rice, S. O. Mathematical analysis of random noise. *Bell Syst. Tech. J.* **23**, 282–332 (1944).
33. Xing, Y., Gaidai, O., Ma, Y., Naess, A. & Wang, F. A novel design approach for estimation of extreme responses of a subsea shuttle tanker hovering in ocean current considering aft thruster failure. *Appl. Ocean Res.* **2022**, 123. https://doi.org/10.1016/j.apor.2022.103179 (2022).
34. Gaidai, O. *et al.* Offshore renewable energy site correlated wind-wave statistics. *Probab. Eng. Mech.* **2022**, 68. https://doi.org/10.1016/j.probengmech.2022.103207 (2022).
35. Sun, J. *et al.* Extreme riser experimental loads caused by sea currents in the Gulf of Eilat. *Probab. Eng. Mech.* **2022**, 68. https://doi.org/10.1016/j.probengmech.2022.103243 (2022).
36. Xu, X. *et al.* Bivariate statistics of floating offshore wind turbine dynamic response under operational conditions. *Ocean Eng.* **2022**, 257. https://doi.org/10.1016/j.oceaneng.2022.111657 (2022).
37. Gaidai, O. *et al.* Improving extreme anchor tension prediction of a 10-MW floating semi-submersible type wind turbine, using highly correlated surge motion record. *Front. Mech. Eng.* **2022**, 51. https://doi.org/10.3389/fmech.2022.888497 (2022).

38. Gaidai, O., Xing, Y. & Xu, X. COVID-19 epidemic forecast in USA East coast by novel reliability approach. *Res. Sq.* https://doi.org/10.21203/rs.3.rs-1573862/v1 (2022).
39. Xu, X. *et al.* A novel multidimensional reliability approach for floating wind turbines under power production conditions. *Front. Mar. Sci.* https://doi.org/10.3389/fmars.2022.970081 (2022).
40. Gaidai, O., Xing, Y. & Balakrishna, R. Improving extreme response prediction of a subsea shuttle tanker hovering in ocean current using an alternative highly correlated response signal. *Results Eng.* https://doi.org/10.1016/j.rineng.2022.100593 (2022).
41. Cheng, Y., Gaidai, O., Yurchenko, D., Xu, X., Gao, S. Study on the dynamics of a payload influence in the polar ship. In *The 32nd International Ocean and Polar Engineering Conference, Paper Number: ISOPE-I-22-342* (2022).
42. Gaidai, O. *et al*. On-board trend analysis for cargo vessel hull monitoring systems. In *The 32nd International Ocean and Polar Engineering Conference, Paper Number: ISOPE-I-22-541* (2022).
43. Gaidai, O. *et al.* Bivariate statistics of wind farm support vessel motions while docking. *Ships Offshore Struct.* **16**(2), 135–143 (2020).
44. Gaidai, O., Yan, P., Xing, Y., Xu, J. & Wu, Y. A novel statistical method for long-term coronavirus modelling. *F1000 Res.* **11**, 1282 (2022).
45. Gaidai, O. *et al.* Novel methods for wind speeds prediction across multiple locations. *Sci. Rep.* **12**, 19614. https://doi.org/10.1038/s41598-022-24061-4 (2022).
46. Gaidai, O. & Xing, Y. Novel reliability method validation for offshore structural dynamic response. *Ocean Eng.* **266**, 5. https://doi.org/10.1016/j.oceaneng.2022.113016 (2022).

Acknowledgements

The authors declare no conflicts of interest. No funding was received. All authors contributed equally. Authors declare their research conformity with journal ethical standards.

Author contributions

O.G.—theory,P.Y.—data analysis,Y.X.—corresponding author.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-27547-x.

Correspondence and requests for materials should be addressed to Y.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.