Cancer Epidemiology 90 (2024) 102576

Contents lists available at ScienceDirect

Cancer Epidemiology

journal homepage: www.elsevier.com/locate/canep

An index of cancer survival to measure progress in cancer control: A tutorial

Manuela Quaresma [a,*,1], Francisco Javier Rubio [b], Bernard Rachet [a]

[a] Inequalities in Cancer Outcomes Network (ICON), Department of Health Services Research and Policy, Faculty of Public Health and Policy, London School of Hygiene & Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH, UK
[b] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

ARTICLE INFO

Keywords:
Cancer index
Cancer policy
Monitoring
Net survival

ABSTRACT

Background: Cancer survival is a key component to assess the overall effectiveness of healthcare systems in their cancer management efforts. A key supporting tool for planning and decision making was introduced with the development of an index of cancer survival that summarises survival for all adults and cancer types into one single estimate, but the implementation details have not been previously described.
Methods: We detail the construction of the index, including the structure, the calculation of 'sex-age-cancer' specific weights and our proposed modelling strategy to estimate net survival. We provide some practical recommendations through an illustration using a synthetic dataset ('Replica') that we generated for this purpose. An example of R code usage to estimate the index using our approach is provided.
Results: The 'Replica' contains 500 000 artificial cancer records that mimic a cohort of adult cancer patients diagnosed with cancer in England between 1980 and 2004. Using this dataset, we estimated an index of cancer survival at one, five, and ten years after diagnosis for five selected periods of diagnosis, and provide an example of interpretation of these results.
Discussion: We propose a flexible penalised regression modelling strategy to estimate the index's 'sex-age-cancer' specific cancer survival components that minimises the estimation challenge of these components. This tutorial will support researchers in constructing an index of cancer survival for their own setting, facilitating the enrichment of existing toolkits of cancer indicators to more effectively measure progress against cancer in their respective regions/countries.

1. Introduction

Cancer is a major public health and economic concern worldwide, with its burden expected to spiral upwards for the foreseeable future. [1] Cancer control measures, aimed at reducing the number of new cancers and premature deaths in a population, are based on the implementation of systematic, equitable and evidence-based strategies for prevention, early diagnosis and treatment. [2] Alongside incidence and mortality, cancer survival trends, in particular, provide key insights into the cancer management effectiveness at the population level. [3] In this setting, since the cause of death is not available for the whole cancer population, cancer survival is estimated under the relative survival framework rather than using the cause-specific framework. This implies that the hazard of death due to other causes is instead accounted for using all-cause mortality rates from general population life tables. [4,5]

In 2010, we were commissioned by the then National Cancer Director in the UK Government's Department of Health to build and deliver a cancer survival indicator which would "serve as a measure of the effectiveness of cancer services at both local and national level". [6] We proposed the cancer survival index, a multivariable extension of the conventional univariable standardisation, to complement existing cancer-specific indicators. [6] The index was envisioned to become an instrumental surveillance tool of strategic value for the government's policy, and to serve as a monitoring tool for regional health service managers using a standardised metric and transparent methodology. As such, the index was included into the Delivery Dashboard of England's NHS Assurance Framework that sits at the top of NHS's accountability tree [7,8] and National Statistics from 2010. [9] The index was also used

[*] Corresponding author.
E-mail address: Manuela.Quaresma@lshtm.ac.uk (M. Quaresma).
[1] ORCID: 0000–0002-5290–279X

https://doi.org/10.1016/j.canep.2024.102576
Received 15 November 2023; Received in revised form 10 April 2024; Accepted 14 April 2024
Available online 1 May 2024

at national level to support the 2015–2030 Cancer Research UK's vision set out in their research strategy [10], fed into numerous public funding campaigns, and into online information blogs. [11] Indicators using a similar technique were further developed in other countries. [12,13]

However, with the exception of technical reports [14–17] and a short methodological section in a peer-reviewed paper [18], the principles, steps and challenges for the construction of such an index have never been described. This article details the steps used to construct the index, and provides practical recommendations through an illustration. An example of R code is provided, together with a synthetic dataset and a set of 'sex-age-cancer' specific weights, to enable the user to replicate the construction of the index, and to apply it to their own setting. We capitalise on recent developments in cancer survival modelling [19], which help to overcome some of the difficulties encountered during the estimation phase. Our steps and recommendations can however be extended to other methods and software.

2. Methods

2.1. Basic principles of the index of cancer survival

The term 'index of cancer survival' was chosen for the all-cancers survival indicator to distinguish it from survival estimates for a specific cancer and to minimise the risk of misinterpretation. The approach adopted is based on an expansion to three factors (age, sex and cancer type) of the classical direct age-standardisation technique. Although more factors could have been included, we chose these three factors because survival varies widely with all of them (and these variables are usually known for all cancer patients). This will ensure that the index is not affected by shifts over time in the cancer-specific distributions of age or sex, or changes in the cancer incidence distributions - for example, a reduction in lung cancer incidence or an increase in breast cancer incidence. Changes observed in the index will reflect improvements (or otherwise) in survival either through earlier detection (i.e. a change in the stage at diagnosis distribution), or improvements in treatment (i.e.

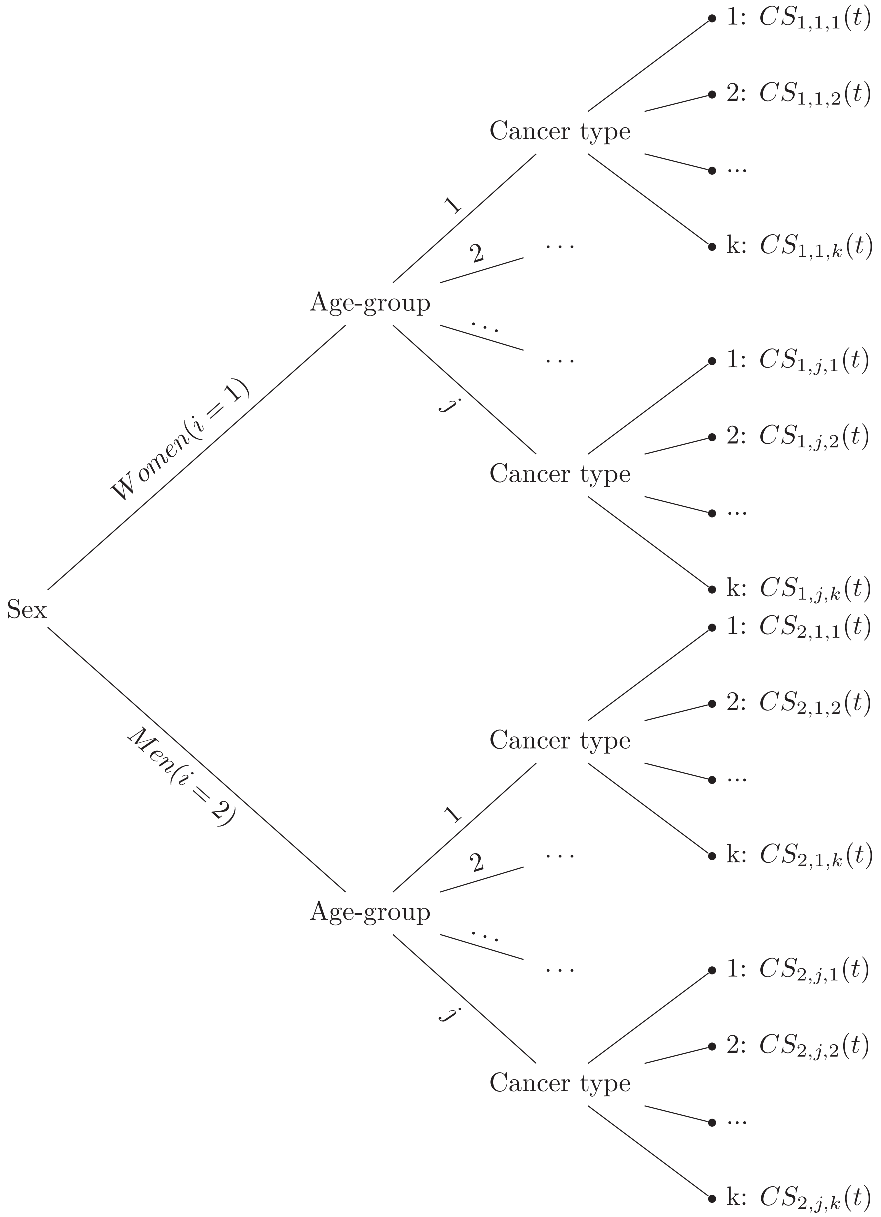

Fig. 1. Generic combinations needed for the estimation of the index of cancer survival using sex i (i=1, 2), age-group j (j=1, 2, …, J) and cancer type k (k=1, 2, …, K).

increases in the proportion of patients receiving treatment with curative intent and/or the use of more efficient treatments).

2.2. Structure of the index of cancer survival

Given a population of cancer patients, we define the index of cancer survival as a weighted average of cancer survival for every pre-specified combination of sex, age group at diagnosis and cancer type (see Fig. 1),

$$\mathrm{ICS}(t) = \sum_{i,j,k} w_{i,j,k} \times \mathrm{CS}_{i,j,k}(t) \quad (1)$$

where, $ICS(t)$ is the index of cancer survival at a given time t after diagnosis, $CS_{i,j,k}(t)$ is the 'sex-age-cancer' specific survival at time t for every combination of sex i (i=1,2), age-group at diagnosis j (j=1,2,…,J) and cancer type k (k=1,2,…,K), and $w_{i,j,k}$ are the 'sex-age-cancer' specific weights. Standard errors [$se(ICS(t))$] can be calculated using Eq. (2), and 95 % confidence intervals can be calculated accordingly using available transformations. [20]

$$se(ICS(t)) = \sqrt{\sum_{i,j,k} w^2_{i,j,k} \times se(CS_{i,j,k})^2} \quad (2)$$

2.3. Estimation of the index of cancer survival

Suitable sets of 'sex-age-cancer' specific weights (if not available) can be created by calculating the proportion of patients in the same predefined combinations as the ones used to estimate the 'sex-age-cancer' specific survival. We recommend using a cancer patient population different from the cancer patient population for which the index is being estimated. This implies that the choice of weights can be seen as arbitrary, since the main purpose of the three-way standardisation is to obtain an index that can be used to monitor changes over time and/or to make comparisons between sub-groups of the population. Once a set of weights is calculated, the same set should be used across all analyses for valid comparisons. To maintain numerical consistency of the estimated index, the sum of weights across all the 'sex-age-cancer' combinations must be one (unity). This implies that an estimate of survival is required for each combination for which the set of weights is defined.

To estimate the 'sex-age-cancer' specific survival components, three choices need to be made regarding: 1) the measure of cancer survival used for the index; 2) the framework under which cancer survival is estimated; and 3) the estimation approach for the survival components. [21] We choose net survival as the measure of cancer survival to estimate an 'Index of Cancer Survival', and we estimate net survival under the 'Relative Survival' framework. [4, 5] Net survival is the most commonly used measure in population-based cancer survival comparisons. It quantifies the survival experienced by patients if cancer was the only possible cause of death. This hypothetical setting is in itself not of interest for an individual cancer patient, but net survival is a useful measure of the effectiveness of a healthcare system in managing cancer treatment and care for the entire population. This is because net survival does not depend on the competing risks of death from other causes, thus allowing for comparisons between populations with different background (all-cause) mortality. In the absence of complete and accurate ascertainment of the cause of death, the relative survival framework enables the estimation of net survival by taking the competing risks of death into account using all-cause mortality rates derived from population life tables. [22] One of the advantages of this approach in comparison with cause-specific approaches is that any excess mortality due to for example yet unknown adverse effects of cancer treatments is accounted for. Within the relative survival framework, several estimation approaches are available, ranging from non-parametric to fully parametric regression modelling approaches (see [19] for a recent review). We chose a modelling approach to estimate the individual net survival components using flexible excess hazard regression models. This approach is better suited for situations of data sparsity compared to non-parametric approaches, when the small number of cases (and events) in some of the defined groups of sex, age and cancer type challenges the estimation of survival leading to unstable estimates (please see Table 1 for some practical estimation tips). When using a modelling approach, net survival of a given group of patients is obtained as the mean of all individual net survival of this group predicted by the model.

2.3.1. Modelling strategy

We use flexible excess hazard regression models, implemented in the R package GJRM [19, 23] to estimate the net survival components for every 'sex-age-cancer' sub-group. These models assume an additive decomposition of the overall hazard function, $h(t|\mathbf{x})$, into two components:

$$h(t|\mathbf{x}) = h_E(t|\mathbf{x}) + h_P(age{+}t) \quad (3)$$

where, $h_E(t|\mathbf{x})$ is the excess hazard function associated with the cancer of interest for an observed event time t and $\mathbf{x}$ a vector of observed covariates. The second component is the hazard function associated with other causes of death, $h_P(age{+}t)$, evaluated at the attained age at death or censoring, age+t, with age the age at diagnosis. This component is typically replaced by the population hazard rate, $h_P(age{+}t|\mathbf{w})$, with $\mathbf{w}$ a subvector of covariates ($\mathbf{w} \subset \mathbf{x}$) obtained from existing population life tables, stratified as finely as possible according to the subset of covariates $\mathbf{w}$. The subset of covariates usually contains less covariates than those available for the cohort of cancer patients, possibly including, in addition to age at death (or censoring), sex and calendar year, socioeconomic status or region of residence.

Based on the decomposition of the hazard function in Eq. (3), we can write the cumulative hazard function as

$$H(t|\mathbf{x}) = H_E(t|\mathbf{x}) + H_P(age{+}t|\mathbf{w}) - H_P(age|\mathbf{w}) \quad (4)$$

The survival function can then be written as:

$$S(t|\mathbf{x}) = \exp\{-H(t|\mathbf{x})\} = \exp\{-H_P(age{+}t|\mathbf{w}) + H_P(age|\mathbf{w})\} \exp\{-H_E(t|\mathbf{x})\} \quad (5)$$

The component, $S_N(t|\mathbf{x}) = \exp\{-H_E(t|\mathbf{x})\}$, is the survival function associated with the excess hazard, and represents the (individual) net survival. We define a link-based net survival model (see [19] for a detailed model specification) for $S_N(t|\mathbf{x})$ as:

$$g\{S_N(t_i|x_i;\beta)\} = \eta_i = \beta_0 + \mathrm{period}_i^T\beta_1 + s_1(\log(t_i)) + s_2(\mathrm{agec}_i) + s_3(\log(t_i), \mathrm{agec}_i) \quad (6)$$

where,

- t_i is the observed event time for individual i, i = 1, …, n and n the population size;
- $S_N(t_i|x_i; \boldsymbol{\beta})$ is the net survival function (conditional on x_i and $\boldsymbol{\beta}$);
- x_i represents a generic vector of patient or tumour characteristics with an associated regression coefficient vector $\boldsymbol{\beta}$;
- g is one of the three allowed link functions (Proportional hazards ('PH'), Proportional Odds ('PO') and Probit ('probit'));
- η_i is the additive predictor;
- period_i represents the period of diagnosis defined on a discrete scale with levels: 1 [1980–1984], 2 [1985–1989], 3 [1990–1994], 4 [1995–1999] and 5 [2000–2004];
- $s_1(.)$ is a monotonic P-spline taken over the logarithm of time;
- age at diagnosis is included as a non-linear and time-dependent effect, with $s_2(.)$ a cubic regression spline taken over the scaled and centered age at diagnosis agec_i, and $s_3(.)$ a tensor product interaction between the scaled and centered age and time, whose marginals are also cubic regression splines.

The variable "period" is included as a fixed effect as we are interested on estimating the specific effect of each period on the net survival. The models were fitted separately for men and women, and for each cancer

Table 1
Practical tips for the estimation of an index of cancer survival.

Topic	Practical tips
Data preparation	• Only first, primary malignant neoplasms are included in line with the types of cancers collected by the majority of population-based cancer registries worldwide (see Discussion for further guidance on this point). • Multiple cancers occurring in different anatomical sites or in the same site are excluded to avoid that two or more cancer records are included for the same patient. • Cancer groupings can be defined according to the International Classification of Diseases [25], the International Classification of Diseases for Oncology [26], or any other relevant classification. • For cancer records where the date of diagnosis is the same as the date of last follow-up (‘true zero survival’), a small-time unit can be added to the follow-up time (for instance, 1 day) to avoid the exclusion of that record from the survival analysis. Cases identified solely by death certification, for which a date of diagnosis could not be retrieved, are excluded. • Each cancer record is matched to a hazard/mortality rate from the general population. These rates are obtained from available population life tables, and the records are merged on calendar year of last follow-up (of death or censoring), age at last follow-up, sex, and any other variables for which life tables are available, as for instance socio-economic status or region of residence (see ‘brate’ in Supplementary Table 1).
‘Sex-age-cancer’ specific weights	• The set of ‘sex-age-cancer’ specific weights does not need to be calculated using the same cancer patient population for which the index of cancer survival is being estimated. • Calculation of the ‘sex-age-cancer’ specific set of weights only needs to be performed once, and the same set of weights should be used throughout the same index analysis to ensure comparability of results, and it can be re-used to estimate different indexes. • Any other suitable number of cancer groupings or age-groups can be chosen to calculate the set of weights instead of those proposed in this article, as long as the sum of the weights remains equal to 1. • It might be of interest to estimate sex-specific, age-specific or cancer-specific survival estimates to present along with the cancer survival index. In that case, those specific estimates can be ‘standardised’ using a calibrated set of the same weights used for the index calculation. For example, sex-specific survival estimates can be standardised by age and cancer. See reference [18] for details.
Estimation of net survival	• When setting-up an excess hazard regression model to predict survival for the required ‘sex- age-cancer’ specific components: – Instead of modelling period of diagnosis as described in the illustration, we can model individual years of diagnosis to estimate an index of cancer survival for each year of diagnosis. – Instead of modelling age-group as a categorical variable, we can model age at diagnosis as a continuous variable, and in the post-estimation step predict net survival for the needed age-groups. • Calculation of the index relies on having an estimate of net survival for each combination of the variables sex, age and cancer. When it is not possible to estimate net survival for each of these combinations, either due to small number of cases (or events) or even zero records in any particular combination, some ad-hoc solutions (in no particular order of preference) include: – If the number of missing ‘sex-age-cancer’ specific net survival combination is less than 10 %, the missing estimate can be replaced by the estimate for the nearest age group for which an estimate was available for a particular cancer-sex combination. – When doing a sub-population analysis, for instance by health-geography, the missing estimate can be replaced by the equivalent ‘sex-age-cancer’ specific estimate for the whole country. – If the number of missing combinations is larger than 10 %, we recommend using broader age groups and fewer cancer groups. This implies that a new set of ‘sex-age-cancer’ specific weights needs to be calculated for the new groupings.

type, and the net survival was estimated by five pre-defined periods (1:[1980–1984], 2:[1985–1989], 3:[1990–1994], 4:[1995–1999] and 5:[2000–2004]), and five pre-defined age-groups (1:[15–44], 2:[45–54], 3:[55–64], 4:[65–74], 5:[75–99]). Including age at diagnosis in every model is crucial to account for the fact that all-cause mortality also varies by age and sex, and thus adjusting for it in the model will allow for the estimation of net survival in the corresponding group. [24]

For each combination of sex (men and women) and cancer type (18 cancer groups as described in ‘Material for illustration’), we fitted three models using the same additive predictor as defined in Eq. (6) but interchanging between three different link functions: Proportional hazards (PH), Proportional Odds (PO) and probit. For each combination, the best fitting model was selected as the one with the smallest Akaike Information Criterion (AIC). After each best fitting model was chosen for each combination of sex and cancer-type, net survival was estimated for each of the five periods of diagnosis and age-groups at one, five and ten years after diagnosis using the post-estimation extraction functions implemented in the GJRM package. These functions are implemented to: 1) predict for each patient their individual net survival function (conditional on their x_i and $\boldsymbol{\beta}$); 2) calculate net survival for sub-groups of the population by averaging the individual net survival functions over all the patients that fall within those sub-groups. For instance, in the model defined in Eq. (6), we model age at diagnosis on a continuous scale but we estimate net survival for each of the five pre-defined age groups, by averaging the predicted individual net survival functions over all the patients whose age falls within a specific age group.

2.4. *Material for illustration and replicating the index estimation*

2.4.1. *The ‘Replica’ dataset*

We generated a synthetic dataset (‘Replica’) containing 500 000 artificial records that mimics the sex, age, cancer patterns of a cohort of adult patients diagnosed in England between 1980 and 2004. A technical summary for the generation of the ‘Replica’ can be found in Supplementary Materials Online. We then used an extract of the ‘Replica’ for the period of diagnosis 2000–2004 to create a set of ‘sex-age-cancer’ specific weights. These were calculated as the proportion of patients in all the combinations of sex, age group and cancer type (Supplementary Table 2).

The ‘Replica’ will enable the user to replicate the estimation of the index, and to use the R code provided on the public repository https://github.com/ManuelaQuaresma/CSI to construct an index using their own data (only rounded values of age at diagnosis and follow-up time available on the data repository). Fig. 2 summarises the key steps for the construction of the index, and Table 1 presents practical estimation tips organised by relevant topics.

3. Results

We emphasise that the results are based on the synthetic dataset ‘Replica’, and do not represent the real distribution of cancer cases, nor the survival trends for patients diagnosed in England. These values are presented for illustrative purposes only.

3.1. *Summary statistics of the ‘Replica’*

Table 2 shows the distribution of cases and deaths, for men and

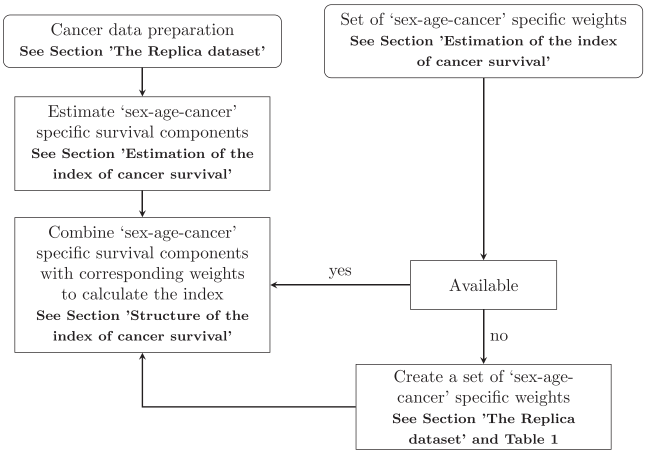
Cancer data preparation
See Section 'The Replica dataset'
Set of 'sex-age-cancer' specific weights
See Section 'Estimation of the index of cancer survival'
Estimate 'sex-age-cancer' specific survival components
See Section 'Estimation of the index of cancer survival'
Combine 'sex-age-cancer' specific survival components with corresponding weights to calculate the index
See Section 'Structure of the index of cancer survival'
yes
Available
no
Create a set of 'sex-age-cancer' specific weights
See Section 'The Replica dataset' and Table 1

Fig. 2. Key steps for the construction of an index of cancer survival.

Table 2
Number of cases (N) and proportion of all-cause deaths (%) within the follow-up period by cancer type for all adult men and women (aged 15–99 years) in the 'Replica' dataset.

	Men		Women	
Cancer type	Cases (N)	Deaths* (%)	Cases (N)	Deaths* (%)
Bladder	18 929	71.0	7352	72.9
Brain	4372	92.3	3188	91.1
Breast (female)	–	–	78 011	51.8
Cervix	–	–	8625	49.6
Colon	20 122	78.8	22 000	77.5
Kidney	6519	76.5	3893	74.7
Leukaemia	6929	82.9	5301	81.6
Lung	54 950	97.3	27 048	97.0
Malignant melanoma	4896	48.6	6947	35.8
Non-Hodgkin lymphoma (NHL)	8685	72.8	7694	71.0
Oesophagus	8092	95.8	5250	96.0
Ovary	–	–	13 192	77.8
Pancreas	6747	98.7	6897	98.8
Prostate	45 119	74.5	–	–
Rectum	15 105	78.5	10 652	76.4
Stomach	14 589	94.7	8490	94.8
Uterus	–	–	11 120	50.3
Other cancers	34 609	70.3	24 677	75.0
Total	249 663	81.8	250 337	69.4

Disclaimer: The results presented in this table are based on the 'Replica' dataset, a cohort of 500 000 artificial cancer records, and they do not represent the real distribution of cancer cases diagnosed in England between 1980 and 2004.

* Deaths from any cause

women, by cancer type. Of the 500 000 records, 249 663 (49.9 %) were men and 250 337 (50.1 %) were women. The number of cases by cancer type ranged between 4372 and 54 950 for men and between 3188 and 78 011 for women. Survival time was defined in years from the date of diagnosis until the date of death or the date of last follow-up. Death (from any cause) was observed for 378 038 (75.6 %) patients over the maximum duration of follow-up of 10.9 years, with a median survival time of 0.81 years among those who died. The mean age at diagnosis was 68.4 years (range=(15.2–99.5), SD=12.8) for men and 66.3 years (range=(15.4–99.7), SD=14.9) for women.

3.2. *Index of cancer survival using the 'Replica'*

We estimated an index of cancer survival at one, five, and ten years after diagnosis for five selected periods of diagnosis: 1980–84, 1985–89, 1990–1994, 1995–1999 and 2000–04. We used the modelling strategy detailed in the 'Modelling strategy' section to estimate the individual 'sex-age-cancer' specific net survival components. Of the 31 sex and cancer specific models fitted, the most chosen model (i.e. model with the smallest AIC) was the model using the Proportional Hazards ('PH') link: 8 models were chosen for cancers in men and 14 models were chosen for cancers in females. All the other chosen models were based on the Proportional Odds ('PO') link for both men and women. Total computing time was 5h48m using the GJRM package (version 0.2–6) in R (version 4.2.2). [23, 27] For time reference, all models were fitted on a 64-bit Operating System Windows server (AMD EPYC 7402 24-Core Processor 2.79 GHz with 1.00 TB of RAM).

The index of cancer survival increased consistently at one, five and ten years after diagnosis between 1980 and 2004 (Fig. 3). The index was estimated at 59.4 % at one-year after diagnosis for patients diagnosed in 1980–1984, reaching 67.6 % in 2000–2004. The index at five years after diagnosis increased from 37.5 % in 1980–1984 to 50.8 % in 2000–2004. Ten-year index reached 46.6 % in 2000–2004 rising from 31.6 % in 1980–1984. Estimates are shown as percentages (0−100) since this is the most common scale cancer survival estimates are presented but they refer to survival probabilities taking values between 0 and 1.

4. Discussion

In this article, we detail the construction of an index of cancer survival, introducing the concept and describing our estimation approach with some practical tips. We illustrate the estimation using a synthetic dataset ('Replica') that we generated to mimic the patterns of cancer survival in England for all adult cancer patients. Paediatric cancers were not included in the generation of the Replica, but the same approach we propose could be used to construct an index of cancer survival only for paediatric cancers, or even an index of cancer survival for cancer patients of all ages.

The index of cancer survival is a single number indicator of cancer survival with a transparent and interpretable construction that conveniently summarises the overall patterns of cancer survival in any one population, in each calendar year (or period), for men and women, and for a wide range of cancers with very different survival. In England, national policymakers have adopted the index as a tool for both national surveillance and local monitoring of cancer services. [7–9] The index is estimated by applying a common sex-, age-, cancer-distribution to the 'sex-age-cancer' specific survival estimates using a set of 'sex-age-cancer' specific weights. This technique is an expansion to three factors (sex, age and cancer type) of the classical direct age-standardisation technique that will ensure that the index is not affected by shifts over time in the cancer-specific distributions of age or

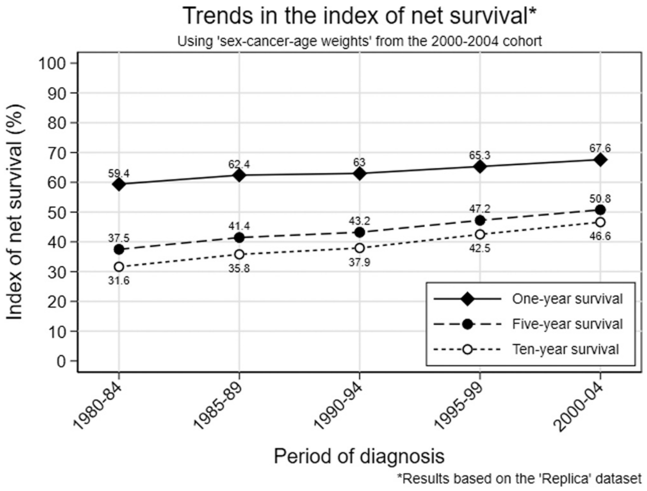

Fig. 3. Trends in the index of net survival for all cancers combined using the 'Replica' dataset.

sex, or changes in the cancer incidence distributions. Similar to direct age-standardisation, the choice of weights is arbitrary given that the main purpose is to monitor changes in the index of cancer survival over time or to make comparisons between populations. Using different sets of weights will result in a shift in the levels of survival but the relationship between estimates being compared remains the same. It is therefore crucial that the same set of weights is used across all analyses for valid comparisons. We note that existing sets of weights for age-standardisation of cancer survival (i.e. weights defined by age groups) are derived from cancer patient populations, differing from the direct age-standardisation of incidence rates, for which sets of weights are derived from general (i.e. not cancer-specific) populations.

Caution is required in the interpretation of the index of cancer survival. [18, 28] We emphasise that the index does not reflect the prospects of survival for any individual cancer patient (or group of patients) since it is based on the estimation of net survival. For example, an estimate of 50 % for the index does not mean that half of all patients will survive. It provides a measure (adjusted for different distributions of cancer patients by age, sex and cancer-type) that is designed to assess and monitor overall progress in the effectiveness of the health system in treating and caring for cancer patients. Measuring progress is a complex multi-factorial challenge, and this strategic surveillance tool is not intended to be used in isolation but to complement existing cancer-specific indicators. [29] It should be seen as a guide to raise questions about the potential for improvement. However, the index can potentially be affected by several factors, as for example, by an increase in the proportion of some early-stage cancers, that in turn can modify the interpretation of time trends. This is the case with prostate cancer where the widespread use of prostate-specific antigen (PSA) testing resulted in the diagnosis of less advanced tumours with a shift of the stage distribution to less advanced, less aggressive and thus less lethal disease. For this reason, the index has been often presented with and without prostate cancer. Changes in pathological disease definitions and cancer registry coding can also affect the interpretation of trends in cancer incidence and survival, such as the shift observed for bladder cancer since the mid-1990 s towards a more restrictive definition of invasive, malignant bladder tumours. For such changes to affect the survival index, a substantial proportion of all cancers needs to be affected, for which prognosis is also very different from that for other cancers. Depending on the setting for which a cancer survival index is being constructed, the impact of such changes in disease definition can be explored by presenting the index with and without the affected cancers, as well as closely examining time series of incidence and survival by stage at diagnosis to aid in the interpretation of the survival index.

In our illustration, we have constructed the survival index only including the first, primary malignant neoplasm of each patient, and thus excluding subsequent cancers. This excludes records that (by definition) will have a shorter survival time for those patients with subsequent cancers, and which are more likely to occur during the most recent periods of diagnosis. This exclusion can potentially introduce a bias when comparing the survival index over time, if the proportion of subsequent cancers is not negligible. For instance, this could be an issue with the increased diagnosis of prostate cancers as subsequent cancers, and requires some thought when defining a meaningful population to construct a cancer survival index, i.e. deciding if only to include first, primary cancers or include the most recent cancer record of each patient. Some sensitivity analysis could be performed to examine the impact on the survival index of only including the most recent cancers.

Although different approaches are available to estimate net survival, we propose a flexible modelling strategy that uses a stable penalised likelihood-based algorithm. In addition, post-estimation is simplified by easy extraction of net survival via user-written functions implemented in the GJRM R package. The same strategy can be used to estimate an index for different sub-populations, and at any level of geographical aggregation, as for instance relevant health geographies. The adopted modelling approach minimises the challenge of estimating net survival for each of the required 'sex-age-cancer' specific combinations. In the illustration, we estimated the indexes' cancer-specific survival components using the 17 most common cancer groups, and we have merged all other cancer types into one single group, which we called "other cancers". When constructing an index, a fine balance needs to be found between the number of cancer groupings that are going to be chosen for the estimation and the stability of those estimates, since these can be affected by the small number of cases (an events) in each group. This choice will depend on the incidence of each cancer type in the population for which the index is being produced, and the stability of the estimates should be checked carefully. Inevitably, this implies that rarer cancers will often be included in the "other cancers" group, which can potentially present heterogeneous levels of survival for the cancers included in that group. However, we emphasise that the index of cancer survival was envisioned as a simple tool to monitor overall progress in cancer outcomes for all cancers combined, and it should not replace other existing cancer-specific indicators (such as outcome indicators for rare cancers), but be used in conjunction with those indicators to draw a complete picture of cancer progress in a population.

The cancer survival index provides a simple tool to measure progress both at national and local area levels. Initially, the index can be constructed for long time series (depending on available data), and subsequently be updated on a yearly basis, adding on to the initial time series. In our illustration, we estimated the index for a long time series spanning over two decades of diagnosis, from 1980 to 2004. Every cancer patient that we have simulated in the synthetic dataset 'Replica' has a full potential follow-up of 10.9 years. We have decided to present the estimates of the survival index by five periods of diagnosis and estimate the index at one-, five-, and ten-years since diagnosis using a cohort approach (Fig. 3). Other estimation options, could include estimating the survival index for individual years of diagnosis, and use other survival estimation approaches, such as a complete, period or hybrid approach to predict long-term survival using the most recent available data, and with a similar call for caution of differences in interpretation of such survival estimates. A modelling approach, similar to the one we present here, could also be explored to predict long-term survival.

In summary, we provided some practical recommendations for researchers to implement an index of cancer survival for their own setting, facilitating the enrichment of existing toolkits of cancer indicators to effectively measure progress against cancer in their respective regions/countries. The successful use of such an index by policymakers requires careful consideration regarding the presentation of results in a simple and effective way for a vast range of audiences of diverse backgrounds.

Research ethics approval

Ethical approval was obtained from the NHS Health Research Authority - Research Ethics Committee (REC reference: 21/LO/0552).

Author contributions

MQ and BR conceptualised and developed the Index of Cancer Survival. FJR and MQ wrote R code for the estimation of the index and MQ performed the analysis. All authors reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by Cancer Research UK Population Research Committee Funding Scheme: Cancer Research UK Population Research Committee-Programme Award C7923/A29018.

CRediT authorship contribution statement

Bernard Rachet: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. **Francisco Javier Rubio:** Software, Writing – review & editing. **Manuela Quaresma:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare no conflict of interest.

Data Availability

The synthetic cancer data set (the Replica) used in this study (with rounded values of age at diagnosis and follow-up time) is available on the public GitHub repository https://github.com/ManuelaQuaresma/CSI, together with a data set containing the set of 'sex-age-cancer' specific weights, and software tools (R code) that allow for reproducibility of our results.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.canep.2024.102576.

References

[1] Global Cancer Observatory: Cancer today [Internet]. International Agency for Research on Cancer. Available from: https://gco.iarc.fr/today.

[2] World Health Organization. National cancer control programmes: policies and managerial guidelines, 2nd ed, World Health Organization, Geneva, 2002.

[3] L. Ellis, L.M. Woods, J. Esteve, S. Eloranta, M.P. Coleman, B. Rachet, Cancer incidence, survival and mortality: explaining the concepts, Int. J. Cancer 135 (8) (2014) 1774–1782.

[4] J. Esteve, E. Benhamou, M. Croasdale, L. Raymond, Relative survival and the estimation of net survival: elements for further discussion, Stat. Med. 9 (5) (1990) 529–538.

[5] M.P. Perme, J. Stare, J. Esteve, On estimation in relative survival, Biometrics 68 (1) (2012) 113–120.

[6] M. Quaresma, S. Walters, E. Gordon, C. Carrigan, M.P. Coleman, B. Rachet. A cancer survival index for primary care trusts, Office for National Statistics, Newport, UK, 2010.

[7] All-Party Parliamentary Group on Cancer. One year cancer survival rates: measuring progress. 2015. [26/07/2023] Available From: ⟨https://www.macmillan.org.uk/dfsmedia/1a6f23537f7f4519bb0cf14c45b2a629/8976-10061/appgc-measuring-progress/⟩.

[8] John Baron M.P. NHS transparency on cancer survival rates will be transformational. 2014. [26/07/2023]. Available From: ⟨https://conservativehome.com/2014/12/16/john-baron-mp-nhs-transparency-on-cancer-survival-rates-will-be-transformational/⟩.

[9] NHS Digital. Cancer survival: index for sub-integrated care boards, 2005 to 2020. 2023. [26/07/2023] Available From: ⟨https://digital.nhs.uk/data-and-information/publications/statistical/cancer-survival-in-england/index-for-sub-integrated-care-boards-2005-to-2020/⟩.

[10] Cancer Research UK. Beating cancer sooner. Our research strategy. 2014. [26/07/2023]. Available From: ⟨https://www.cancerresearchuk.org/sites/default/files/cruk_research_strategy.pdf/⟩.

[11] Jones G, Why are cancer rates increasing? 2015. [26/07/2023]. Available From: ⟨https://news.cancerresearchuk.org/2015/02/04/why-are-cancer-rates-increasing/⟩.

[12] C.J. Johnson, H.K. Weir, A. Mariotto, R. Wilson, D. Nishri, Construction of a North American cancer survival index to measure progress of cancer control efforts, Prev. Chronic Dis. 14 (2017). E81.

[13] L.F. Ellison, The cancer survival index: measuring progress in cancer survival to help evaluate cancer control efforts in Canada, Health Rep. 32 (9) (2021) 14–26.

[14] M. Quaresma, S. Whitehead, N. Bannister, M.P. Coleman, B. Rachet. Index of cancer survival for clinical commissioning groups in England; patients diagnosed 1996-2011 and followed up to 2012, Office for National Statistics, Newport, UK, 2013.

[15] M. Quaresma, R. Drummond, S. Rowlands, P. Brown, N. Bannister, M.P. Coleman, et al.. Index of cancer survival for clinical commissioning groups in England: adults diagnosed 1997-2012 and followed up to 2013, Office for National Statistics, Newport, UK, 2014.

[16] M. Quaresma, J. Jenkins, N. Bannister, R. Murphy, J. Kaur, M. Peet, et al.. Index of cancer survival for clinical commissioning groups in England: adults diagnosed 1999-2014 and followed up to 2015, Office for National Statistics, Newport, UK, 2016.

[17] M. Quaresma, E. Nash, N. Bannister, M.P. Coleman, B. Rachet. Index of cancer survival for clinical commissioning groups in England: adults diagnosed 1998-2013 and followed up to 2014, Office for National Statistics, Newport, UK, 2016.

[18] M. Quaresma, M.P. Coleman, B. Rachet, 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: a population-based study, Lancet 385 (9974) (2015) 1206–1218.

[19] A. Eletti, G. Marra, M. Quaresma, R. Radice, F.J. Rubio, A unifying framework for flexible excess hazard modelling with applications in cancer epidemiology, J. R. Stat. Soc. Ser. C 71 (2022) 1044–1062.

[20] M. Quaresma, M.P. Coleman, B. Rachet, Funnel plots for population-based cancer survival: principles, methods and applications, Stat. Med. 33 (6) (2014) 1070–1080.

[21] M. Pohar Perme, L.C. de Wreede, D. Manevski, What is relative survival and what is its role in haematology? Best. Pract. Res. Clin. Haematol. 36 (2) (2023) 101474.

[22] Office for National Statistics, (2021). National life tables - Life expectancy in the UK: 2018 to 2020.

[23] Marra G., Radice R. R package GJRM: Generalised Joint Regression Modelling. 0.2-6.4 ed2023.

[24] C. Danieli, L. Remontet, N. Bossard, L. Roche, A. Belot, Estimating net survival: the importance of allowing for informative censoring, Stat. Med. 31 (8) (2012) 775–786.

[25] World Health Organization, International Statistical Classification of Diseases and Related Health Problems [electronic resource]. 10th rev., edition 2008 ed, World Health Organization, Geneva, 2009.

[26] World Health Organization. International Classification of Diseases for Oncology (ICD-O), 3rd ed., World Health Organization, Geneva, 2013, 1st revision ed.

[27] R Core Team, R: A language and environment for statistical computing, R foundation for statistical computing, Vienna, Austria, 2021.

[28] M.J. Rutherford, Care needed in interpretation of cancer survival measures, Lancet 385 (9974) (2015) 1162–1163.

[29] A. Belot, A. Ndiaye, M.A. Luque-Fernandez, D.K. Kipourou, C. Maringe, F.J. Rubio, et al., Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data, Clin. Epidemiol. 11 (2019) 53–65.