Published in final edited form as:

Stat Med. 2019 July 30; 38(17): 3184–3203. doi:10.1002/sim.8170.

Partitioning of Time Trends in Prevalence and Mortality of Lung Cancer

Igor Akushevich¹, Julia Kravchenko², Arseniy P. Yashkin¹, Fang Fang³, Anatoliy I. Yashin¹

¹Biodemography of Aging Research Unit, Center for Population Health and Aging, Duke University, Durham, NC, United States.

²Department of Surgery, Duke University School of Medicine, Durham, NC, United States.

³Center for Genomics in Public Health and Medicine, RTI International, Research Triangle Park, NC, United States.

Abstract

Background: Time trends of lung cancer prevalence and mortality are the result of three competing processes: changes in the incidence rate, stage-specific survival, and ascertainment at early stages. Improvements in these measures act concordantly to improve disease-related mortality, but push the prevalence rate in opposite directions making a qualitative interpretation difficult. The goal of this paper is to evaluate the relative contributions of these components to changes in lung cancer prevalence and mortality.

Methods: Partitioning of prevalence and mortality trends into their components using SEER data for 1973–2013.

Results: The prevalence of lung cancer increases for females and decreases for males. In 1998 the former was due to increased incidence (45–50% of total trend), improved survival (40–45%) and increased ascertainment at early stages (10–15%). In males a rapidly declining incidence rate overpowered the effects of survival and ascertainment resulting in an overall decrease in prevalence over time. Trends in lung cancer mortality are determined by incidence during 1993–2002 with noticeable contribution of survival after 2002.

Conclusion: Lung cancer incidence was the main driving force behind trends in prevalence and mortality. Improved survival played essential role from 2000 onwards. Trends in stage ascertainment played a small but adverse role. Our results suggest that that further improvement in lung cancer mortality can be achieved through advances in early stage ascertainment, especially for males, and that in spite of success in treatment, adenocarcinoma continues to exhibit adverse trends (especially in female incidence) and its role among other histology-specific lung cancers will increase in the near future.

1. Introduction

Lung cancer remains one of the most frequent cancer types worldwide accounting for 12.9% of total incident cases and 19.4% of all cancer-related deaths¹. In the U.S., lung cancer has replaced breast cancer as the leading cause of cancer-related death in females 60+ and is the leading cause of cancer-related death in males aged 40+². In general, trends of lung cancer incidence reflect the decrease in smoking prevalence in the U.S., with males being the primary beneficiaries². However, the 5-year relative survival is still low at about 18% and only 4% if lung cancer is diagnosed at distant stage² (over 50% of all diagnoses).

Time trends in prevalence and mortality are the result of three competing processes: changes in incidence rate, stage-specific survival, and cancer ascertainment at early stages. Improvements in these measures due to health interventions or other reasons act concordantly to improve disease-related mortality. However, these same improvements will push the prevalence in opposite directions making a qualitative evaluation of the prevalence trend difficult. In this paper, we perform a partitioning of the historical time trends in lung cancer prevalence and mortality to evaluate how changes in incidence, survival and ascertainment over the 1973–2013 period affected these trends in males and females and identify the effects of progress in lung cancer diagnosis and treatment.

The partitioning approach used in this paper is a further development of the one used by Akushevich et al.³ to analyze prevalence trends of diabetes mellitus using Medicare data. We expand on this method to account for the different structure of cancer data in the Surveillance, Epidemiology, and End Results (SEER) dataset and for the effect of changes in disease ascertainment which is largely determined by stage at diagnosis that is not available in Medicare data. Furthermore, the wide range of age groups present in SEER (Medicare is effectively limited to 65+) allows for the decomposition to be applied directly for all ages or, to improve the prediction of initial prevalence at 65 necessary for the base method³. The difference in data structure requires different approaches for parameter estimation. The resulting method noticeably improves on the one developed by Akushevich and colleagues, overcoming many of its limitations and improving the accuracy of the estimates achievable with Medicare data alone.

2. Theory of partitioning of cancer trends in prevalence and mortality

Data collected in cancer registries represent information on individuals resident in a certain geographic area with cancer diagnosed at a certain year. For each individual a spectrum of tumor characteristics (stage, behavior, histology, grade, etc.) is available as well as time till death or censoring (follow-up cutoff date). In addition, population counts in one-year groups for the calculation of age- and time-specific incidence rates are usually available. In this study, we use data on lung cancer collected in a national SEER dataset (1973–2013)⁴. Three types of empiric estimates are used for comparison and cross checks: age-adjusted prevalence of lung cancer produced by SEER*Stat software⁵, recent prevalence estimates using Medicare data⁶, and estimates of mortality rates using the Multiple Cause-of-Death (MCD) dataset (1968–2013)⁷.

Figure 1 presents the Lexis diagram in the plane over age (in years; denoted by x) and calendar time (in years; denoted by y). Each of the dashed lines in the Lexis diagram uniquely corresponds to a birth cohort with the birth time $y_b = y - x$ for any point (x, y) belonging to the cohort-specific dashed line. Therefore, epidemiologic characteristics at a given point of time are defined by the history of the cohort represented by a leftward move along the respective line in the Lexis diagram down to bounds of the available region. In this analysis we partition age-adjusted lung cancer prevalence and mortality for ages 65 and above over the 1993–2010 time period. This area is shown in Figure 1 as a rectangle. Information for the cohorts beyond the left-bottom bound is used to produce accurate initial values for all epidemiologic measures using in the analysis.

Below we derive the formulas for partitioning of age-adjusted prevalence (subsection 2.1) and mortality (subsection 2.2). The mathematics involved are most closely related to the methods in deterministic modeling of infectious diseases^{8–10}. Then we discuss the properties of the obtained formulas (subsection 2.3) and develop the estimation procedure for this model (Section 3). The mathematical notation used throughout this article is presented in Table 1.

2.1. Model for Cancer Prevalence

The probability of being prevalent $P_c(x)$ at age x requires being incident at an earlier age τ , $\tau < x$ (represented by incidence rate $I_c(\tau)$) and having survival time larger than $x - \tau$ (represented by survival probability $S(x - \tau, \tau)$ of patient diagnosed at age τ). Thus

$$P_c(x) = \int_{\bar{x}_0}^x I_c(\tau) S(x - \tau, \tau) d\tau \quad (1)$$

where we integrate over all possible ages of incidence. This formula can be understood in terms of the numbers of individuals. If N_0 is the size of a birth cohort and $N_d(\tau)$ is the number of individuals with disease onset at age period $[\tau, \tau + \Delta\tau]$, the total number of sick (and alive) individuals at age x ($N_d(x)$) is the sum of all individuals who survived to age x after diagnoses at τ over all age periods, i.e., $N_d(x) = \sum_n N_d(\tau_n) S(x - \tau_n, \tau_n)$, where $S(x - \tau_n, \tau_n)$ is the survival function of individuals diagnosed at age period $[\tau_n, \tau_n + \Delta\tau_n]$ who survived to age x . The formula (1) is obtained when we consider infinitely small age periods (i.e., $\Delta\tau \rightarrow 0$) and define $P_c(x) = N_d(x)/N_0$ and $I_c(\tau) = N_d(\tau)/N_0$.

The eqn. (1) is written in terms of the cohort functions that are defined along the dashed lines in Fig 1. The denominator N_0 is the number of individuals at the time of cohort forming. The initial age is denoted by $\bar{x}_0$. It can be age of birth ($\bar{x}_0 = 0$) or age when prevalence of lung cancer is still negligible. The second option allows us to avoid modeling of patterns of child mortality which is not related to the effects of a considered disease. We use $\bar{x}_0 = 30$ years in this paper.

Cancer can be detected at different stages (e.g., local, regional, distant). If data on stages at diagnoses are available the formula for cohort prevalence is upgraded as:

$$P_c(x) = \int_{\bar{x}_0}^x I_c(\tau) \left(\sum_s \pi_s(\tau) S_s(x - \tau, \tau) \right) d\tau$$

where $\pi_s(\tau)$ are the probability to detect stage at diagnosis, s , and $S_s(x - \tau, \tau)$ is the stage-specific survival probability.

Incidence, stage at diagnosis probability, and survival after disease onset can be cohort dependent. Therefore we use

$$P_c(x, y_b) = \int_{\bar{x}_0}^x I_c(\tau, y_b) \left(\sum_s \pi_s(\tau, y_d) S_s(x - \tau, \tau, y_d) \right) d\tau \quad (2)$$

where y_b is the birth year of a cohort and $y_d = y_b + \tau$ is the year of diagnosis and therefore the year when a cohort of patient is formed. The exact definition of $P_c(x, y_b)$ is the fraction of individuals living with lung cancer at age x and born in year y to the total number of individuals born in year y_b . Similarly, $I_c(\tau, y_b)$ is the cohort incidence defined through the number of new cases per cohort size (i.e., the number of individuals born in year y_b). However, the cohort size for the studied population is not usually known with sufficient accuracy. What is known (or can be estimated) is the current population at risk, i.e., the population currently living in the same age and year. Therefore we avoid dealing with the cohort prevalence and incidence and use their definitions involving the population at risk rather than birth cohort size. The eqn. (2) within these definitions is rewritten as,

$$P(x, y_b) = S_t^{-1}(x, y_b) \int_{\bar{x}_0}^x I(\tau, y_b) S_t(\tau, y_b) \left(\sum_s \pi_s(\tau, y_d) S_s(x - \tau, \tau, y_d) \right) d\tau \quad (3)$$

What is new here are the survival functions $S_s(x, y_b)$ in the general population. Since $S_t^{-1}(x, y_b) S_t(\tau, y_b)$ is the reciprocal of survival probability at x for the cohort of patients formed at τ , the factor containing all three survival functions in (3) is the relative survival for patients diagnosed at τ :

$$S_s^r(x - \tau, \tau, y_d) = \frac{S_t(\tau, y_b)}{S_t(x, y_b)} S_s(x - \tau, \tau, y_d) \quad (4)$$

In addition to the cohort effect, we are interested in prevalence as a function of the calendar year ($y = y_b + x$). The cohort prevalence and incidence are expressed through accepted definitions of prevalence and incidence: $P_c(x, y_b) = P(x, y) S_s(x, y_b)$ and $I_c(x, y_b) = I(x, y_d) S_s(x, y_b)$, where $S_s(x, y_b)$ is the survival function of the cohort born at year y_b . For prevalence dependent on x and y we obtain,

$$P(x, y) = \int_{\bar{x}_0}^x I(\tau, y_d) \left(\sum_s \pi_s(\tau, y_d) S_s^r(x - \tau, \tau, y_d) \right) d\tau,$$

where S_s^r are the relative stage-specific survivals (ratios of survival probabilities for patients to those for the general population), e.g., $S^r(x - \tau, \tau, y_d)$ is the relative survival at $x - \tau$ for patients diagnosed at $\bar{x}_0$ and diagnosed at age τ and year y_d .

The quantity of interest is the time trend of age adjusted prevalence (over the age region $(x_0, x_{\max})$ or (x_0, ∞)) and its partitioning. Age-adjusted prevalence is:

$$P(y) = \int_{x_0}^{\infty} \left(\int_{\bar{x}_0}^x I(\tau, y_d) (\sum_s \pi_s(\tau, y_d) S_s^r(x - \tau, \tau, y_d)) d\tau \right) p(x) dx$$

where $p(x)$ is the density of age distribution in a standard year. We see that the time trend in $P(y)$ can be due to changes in the incidence rate, stage-specific survival, and ascertainment at early stages.

Explicit calculation of the derivative with respect to y results in a time trend formally defined as $P'(y)/P(y)$

$$\begin{aligned} \frac{P'(y)}{P(y)} &= \frac{1}{P(y)} \int_{x_0}^{\infty} \int_0^{x - \bar{x}_0} \sum_s (I'(\tau, y_d) \pi_s(\tau, y_d) S_s^r(x_\tau, \tau, y_d) + I(\tau, y_d) \pi'_s(\tau, y_d) S_s^r(x_\tau, \tau, y_d) \\ &+ I(\tau, y_d) \pi_s(\tau, y_d) S_s^{r'}(x_\tau, \tau, y_d)) p(x) dx_\tau dx \end{aligned}$$

where we changed the internal integration variable $\tau \rightarrow x_\tau \equiv x - \tau$. Thus, the time trend is determined through three terms representing the partitioning of time trends for prevalence rates:

$$P'(y)/P(y) = T_{inc}(y) + T_{\pi}(y) + T_S(y).$$

Explicitly the expressions for $T_{\cdot}(y)$ are:

$$\begin{aligned}
T_{inc}(y) &= \frac{1}{P(y)} \int_{x_0}^{\infty} \left(\int_0^{x-\bar{x}_0} I'(x-x_\tau, y_d) \left(\sum_s \pi_s(x-x_\tau, y_d) S_d^r(x_\tau, x-x_\tau, y_d) \right) dx_\tau \right. \\
&\quad \left. \right) p(x) dx \\
T_\pi(y) &= \frac{1}{P(y)} \int_{x_0}^{\infty} \left(\int_0^{x-\bar{x}_0} I(x-x_\tau, y_d) \left(\sum_s \pi'_s(x-x_\tau, y_d) S_d^r(x_\tau, x-x_\tau, y_d) \right) dx_\tau \right) p \\
&\quad (x) dx \\
T_S(y) &= \frac{1}{P(y)} \int_{x_0}^{\infty} \left(\int_0^{x-\bar{x}_0} I(x-x_\tau, y_d) \left(\sum_s \pi_s(x-x_\tau, y_d) S_d^{r'}(x_\tau, x-x_\tau, y_d) \right) dx_\tau \right) p \\
&\quad (x) dx
\end{aligned} \tag{5}$$

2.2. Model for Cancer Mortality

Similarly, for mortality (we consider incidence-based mortality, i.e., mortality after disease onset¹¹) the probability of dying in the age interval $(x, x + dx)$ requires having death in the interval $(x, x + dx)$ and being incident at an earlier age τ . The same consideration as for eqn. (3) results in:

$$M(x, y_b) = S_t^{-1}(x, y_b) \int_{\bar{x}_0}^x I(\tau, y_b) S_t(\tau, y_b) \left(\sum_s \pi_s(\tau, y_d) f_s(x-\tau, \tau, y_d) \right) d\tau$$

where $d_s(x-\tau, \tau, y_b)$ is the density of death events. Since $S_t^{-1}(x, y_b) S_t(\tau, y_b) = S_t^{-1}(x-\tau, \tau, y_b)$, combining survival probabilities for mortality results in the ratio of density f_c to the survival probability in the general population, that can be further transformed as

$$\begin{aligned}
\frac{f(x-\tau, \tau, y_d)}{S_t(x-\tau, \tau, y_d)} &= - \frac{S'_t(x-\tau, \tau, y_d)}{S_t(x-\tau, \tau, y_d)} = - \frac{(S_t(x-\tau, \tau, y_d) S_t^r(x-\tau, \tau, y_d))'}{S_t(x-\tau, \tau, y_d)} = \\
&= f^r(x-\tau, \tau, y_d) + S_t^r(x-\tau, \tau, y_d) \mu(x, y)
\end{aligned}$$

The function $\mu(x, y)$ is the mortality hazard function in the general population. Note, since $S_t(x-\tau, \tau, y_d) = \exp(-\int_\tau^x \mu(v, y_d + v - \tau) dv)$, then $S'_t(x-\tau, \tau, y_d) = \mu(x, y_d + x - \tau) S_t(x-\tau, \tau, y_d)$. We see that all terms at $\mu(x, y)$ give the prevalence. Finally we obtain the incidence-based mortality in terms of $\mu(x, y)$ and the functions used for partitioning of prevalence trends. Then the age-specific mortality rate is

$$M(x, y) = \mu(x, y)P(x, y) + \int_{\bar{x}_0}^x I(\tau, y_d) \left(\sum_s \pi_s(\tau, y_d) f_d^r(x - \tau, \tau, y_d) \right) d\tau$$

The age-adjusted rate is

$$M(y) = \int_{x_0}^{\infty} \left(\mu(x, y)P(x, y) + \int_0^{x - \bar{x}_0} I(\tau, y_d) \left(\sum_s \pi_s(\tau, y_d) f_d^r(x_\tau, \tau, y_d) \right) dx_\tau \right) p(x) dx$$

The derivative that represents the partitioning of the incidence-based mortality have 5 terms:

$$M'(y)/M(y) = \hat{T}_P(y) + \hat{T}_\mu(y) + \hat{T}_{inc}(y) + \hat{T}_\pi(y) + \hat{T}_S(y) \quad (6)$$

The term $\hat{T}_P(y)$ reflects the contribution to mortality from the change in prevalence of cancer patients. Explicitly it is:

$$\begin{aligned} \hat{T}_P(y) = \frac{1}{M(y)} \int_{x_0}^{\infty} \int_0^{x - \bar{x}_0} \sum_s (I'(\tau, y_d) \pi_s(\tau, y_d) S_s^r(x_\tau, \tau, y_d) \\ + I(\tau, y_d) \pi_s'(\tau, y_d) S_s^r(x_\tau, \tau, y_d) + I(\tau, y_d) \pi_s(\tau, y_d) S_s^{r'}(x_\tau, \tau, y_d)) \mu(x, y) p(x) dx_\tau dx \end{aligned} \quad (7)$$

Respectively, $\hat{T}_\mu(y)$ is:

$$\hat{T}_\mu(y) = \frac{1}{M(y)} \int_{x_0}^{\infty} \left(\int_0^{x - \bar{x}_0} I(\tau, y_d) \sum_s (\pi_s(\tau, y_d) S_s^r(x_\tau, \tau, y_d)) dx_\tau \right) \mu'(x, y) p(x) dx \quad (8)$$

The remaining expressions for $\hat{T}_\pi(y)$ are similar to the respective terms in the expressions for prevalence partitioning:

$$\begin{aligned}
\hat{T}_{inc}(y) &= \frac{1}{M(y)} \int_{x_0}^{\infty} \left(\int_0^{x-\bar{x}_0} I(x-x_\tau, y_d) \left(\sum_s \pi_s(x-x_\tau, y_d) f_d^r(x_\tau, x-x_\tau, y_d) \right) dx_\tau \right) p(x) dx \\
\hat{T}_\pi(y) &= \frac{1}{M(y)} \int_{x_0}^{\infty} \left(\int_0^{x-\bar{x}_0} I(x-x_\tau, y_d) \left(\sum_s \pi'_s(x-x_\tau, y_d) f_d^r(x_\tau, x-x_\tau, y_d) \right) dx_\tau \right) p(x) dx \\
\hat{T}_S(y) &= \frac{1}{M(y)} \int_{x_0}^{\infty} \left(\int_0^{x-\bar{x}_0} I(x-x_\tau, y_d) \left(\sum_s \pi_s(x-x_\tau, y_d) f_d^{r'}(x_\tau, x-x_\tau, y_d) \right) dx_\tau \right) p(x) dx
\end{aligned} \tag{9}$$

2.3. Discussion of the partitioning formulas

The partitioning approach developed in this study is based on an explicit representation of prevalence at ages 65+ with no simplifying assumptions. The resulting formulae for age-specific and age-adjusted prevalence for ages 65 and above are (refer to notation previously introduced in Table 1):

$$\begin{aligned}
P(x, y) &= \int_{\bar{x}_0}^x I(\tau, y_d) \left(\sum_s \pi_s(\tau, y_d) S_d^r(x_\tau, \tau, y_d) \right) d\tau \\
P(y) &= \int_{x_0}^{\infty} \left(\int_0^{x-\bar{x}_0} I(\tau, y_d) \left(\sum_s \pi_s(\tau, y_d) S_d^r(x_\tau, \tau, y_d) \right) dx_\tau \right) p(x) dx
\end{aligned} \tag{10}$$

Similarly, the resulting formulae for age-specific and age-adjusted prevalence for ages 65 and above are:

$$\begin{aligned}
M(x, y) &= \mu(x, y) P(x, y) + \int_{\bar{x}_0}^x I(\tau, y_d) \left(\sum_s \pi_s(\tau, y_d) f_d^r(x-\tau, \tau, y_d) \right) d\tau \\
M(y) &= \int_{x_0}^{\infty} \left(\mu(x, y) P(x, y) + \int_0^{x-\bar{x}_0} I(\tau, y_d) \left(\sum_s \pi_s(\tau, y_d) f_d^r(x_\tau, \tau, y_d) \right) dx_\tau \right) p(x) dx
\end{aligned} \tag{11}$$

The formulas (10) and (11) reflect that the probability of being prevalent at age x requires being incident at an earlier age τ , $\tau < x$ and having survived longer than $x - \tau$. Similarly, for incidence-based mortality the probability of dying in the age interval $(x, x + dx)$ requires having death in the interval $(x, x + dx)$ and either being prevalent at the boundary point for

this cohort or being incident at an earlier age $x - \tau$. Integration over all ages of diagnosis (τ) results in the formula for age-specific prevalence presented in eqn. (10). For mortality these calculations result in i) use of the density f_d^r of the relative survival function and ii) the introduction of a separate term containing disease prevalence and mortality of the general population. Integration of age-specific prevalence and mortality over all ages result in the age-adjusted formulae for a specific year y . In formulae (10) and (11) the minimal age at diagnosis ($\bar{x}_0$) is much lower than the minimal age for calculation of the age-adjusted rates prevalence and mortality (x_0); specifically, $\bar{x}_0 = 30$ years and $x_0 = 65$ years in the example for lung cancer considered below. The use of different values for $\bar{x}_0$ and x_0 is necessary to capture possible non-zero survival to age $x > x_0$ of individuals diagnosed at $\bar{x}_0 \leq \tau \leq x_0$. Therefore, evaluation of age-adjusted prevalence and mortality require models for the incidence, stage at diagnoses, and stage-specific relative survival involving earlier ages and earlier years (along the cohort specific lines in Figure 1).

The mortality model (11) includes two terms: $M(y) = R_p + R_{is}$, where R_p represents the effect of mortality in the general population for individuals with lung cancer and R_{is} represents the effect of trends in cancer incidence, stage ascertainment, and cancer survival relative to general population survival. If relative survival is close to 1, then R_{is} would be small and $M(y) \approx R_p$, i.e., individuals with the disease would have mortality (possibly from other causes) as in the general population. The situation is opposite for lung cancer and we expect $R_p \ll R_{is}$.

The time trends in $P(y)$ and $M(y)$ are determined by the trends in the incidence rate, stage-specific survival, and in ascertainment at early stages as well as mortality in the general population in the case of $M(y)$. Formally the time trend is defined as the derivative of $P(y)$ and $M(y)$ with respect to y . The explicit calculation of this derivative results in:

$$\begin{aligned} \frac{P'(y)}{P(y)} = \frac{1}{P(y)} \int_{x_0}^{\infty} \int_0^{x-\bar{x}_0} \sum_s (I'(\tau, y_d) \pi_s(\tau, y_d) S_s^r(x_\tau, \tau, y_d) \\ + I(\tau, y_d) \pi'_s(\tau, y_d) S_s^r(x_\tau, \tau, y_d) + I(\tau, y_d) \pi_s(\tau, y_d) S_s^{r'}(x_\tau, \tau, y_d)) p(x) dx_\tau dx \end{aligned} \quad (12)$$

Thus, the time trend of disease prevalence is determined through three terms within the brackets of eqn. (12). Each of these terms contains a derivative of the epidemiologic measures associated with incidence, ascertainment at early stages, and patient survival, providing time trends in these measures. Therefore these terms can be interpreted as contributions of incidence, $T_{inc}(y)$, ascertainment at early stages, $T_\pi(y)$, and patient survival, $T_S(y)$, to the total trend of cancer prevalence and the eqn. (12) represents the partitioning of time trends in the prevalence:

$$P'(y)/P(y) = T_{inc}(y) + T_\pi(y) + T_S(y) \quad (13)$$

Explicit expressions for $T(y)$ are given by formulae (5). Negative $T_{inc}(y)$ would mean that prevalence decreases because of decreased incidence (and *vice versa*). Similarly, positive/negative $T_S(y)$ would indicate prevalence changes due to improvement/degradation in incidence-related survival. What is less obvious is whether positive or negative $T_\pi(y)$ is beneficial. Increase of prevalence due to positive $T_\pi(y)$ results from higher fractions of less advanced stages with better survival. Therefore, a positive $T_\pi(y)$ results in increasing prevalence due to a beneficial effect of changes in stage ascertainment (i.e. an increased proportion of cases diagnosed at earlier stages).

Similarly, we obtain the time trend partitioning for mortality:

$$M'(y)/M(y) = \hat{T}_p(y) + \hat{T}_\mu(y) + \hat{T}_{inc}(y) + \hat{T}_\pi(y) + \hat{T}_S(y) \quad (14)$$

Explicit expressions for $\hat{T}(y)$ are given in eqs. (8) and (9). What is new here compared to the case of prevalence are two additional terms associated with change in disease prevalence, $\hat{T}_p(y)$, and mortality in the general population, $\hat{T}_\mu(y)$. These two terms would be responsible for trends in incidence-based mortality if mortality from a considered disease is negligible: in this case change in mortality would be only from changing disease prevalence and mortality (possibly from other causes) in the general population. The effects of the terms $\hat{T}_{inc}(y)$, $\hat{T}_\pi(y)$, and $\hat{T}_S(y)$ describe the additional effects of the prevalence components. Since the expression for $\hat{T}_p(y)$ (eq. (7)) contains three terms associated with the partitioning components, they can be joined with $\hat{T}_{inc}(y)$, $\hat{T}_\pi(y)$, and $\hat{T}_S(y)$ resulting in new terms $\tilde{T}_{inc}(y)$, $\tilde{T}_\pi(y)$, and $\tilde{T}_S(y)$ and the following representation of the trend in incidence based mortality:

$$M'(y)/M(y) = \hat{T}_\mu(y) + \tilde{T}_{inc}(y) + \tilde{T}_\pi(y) + \tilde{T}_S(y) \quad (15)$$

Negative values of these components reflect beneficial changes in public health. Thus we expect the same signs of $T_{inc}(y)$ and $\tilde{T}_{inc}(y)$ but different signs for $T_S(y)$ and $\tilde{T}_S(y)$ as well as for $T_\pi(y)$ and $\tilde{T}_\pi(y)$.

The percentage of the trends in prevalence and mortality explained by a term in (13) and (15) is given by

$$\begin{aligned} R_{inc} &= \frac{T_{inc}(y)}{[T_{inc}(y) + T_\pi(y) + T_S(y)]} \cdot 100\%, \quad \tilde{R}_{inc} \\ &= \frac{\tilde{T}_{inc}(y)}{[\hat{T}_\mu(y) + \tilde{T}_{inc}(y) + \tilde{T}_\pi(y) + \tilde{T}_S(y)]} \cdot 100\% \end{aligned} \quad (16)$$

and similarly for the components associated with survival ($R_S, \tilde{R}_S$), stage ascertainment ($R_\pi, \tilde{R}_\pi$), and mortality in the general population (R_μ). Note that the sums $R_{inc}(y) + R_\pi(y) + R_S(y)$ and could be 100% or -100% for increasing and decreasing prevalence and mortality,

respectively. This definition allows us to keep the same sign in terms (16) after change the trend direction in prevalence and mortality (i.e., when before and after $P'(y) = 0$ or $M'(y) = 0$) and make interpretation simpler and easier.

3. Model Estimation

The quantities of interest (i.e., $T_{\dots}(y)$ and $\hat{T}_{\dots}(y)$) are expressed in terms of the derivatives of the respective functions with respect to time. In our approach, we use explicit analytic parameterizations for all functions for which derivatives are needed. Therefore, we calculate the derivatives analytically. This allows us to avoid dealing with possible numeric instabilities occurring when derivatives are evaluated numerically. For any function we use a two-stage approach: i) we parameterize the function and fit data for each year of diagnosis, and ii) use B-splines¹² to model relationships between year-specific model parameters and evaluate y -dependences of the function. B-splines allow explicit calculation of derivatives without requiring additional simplifying assumptions as discussed in Subsection 3.1. The analysis involves the design and estimation of four models: i) for the incidence rate (Subsection 3.2), ii) for frequencies of stage at onset (Subsection 3.3), iii) for the probability of relative survival after cancer diagnosis (Subsection 3.4), and iv) for mortality in the general population (Subsection 3.5). The distribution of age and time after onset was modeled using the Armitage-Doll model with additional individual predisposition modeled by gamma or inverse Gaussian distributions¹³ (for incidence), the Weibull model for time after disease onset (for relative survival), and the Gompertz model (for mortality). Time trends were modeled by estimating model parameters for each year and applying B-splines to fit the time patterns of obtained parameters.

3.1. B-splines

We need to develop and estimate models for four functions (i.e., incidence, stage at diagnoses, relative survival, mortality in the general population) which are involved in the expressions for prevalence and mortality. For the majority of these, we have available established models for age and time after onset distribution. What we do not have are well-established models for time trends. Our approach involves B-splines for models of time trends. We estimate model parameters for each year and then apply B-splines to fit time patterns of obtained parameters. The important feature of B-splines necessary for our study is that they provide opportunities to calculate derivatives explicitly and without additional simplifying assumptions. Specifically, y -dependence of model parameters are modeled as $p(y) = \sum_i \alpha_i B_{i,n}(y)$, where n is the degree of B-splines ($n = 3$ in our analysis) and i runs over all B-splines the number of which is defined by the number of used knots. A feature important for our analyses is that first derivative of a B-spline is explicitly calculated because $B'_{i,n}(y)$ is represented in terms of B-splines of lower degree.

3.2. Incidence

Many popular carcinogenesis models (e.g., the two-stage clonal expansion model) are not able to appropriately describe the age-pattern of cancer incidence especially at advanced ages. Based on our prior studies^{13,14} we use the model based on: i) the Armitage-Doll

assumption that cancer is the result of the accumulation of a critical number of mutations and ii) the individual predisposition Z that could be modeled by gamma or inverse Gaussian distributions: $h(x, Z) = Zx^{m-1}$. Statistical averaging over the population results in a risk function in the form of a four-parameter model

$$I(x) = \frac{x^{m-1}}{c^m \left(1 + n\sigma^2 c^{-m} x^m\right)^{1/n}}$$

where four model parameters have the following meaning: m is the number of carcinogenesis stages; c (in years) is the scale parameter; σ^2 is the variance of the random predisposition to cancer risk; and n is the parameter describing the shape of frailty distribution ($n = 1$ for gamma-distribution or $n = 2$ for inverse Gaussian distribution). A useful statistical property of this model is that the gamma (or inverse Gaussian) distribution remains the gamma (or inverse Gaussian) distribution after removal of individuals due to hazard $h(t)$. Note, that the parameter c is related to the age of maximal incidence, a_0 , as $c^m = \sigma^2 a_0^m (m + n - mn)/(m(m-1))$

The four model parameters (m, n, a_0, σ^2) are estimated for each calendar year using the age- and sex-specific incidence rates evaluated from the SEER data and the least squares approach. Then these estimates are approximated by B-splines allowing us to have continuous function of incidence $I(x, y)$ and its first derivative $I'_y(x, y)$:

$$I'_y(x, y) = I'_m(x, y)m'_y + I'_n(x, y)n'_y + I'_{a_0}(x, y)(a_0)'_y + I'_{\sigma^2}(x, y)(\sigma^2)'_y$$

Derivatives of model parameters are given by B-splines (we used 1984, 1993, and 2002 as knots and 1973 and 2012 as external knots), while derivatives of $I(x, y)$ over model parameters are calculated explicitly:

$$\frac{I'_m(x, y)}{I(x, y)} = \frac{1}{m} - \left(1 - \frac{(m-1)x^m}{\tilde{n}}\right) \left(\log \frac{a_0}{x} + \frac{1}{\tilde{m}(m-1)}\right)$$

$$\frac{I'_n(x, y)}{I(x, y)} = \frac{1}{n^2} \log \left(1 - \frac{n(m-1)x^m}{\tilde{m}a_0^m}\right) - \frac{m-1}{\tilde{m}} \left(1 - \frac{mx^m}{n\tilde{n}}\right)$$

$$\frac{I'_{a_0}(x, y)}{I(x, y)} = -\frac{m}{a_0} \left(1 - \frac{(m-1)x^m}{\tilde{n}}\right)$$

$$\frac{I'_{\sigma^2}(x, y)}{I(x, y)} = -\frac{1}{\sigma^2}$$

where $\tilde{m} = m n - m - n$, $\tilde{n} = \tilde{m}(x^m - a_0^m) + m x^m$, and model parameters are y-dependent.

3.3. Frequencies of stages at cancer diagnosis

Three stages are distinguished in SEER data for a cancer: local ($s = 1$), regional ($s = 2$), and distant ($s = 3$). Frequencies of being diagnosed at stage $s = 1$ and $s = 2$ are modeled for each year as quadratic functions of age $f_s = f_{s0} + f_{s1}\tau + f_{s2}\tau^2$ and $f_3 = 1 - f_1 - f_2$. Derivatives are $f'_s = f'_{s0} + f'_{s1}\tau + f'_{s2}\tau^2$ and $f'_3 = -f'_1 - f'_2$. Year dependence of f_{si} are subject of modeling by B-splines.

3.4. Relative survival after cancer diagnosis

The survival function for cancer patients is estimated directly from SEER data using the Kaplan-Meier approach. Survival in the general population is taken from the life-tables available in the Human Mortality Database. Then the relative survival is estimated and represented in $\log(-\log S(x_\tau))$ vs. $\log(x_\tau)$ in Supplementary Figure 1. We see that the linear approximation for these curves is reasonable, and therefore the Weibull model for relative survival is appropriate.

The parameters for the relative survival model for a specific year are estimated through maximizing the likelihood. Using the standard likelihood for total survival $L = \prod_i h_i^{d_i} S(x_{0i}, x_i)$ and the definitions of relative survival, $S(x_{0i}, x_i) = S_{pop}(x_{0i}, x_i) S_{rel}(x_{0i}, x_i)$, or for respective hazard functions $h(x) = h_{pop}(x) + h_{rel}(x, \beta)$, we construct the log likelihood as following Dickman et al.¹⁵

$$l(\beta) = - \sum_i \int_{x_{0i}}^{x_i} h_{pop}(u) du - \sum_i \int_{x_{0i}}^{x_i} h_{rel}(u, \beta) du + \sum_i d_i \log(h_{pop}(x_i) + h_{rel}(x_i, \beta))$$

Here x_{0i} and x_i are the initial and final ages of individual follow-up, d_i is the death indicator and β is the set of parameters of interest. The first term does not depend on β and therefore can be omitted. The only thing what we need to know about the general population is the population hazards at the age of death for all individuals in the data. This information is obtained from the Human Mortality Database.

Relative survival probability is modeled using the Weibull distribution for time to event. In this case the survival and hazard functions are $S_{rel} = \exp(-\alpha x_\tau^\gamma)$ and $h_{rel} = \gamma \alpha x_\tau^{\gamma-1}$, ($x_\tau = x - \tau$ is the survival time) where γ is the Weibull shape parameter, and $\alpha = \exp(-\gamma\mu) = \lambda^{-\gamma}$ (i.e., $\lambda = \exp(\mu)$) and λ is the Weibull scale parameter. Parameter estimates in age groups show that parameters can have quadratic dependence on age. Only two parameters are independent and the likelihood ratio test shows that the largest likelihood is for $\sigma = 1/\gamma$ and μ . The

survival function in these parameters is: $S_{rel} = \exp(-\exp(\sigma^{-1}(\log x_\tau - \mu)))$. Thus the effects of age (or age at diagnosis) are modeled as

$$\mu = \mu_0 + \mu_1\tau + \mu_2\tau^2, \quad \sigma = \sigma_0 + \sigma_1\tau + \sigma_2\tau^2$$

Thus $\beta = \{\mu_{0-2}, \sigma_{0-2}\}$. The set of parameter together with their first derivatives $\beta' = \{\mu'_{0-2y}, \sigma'_{0-2y}\}$ over y is estimated for each year and then B-spline parameterization is fitted using 1994, 2000, and 2006 as knots and 1986 and 2014 as external knots.

The first derivative of S_{rel} with respect to y is (subscript “rel” is dropped for convenience) defined by the expressions: $S'_y(x, y) = S'_\mu(x, y)\mu'_y + S'_\sigma(x, y)\sigma'_y$, where $S'_\mu(x, y) = \gamma\alpha x_\tau^\gamma S_{rel}(x, y)$, $S'_\sigma(x, y) = (\log(x_\tau) - \mu)\gamma^2\alpha x_\tau^\gamma S_{rel}(x, y)$, $\mu'_y = \mu'_{0y} + \mu'_{1y}\tau + \mu'_{2y}\tau^2$, $\sigma'_y = \sigma'_{0y} + \sigma'_{1y}\tau + \sigma'_{2y}\tau^2$

3.5. Mortality in the general population

We use the Gompertz model to parameterize year specific mortality patterns: $\mu(x, y) = \exp(a(y) + b(y)x)$. Again, the B-spline model is used to create a continuous functions of $a(y)$ and $b(y)$.

4. Application to Lung Cancer

Our incidence model is shown to fit the empirical age-specific incidence rates (Figure 2) and empirical age-adjusted incidence (Figure 3) with high precision. Lung cancer incidence reaches a peak between ages 75 and 80 for both genders. Age-adjusted incidence in males reaches a peak on or about 1985 and has been declining since. For women, a peak does not appear until 2008—more than two decades later. Changes in the relative frequencies of cancer stage (local, regional, distant) at diagnosis (Figure 4) show that patterns for males and females are similar, e.g., both genders have a maximum in the frequency of distant stage at 2005 possibly due to the National Lung Screening Trial launched in 2002¹⁶. However, the general tendencies are different: the trends of “average stage” at diagnosis show that in females lung cancer is being diagnosed at earlier stages and the opposite is true for males. The relative survival is parameterized using the Weibull model. This choice is justified by the linear nature of the empiric rates presented in $\log(-\log S(x_\tau))$ vs. $\log(x_\tau)$ (Supplementary Figure 1). The resulting model (Figure 5) shows an adequate fit of relative survival compared to empirically observed estimates and the estimates calculated using SEER*Stat software.

Figure 6 presents age-adjusted lung cancer prevalence obtained using our model (10) in comparison with estimates obtained using SEER*Stat for 18 and 23 years of look-back period as well as that observed in SEER-Medicare for a 6-month look-back period. The agreement between empirical data and the model is good: the shape of the model mirrors empirical patterns, and differences in magnitude are minor. The results of our mortality model are presented in Figure 7. Our model (11) shows good agreement with rates derived from the MCD even though the definitions of mortality between the two methods differ (incidence-based vs. death-certificate-based). As expected, $R_p \ll R_{is}$, i.e., R_{is} dominates in

determining mortality. The results show that incidence-based mortality has been decreasing for males and is starting to decrease for females after a relatively long plateau.

Partitioning of disease prevalence and incidence-based mortality shown in Figure 8 contain three (trends of incidence, stage ascertainment, and relative survival) and five (trends of incidence, stage ascertainment, relative survival, and prevalence/mortality in the general population) contributing components respectively. The increase in lung cancer prevalence for females is due to increased incidence (explains approximately 45–50% of total trend) and improved survival (40–45% of total trend). The remaining 10–15% are explained by increased ascertainment at early stages. These percentages are valid for 1998. Before 1998 increased prevalence was (up to 80%) due to trends in incidence and after 1998 trends in survival become dominant. Similarly, the increase in mortality for females up to 2002 is due to the effect of increased incidence. After 2002 the trend goes to zero and changes sign (i.e., mortality decreases). This happened due to the increased impact of survival after 2002. For males, the effect of increased survival is compensated by a rapidly declining incidence rate resulting in an overall decrease in prevalence over time. Stage ascertainment plays an adverse role in prevalence for males with maximal contribution in 2005 (it compensates about 30% of the contribution of decreased incidence). Incidence is the main factor in decreasing mortality for males, though improved survival also pushes mortality down starting from 2002. For example, in 2005 the contribution of relative survival and incidence decreased mortality were about 30% and 60%. Remaining 10% are due to decreased mortality in the general population.

Application of the partitioning approach to histotype-specific prevalence for adenocarcinoma, squamous-cell carcinoma, and small-cell carcinoma showed that the patterns of partitioning observed for all histotypes (Figure 9) hold for histotype-specific cancers (Supplementary Figures 2&3), e.g., the trend in incidence is the dominant contribution in determining the trends of histotype-specific lung cancer prevalence and incidence-based mortality. However, some effects not in line with the general trend were observed. First, starting with 2005, incidence for adenocarcinoma increased for males resulting in increased prevalence of adenocarcinoma. This is reflected in mortality as well: starting from 2005 the slope of mortality trend rapidly increases. The same pattern is observed for females (minimal slope in 2005), though both prevalence and mortality in females increased during the entire observation period. Second, compared to other histotypes, ascertainment at early stages plays a more important role for squamous-cell and small-cell carcinoma than for adenocarcinoma, however, the direction of the effect is not stable over time.

Further steps in analyses of the constituent components contributing to partitioning of lung cancer prevalence and mortality can be performed based on the relative contributions of model components given by eqns. (16) that are presented in Supplementary Table 1. The estimates show that for all histotype combined in 2000: i) declining prevalence in males is due to decreased incidence (–136%), improved relative survival (66%), and degraded stage ascertainment (–30%); ii) increasing prevalence in females is due to increased incidence (33%), improved survival (62%), and stage ascertainment (5%); iii) decreased mortality in males are basically due to incidence (–102%) while other contributions are below 10%; and

iv) increase mortality for females is due to increased incidence (139%), improved survival (−53%), disimproved stage ascertainment (25%), and trend in mortality in the general population (−11%).

Maximal relative increase in prevalence ($P'(y)/P(y)$) of adenocarcinoma was detected in the last years of study period: in 2008 it was i) for males due to increased decreased incidence (−31%), survival (125%), and improved stage ascertainment (6%) and ii) for females due to increased incidence (42%), increased survival (55%), and improved stage ascertainment (4%). Maximal decline in adenocarcinoma relative mortality in males was detected in period of 2001–2004: decreased incidence (−52%), improved survival (−23%), improved stage ascertainment (−3%), and mortality in the general population (−22%); estimates are given for 2003. For females maximal increase was in the beginning of study period almost completely due to increased incidence, e.g., in 1996 the contribution of increased incidence was 95% and other contributions were below 10%.

Decreased incidence of squamous cell carcinoma is main driving force of decline of prevalence and mortality in males, though survival and stage ascertainment gave noticeable contributions especially in the last years: if the contributions to the relative prevalence from incidence, survival stage ascertainment were 124%, 13%, and 11% in 2000, then they became −234%, 182%, and −48% in 2008. Similar results are obtained for mortality. In females prevalence increased in two periods before 1998 and after 2004 although because of different reasons: the contributions were 73%, −31%, and 39% in 1996 while in 2006 the contributions were −32%, 146% and −13%, i.e., if before 2000 prevalence increased because of increased incidence (what is disadvantages), after 2000 increased prevalence is largely due to improved survival (what is beneficial). These conclusions are confirmed by the analyses of mortality.

Trends in prevalence and mortality of small cell carcinoma are the results of a trade-off between two beneficial tendencies: decreased incidence and improved survival. The result of this trade-off is however modulated by the noticeable contributions of stage ascertainment, which is beneficial for males but not for females. For example, in 2006 the contributions to male prevalence were −128%, 111%, and 117%, and −45%, 205%, and −60% for female. After 2006 the stage ascertainment for males began to worsen. Extreme decrease in mortality was detected in 2000–2003. The contributions from incidence, survival, stage ascertainment, and mortality in the general population in 2002 were −66%, −7%, −24%, and −2% for males and −79%, −22%, 3%, and −2% for females.

5. Discussion

The approach developed in this paper provides high precision estimates of the time-trends of lung cancer prevalence and mortality and explains them in terms of trends of lung cancer incidence rate, stage-specific relative survival, and ascertainment at early stages. Lung cancer prevalence has been decreasing for males and increasing for females over the study period. Partitioning shows that in females this increase was due to increasing incidence rates (explain up to 45–50% of prevalence trend in 2000) and improved survival (up to 40–45% of the trend in 2000); in males, the effect of improved survival was overpowered by the

substantially declining incidence rate leading to an overall prevalence decline. This means that the dynamic of lung cancer incidence is an important factor that causes sex-specific differences in lung cancer prevalence. In males lung cancer prevalence consistently decreases till 2005 with decreasing incidence being the primary contributor to this outcome during this time period. The prevalence in females increased throughout the entire period. It was determined by increased incidence up until 1998, after which improved survival became the primary contributing factor.

Trends in male mortality have been decreasing over the study period, and such a trend is not observed in females (Figure 6). This is in agreement with other epidemiological studies that showed that in males lung cancer death rates in the U.S. declined 38% in 1990–2012 and in females mortality started declining (up to 13%) between 2002 and 2012². Partitioning of mortality shows that changes in mortality are determined by changes in the incidence rate and improvement in survival plays a more notable role from 2000 onwards for both genders, wherein the role of factors contributing to lung cancer mortality trends differs in males and females. While the increase in mortality in females before 2002 can be predominantly explained by increasing incidence and decreasing mortality, after 2002 it is largely due to improved survival. Declining mortality in males is predominantly due to decreasing incidence (up to 60%) and to a lesser degree due to improving survival (up to 30%) and decreasing mortality in the general population (10%). As we discuss below sex-specific differences in dynamics of tobacco smoking have been suggested as the main explanation for these trends.

Partitioning analysis requires estimation and modeling of age- and time patterns of the constituent partitioning components, i.e., incidence, stage at diagnosis, and stage-specific relative survival. All empirically estimated age- and time-patterns of constituent components are in agreement to those known in literature. Incidence of lung cancer in males has been declining since the mid-1990s. Female lung cancer incidence increased until 2007, and has been decreasing at an annual percentage change of 2.6¹⁷ since then. These findings are consistent with the literature which shows that gender-specific differences in lung cancer incidence reflect historical differences in tobacco use when males started showing declining trends earlier than females^{18,19}, and suggest that studies predicting a major increase in lung cancer incidence in females by 2020²⁰ may turn out to be overly pessimistic. If currently observed trends continue, we would expect to see an additional decrease in the incidence of histological types strongly dependent on smoking, notably small-cell, squamous-cell carcinoma, and large-cell carcinomas, while adenocarcinoma trend likely remains constant or continues to increase¹⁷. The results on histotype-specific incidence were in agreement with conclusions of other studies that showed that since 1970s in both males and females SSCs incidence decreased and ACs increased¹⁷. These changes were suggested to be associated with tobacco smoking prevalence²¹, histology-specific differences in smoking dose-relationships, the latency period between exposure and outcome, and impact of changes in cigarette composition^{22–24}.

We found that survival from lung cancer increases the prevalence for both genders. This demonstrates success of health care in lung cancer treatment²⁵. Partitioning of prevalence and mortality shows that the contribution of stage ascertainment has an adverse effect till

2005, although this effect is small when compared to incidence and survival. Since stage at diagnosis is one of the primary determinants of lung-cancer mortality (the 5-year survival rate is 55% for cases detected when the disease is still localized, 27% for regional disease, and 4% for distant stage disease²⁶) this indicates the need for improved screening methods (e.g., a spiral computed tomography²⁷), to improve early stage ascertainment.

The partitioning approach we used allows researchers to look deeper into the trends affected by cancer screening; without application of partitioning analysis some of these trends remain undistinguished and unidentified. For example, it is known that screening (as well as other changes in cancer detection practices) introduces the lead time bias (lead time is the length of time between the detection of a disease based on new/experimental method/criteria and its usual clinical presentation and diagnosis based on traditional method/criteria) in survival rates, thereby potentially affecting quality of evaluation of progress against cancer in epidemiological studies as well as understanding of resulting cancer prevalence and mortality²⁸. With partitioning approach developed in this study, the percent of the contribution of early diagnosis and survival to cancer incidence and mortality can be evaluated separately.

Because the method of partitioning provides insight into the components that impact observed cancer trends in prevalence and mortality (through the contribution of dynamics in cancer incidence, early-stage diagnoses, survival, and prevalence and mortality in general population), this method can be used for projection of trends of cancer prevalence and mortality with (or without) considering specific interventions. For example, our studies showed that if currently observed trends continue, we would expect to see increased prevalence for both genders and this increase would be due to improved survival. We would expect additional decreases in the incidence of histological types strongly dependent on smoking, notably small-cell, squamous-cell carcinoma, and large-cell carcinomas, while trends in adenocarcinoma are likely to remain constant or continue to increase¹⁷. Trends in lung cancer incidence largely determine the trend in mortality. Only in recent years the contribution from survival became more important and sometimes comparable in magnitude to the effect of incidence. Thus, based on our analysis, continuation of current trends is likely to result in even higher declines in mortality than is predicted by the trends in incidence.

Furthermore, our methodology allows for exploring the possible effect of health campaigns (such as anti-smoking efforts). For example, if the peak in smoking prevalence for males and females occurred at the same time (in reality they differ by 10 years¹⁸) we would expect the same peak in lung cancer incidence and mortality. However, our estimates show that incidence for females declined much slower than for males which reflects the real-world change is smoking prevalence and is confirmed by other studies^{20,25,29}. Better estimates of interventions involving smoking patterns would be possible by precise estimating age- and cohort-specific effect of smoking on lung cancer incidence and incorporating in the models for lung cancer incidence we use. We note that the use of the empiric parametric models is a limitation of our approach. The calculations involve the models for age-dependence of incidence and fractions of stage at diagnosis as well as age- and time-after-diagnosis dependence for survival. Although we carefully constructed such models based on their

empiric properties or widely accepted modeling approaches, statistical uncertainties may still remain.

5.1. Methodological development

In this paper we presented a new approach for cancer epidemiology allowing researchers to partition the influence of changes in rates of incidence, stage-specific survival, and stage of case ascertainment on overall cancer prevalence and mortality. This approach expands on the basic decomposition approach introduced by Akushevich et. al.³ Thus, we demonstrated that the methodology is easily expandable to incorporate greater levels of detail as the data becomes available and implement a number of methodological innovations including: i) application of calculus approaches to obtain analytic expressions for prevalence and mortality, ii) development of empiric models for incidence, fractions of stage at diagnoses, and survival as well as validating model fit, iii) use of B-splines for evaluation of parameters responsible for time evolution of incidence and survival, and iv) numeric calculation of two-dimensional integrals dependent on parameters of B-splines. In addition, extrapolations of the time trends modeled by this approach provide forecasts of disease prevalence and mortality. Such forecasts could be used in disease burden analyses including predictions of the numbers of hospitals/beds, amount of necessary medicine for lung cancer treatment, as well as for development of design of screening strategies in certain subpopulations and regions.

There are no approaches in the literature that address the problem of partitioning the influence of changes in rates of incidence, stage-specific survival, and stage at time of diagnosis on overall cancer prevalence and mortality, however, there is a series of approaches that deal with other kinds of decomposition and partitioning of health related data. Partitioning of breast cancer mortality was analyzed by Chu et al.^{11,30} who developed and applied an approach which allows the partitioning of mortality by variables associated with cancer onset such as stage and age at diagnosis. No connection between trends in incidence and survival were performed. Another related approach is developed in ref³¹ in which the total excess in mortality experienced by patients diagnosed with breast cancer is partitioned into excess cardiovascular mortality and remaining cancer excess mortality. This approach was applied to the case of Hodgkin Lymphoma³². This approach involves flexible parametric models^{33,34} for survival analysis that were developed by Royston and Parmar³⁵ and extended to relative survival by Nelson et al.³⁶.

These models generalized the class of parametric models that we used for modeling of relative survival in the partitioning approach and can be used in future generalizations to improve the accuracy of partitioning analyses developed in this study. Another method used for decomposition of cancer outcome measures is the standardization and decomposition analyses (SDA)^{37–39}. The idea of the decomposition is similar to the Blinder-Oaxaca decomposition that is widely used in economics but is inherently limited by the assumptions made by the regression/logistic models used for estimation⁴⁰. In SDA, the observed outcome difference between populations is split into component effects that are attributed to: i) the “real” outcome difference; and ii) compositional difference of specific confounding factors between populations, e.g., age structure or racial/ethnicity disparities⁴¹. In the future the

ideas of SDA and the partitioning analysis presented in this paper can be combined for analyses of disparities. The three partitioning components (incidence, distribution of stage at diagnosis, and relative survival) can be considered as the specific confounding factors between race-specific populations. Difference in the race-specific prevalence or mortality rates can be decomposed into three respective terms using formulae (2.19–2.22) of ref.³⁷ resulting in three effects (coming from each component) affecting the disparities. Then application of partitioning formulas to each component will allow us to explain changes in time trends of the disparities in terms of these determining components.

As we discussed in the introduction, our approach for partitioning was originally developed for another structure of data (Medicare claims data)^{3,42,43}. The development of the original model was motivated by the necessity to improve and generalize for chronic diseases a partitioning of mortality trends through the use of an approximate formula for the simple decomposition of the annual percent change (APC) for mortality from acute conditions as a sum of APC's of cardiovascular disease incidence and case fatality (percentage of 28-day fatalities)⁴⁴. Other methods of decomposition used in demography and epidemiology (see refs.^{45–47} for a comprehensive review), are not directly related to the decomposition of cancer prevalence and mortality into its constituent epidemiologic components. For example, the so-called recursive partitioning analysis based on decision tree classification was also applied to analyses of cancer data^{48–53}. The role of the recursive partitioning analysis in identifying Non-Small Cell Lung Cancer Patients with Brain Metastases was discussed in⁵⁴.

6. Conclusion

In this paper we developed a new statistical approach providing quantitative information on factors contributing to cancer trends, a clear interpretation of the obtained results that can form the basis for both predictive inference and formal forecasting models. Specifically, partitioning reveals the mechanisms of the interrelations between cancer incidence, prevalence, and survival—an important step in assessing the potential burden posed by a disease, potential problem areas in need of attention, as well as how these important epidemiological characteristics respond to implemented health interventions over time. The developed approach allows for numeric estimates of percentages of the contributions of changes in incidence, stage at diagnoses, and relative survival to the overall change in cancer prevalence and mortality. In the case of lung cancer, the main driving forces underlying the observed trends in prevalence and mortality were: i) declining male and increasing female incidence (reflect changes in smoking rates in past); ii) improved stage-specific survival—after 2000 survival is the primary determinant of female prevalence trends on par with the size of the contribution of incidence in males; and iii) trends in stage ascertainment playing low but adverse role in the trends in males (all years) and females (between 2002–2008) with maximal adverse effect in 2005 for both genders. The partitioning analysis of histotype-specific prevalence and mortality of lung cancer showed the increased role of adenocarcinoma incidence since 2005 and higher relative contributions stage ascertainment to the prevalence and mortality trends of squamous-cell and small-cell carcinomas. Based on the results of our analysis we infer that further improvement in lung cancer mortality can be achieved by improvement of early stage ascertainment especially for males, and that in spite of success in its treatment, adenocarcinoma still has disadvantageous trends (especially in

female incidence) and its role among other histology-specific lung cancers will increase in near future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

This work was supported by the National Institute on Aging (grants R01-AG046860, P01-AG043352).

References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*. 2015;136(5).
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. *CA: a cancer journal for clinicians*. 2016;66(1):7–30. [PubMed: 26742998]
3. Akushevich I, Yashkin A, Kravchenko J, et al. Theory of partitioning of disease prevalence and mortality in observational data. *Theoretical Population Biology*. 2017;114:117–127. [PubMed: 28130147]
4. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973–2015), National Cancer Institute, DCCPS, Surveillance Research Program, released April 2018, based on the November 2017 submission.
5. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Populations - Total U.S. (1969–2016) <Katrina/Rita Adjustment> - Linked To County Attributes - Total U.S., 1969–2016 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released November 2017.
6. Akushevich I, Kravchenko J, Yashkin AP, Yashin AI. Time trends in the prevalence of cancer and non-cancer diseases among older US adults: Medicare-based analysis. *Experimental Gerontology*. 2018;110:267–276. [PubMed: 29932968]
7. United States Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS), Underlying Cause of Death 1999–2017 on CDC WONDER Online Database, released 2018. Data are compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.
8. Diekmann O, Heesterbeek JAP. Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. Vol 5: John Wiley & Sons; 2000.
9. Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. Princeton University Press; 2011.
10. Hethcote HW. The mathematics of infectious diseases. *SIAM review*. 2000;42(4):599–653.
11. Chu KC, Miller BA, Feuer EJ, Hankey BF. A method for partitioning cancer mortality trends by factors associated with diagnosis: An application to female breast cancer. *Journal of clinical epidemiology*. 1994;47(12):1451–1461. [PubMed: 7730854]
12. De Boor C, De Boor C, Mathématicien E-U, De Boor C, De Boor C. A practical guide to splines. Vol 27: Springer-Verlag New York; 1978.
13. Kravchenko J, Akushevich I, Abernethy AP, Lyerly HK. Evaluating the number of stages in development of squamous cell and adenocarcinomas across cancer sites using human population-based cancer modeling. *PloS one*. 2012;7(5):e37430. [PubMed: 22629394]
14. Manton KG, Akushevich I, Kravchenko J. Cancer mortality and morbidity patterns in the US population: an interdisciplinary approach. Springer Science & Business Media; 2008.
15. Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Statistics in medicine*. 2004;23(1):51–64. [PubMed: 14695639]
16. Sharma D, Newman TG, Aronow WS. Lung cancer screening: history, current perspectives, and future directions. *Archives of medical science: AMS*. 2015;11(5):1033. [PubMed: 26528348]

17. Meza R, Meernik C, Jeon J, Cote ML. Lung cancer incidence trends by gender, race and histology in the United States, 1973–2010. *PloS one*. 2015;10(3):e0121323. [PubMed: 25822850]
18. Harris JE. Cigarette Smoking Among Successive Birth Cohorts of Men and Women in the United States During 1900–1980 2. *Journal of the National Cancer Institute*. 1983;71(3):473–479. [PubMed: 6577223]
19. Jemal A, Ma J, Rosenberg PS, Siegel R, Anderson WF. Increasing lung cancer death rates among young women in southern and midwestern States. *Journal of Clinical Oncology*. 2012;30(22):2739–2744. [PubMed: 22734032]
20. Weir HK, Thompson TD, Soman A, Møller B, Leadbetter S. The past, present, and future of cancer incidence in the United States: 1975 through 2020. *Cancer*. 2015;121(11):1827–1837. [PubMed: 25649671]
21. Kenfield SA, Wei EK, Stampfer MJ, Rosner BA, Colditz GA. Comparison of aspects of smoking among four histologic types of lung cancer. *Tobacco control*. 2008;tc.2007.022582.
22. Khuder SA, Mutgi AB. Effect of smoking cessation on major histologic types of lung cancer. *Chest*. 2001;120(5):1577–1583. [PubMed: 11713137]
23. US Department of Health Human Services. The health consequences of smoking—50 years of progress: a report of the Surgeon General. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health. 2014;17.
24. Burns DM, Anderson CM, Gray N. Do changes in cigarette design influence the rise in adenocarcinoma of the lung? *Cancer Causes & Control*. 2011;22(1):13–22. [PubMed: 20967496]
25. Edwards BK, Noone AM, Mariotto AB, et al. Annual Report to the Nation on the status of cancer, 1975–2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. *Cancer*. 2014;120(9):1290–1314. [PubMed: 24343171]
26. Miller KD, Siegel RL, Lin CC, et al. Cancer treatment and survivorship statistics, 2016. *CA: a cancer journal for clinicians*. 2016;66(4):271–289. [PubMed: 27253694]
27. Team NLSTR. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011;2011(365):395–409.
28. Welch HG, Schwartz LM, Woloshin S. Are increasing 5-year survival rates evidence of success against cancer? *Jama*. 2000;283(22):2975–2978. [PubMed: 10865276]
29. Jemal A, Thun MJ, Ries LA, et al. Annual report to the nation on the status of cancer, 1975–2005, featuring trends in lung cancer, tobacco use, and tobacco control. *JNCI: Journal of the National Cancer Institute*. 2008;100(23):1672–1694. [PubMed: 19033571]
30. Chu KC, Tarone RE, Kessler LG, et al. Recent trends in US breast cancer incidence, survival, and mortality rates. *JNCI: Journal of the National Cancer Institute*. 1996;88(21):1571–1579. [PubMed: 8901855]
31. Eloranta S, Lambert PC, Andersson TM, et al. Partitioning of excess mortality in population-based cancer patient survival studies using flexible parametric survival models. *BMC medical research methodology*. 2012;12(1):86. [PubMed: 22726307]
32. Eloranta S, Lambert PC, Sjöberg J, Andersson TM, Björkholm M, Dickman PW. Temporal trends in mortality from diseases of the circulatory system after treatment for Hodgkin lymphoma: a population-based cohort study in Sweden (1973 to 2006). *Journal of Clinical Oncology*. 2013;31(11):1435–1441. [PubMed: 23439756]
33. Andersson TM, Dickman PW, Eloranta S, Lambert PC. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC medical research methodology*. 2011;11(1):96. [PubMed: 21696598]
34. Andersson TML, Dickman PW, Eloranta S, Lambe M, Lambert PC. Estimating the loss in expectation of life due to cancer using flexible parametric survival models. *Statistics in medicine*. 2013;32(30):5286–5300. [PubMed: 24038155]
35. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine*. 2002;21(15):2175–2197. [PubMed: 12210632]

36. Nelson CP, Lambert PC, Squire IB, Jones DR. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in medicine*. 2007;26(30):5486–5498. [PubMed: 17893893]
37. Gupta PD. Standardization and decomposition of rates: A user's manual. US Department of Commerce, Economics and Statistics Administration, Bureau of the Census; 1993.
38. Yuan C, Wei C, Wang J, et al. Self-efficacy difference among patients with cancer with different socioeconomic status: Application of latent class analysis and standardization and decomposition analysis. *Cancer epidemiology*. 2014;38(3):298–306. [PubMed: 24656649]
39. Wang J, Carlson RG, Falck RS, Leukefeld C, Booth BM. Multi-sample standardization and decomposition analysis: an application to comparisons of methamphetamine use among rural drug users in three American states. *Statistics in medicine*. 2007;26(19):3612–3623. [PubMed: 17243192]
40. Fairlie RW. An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. *Journal of economic and social measurement*. 2005;30(4):305–316.
41. Wang J. Standardization and Decomposition Analysis: A Useful Analytical Method for Outcome Difference, Inequality and Disparity Studies. *Innovative Statistical Methods for Public Health Data*: Springer; 2015:251–264.
42. Akushevich I, Yashkin AP, Kravchenko J, et al. Identifying the causes of the changes in the prevalence patterns of diabetes in older US adults: A new trend partitioning approach. *Journal of Diabetes and Its Complications*. 2018;32(4):362–367. [PubMed: 29433960]
43. Akushevich I, Yashkin A, Kravchenko J, et al. A Forecasting Model of Disease Prevalence Based on the McKendrick-von Foerster Equation. *Mathematical biosciences*. 2019;311:31–38. [PubMed: 30597156]
44. Tunstall-Pedoe H, Kuulasmaa K, Mähönen M, Tolonen H, Ruokokoski E, Amouyel P. Contribution of trends in survival and coronary y-event rates to changes in coronary heart disease mortality: 10-year results from 37 WHO MONICA Project populations. *The Lancet*. 1999;353(9164):1547–1557.
45. Canudas-Romo V. *Decomposition methods in demography*. Rozenberg Publishers Amsterdam; 2003.
46. Horiuchi S, Wilmoth JR, Pletcher SD. A decomposition method based on a model of continuous change. *Demography*. 2008;45(4):785–801. [PubMed: 19110897]
47. Vaupel JW, Romo VC. Decomposing change in life expectancy: A bouquet of formulas in honor of Nathan Keyfitz's 90th birthday. *Demography*. 2003;40(2):201–216. [PubMed: 12846129]
48. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*. 2006;15(3):651–674.
49. Hwang SS, Scott CB, Chang VT, Cogswell J, Srinivas S, Kasimis B. Prediction of survival for advanced cancer patients by recursive partitioning analysis: role of Karnofsky performance status, quality of life, and symptom distress. *Cancer investigation*. 2004;22(5):678–687. [PubMed: 15581048]
50. Gaspar LE, Scott C, Murray K, Curran W. Validation of the RTOG recursive partitioning analysis (RPA) classification for brain metastases. *International Journal of Radiation Oncology* Biology* Physics*. 2000;47(4):1001–1006.
51. Werner-Wasik M, Scott C, Cox JD, et al. Recursive partitioning analysis of 1999 Radiation Therapy Oncology Group (RTOG) patients with locally-advanced non-small-cell lung cancer (LA-NSCLC): Identification of five groups with different survival. *International Journal of Radiation Oncology* Biology* Physics*. 2000;48(5):1475–1482.
52. Kattan MW, Hess KR, Beck JR. Experiments to determine whether recursive partitioning (CART) or an artificial neural network overcomes theoretical limitations of Cox proportional hazards regression. *Computers and biomedical research*. 1998;31(5):363–373. [PubMed: 9790741]
53. Cooper JS, Farnan NC, Asbell SO, et al. Recursive partitioning analysis of 2105 patients treated in Radiation Therapy Oncology Group studies of head and neck cancer. *Cancer: Interdisciplinary International Journal of the American Cancer Society*. 1996;77(9):1905–1911.
54. Liu S, Chen P, Liu Y-W, Gu X-N, Qiu X-G, Li B. Role of Recursive Partitioning Analysis and Graded Prognostic Assessment on Identifying Non-Small Cell Lung Cancer Patients with Brain

Metastases Who May Benefit from Postradiation Systemic Therapy. Chinese medical journal. 2018;131(10):1206. [PubMed: 29722340]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1.

Lexis diagram showing the age-time area for which data are available, and representing events (such as disease onset or death) that occur in individuals belonging to different cohorts. Calendar time is on the horizontal axis, while age is on the vertical axis. Dashed lines show time/age points for specific cohorts. The area in which partitioning components are evaluated is denoted by a rectangle. Information for the cohorts beyond the lower-left bound is used to produce accurate initial values for epidemiologic quantities used in the study.

Fig 2. Age-patterns of lung cancer incidence rates: empiric estimates using the SEER data (open dots with errors bars) and the estimated Armitage-Doll model with random predisposition.

Fig 3.
Empiric estimates using SEER data (open dots with errors bars) and the B-spline model for calendar year dependence (curves) for age-adjusted (65+) lung cancer incidence rates.

Fig 4. Empiric estimates (open dots with error bars) and B-spline model (curves) for the year dependence of frequencies of diagnoses at a specific stage (f_{si})

Fig 5. Model predictions (solid line), empiric estimates of relative survival (closed points) with 95% confidence intervals (grey area), and SEER*Stat estimates of relative survival (open points with error bars). Three sets of points at each plot correspond to local, regional, and distant stages (from top to bottom).

Fig 6. Lung cancer prevalence i) predicted using our model (solid line), ii) calculated using SEER*Stat with an 18-year lookback period (closed points), iii) calculated using SEER*Stat with a 23-year lookback period (squares), and iv) observed in SEER-Medicare with a 6-month look-back period (open points).

Fig 7.

Lung cancer mortality i) predicted using our model (solid line) and two terms R_p (dotted line) and R_{is} (dashed line), ii) calculated from MCD data using underlying cause of death only (open points), iii) calculated from MCD data using underlying and secondary causes of death (closed points).

Fig 8.

Partitioning of the prevalence trend into its contributors: $P'(y)/P(y) = T_{inc}(y) + T_{st}(y) + T_{surv}(y)$

Fig 9.

Partitioning of the mortality trends into its contributors:

$$M'(y)/M(y) = \hat{T}_P(y) + \hat{T}_\mu(y) + \hat{T}_{inc}(y) + \hat{T}_\pi(y) + \hat{T}_S(y)$$

Table 1.

Summary of Mathematical Functions

Ages, cancer stage, and calendar times used as arguments/indices in the survival functions			
x	Current age in years	y	Current (calendar) time in years
x_0	Minimal age for calculation age-adjusted rates	$y_d = y - x + \tau$	Time of diagnosis
$\bar{x}_0$	The minimal age at diagnosis (see Figure 1)	$y_b = y - x$	Time of birth for a cohort
$\tau, \bar{x}_0 < \tau \leq x$	Age at diagnosis	$x_\tau = x - \tau = y - y_d$	Time after diagnosis
s	Stage at diagnosis		

Survival analysis functions for age-specific prevalence and mortality	
$P_c(x, y_b)$	Probability of being prevalent at age x in cohort with birth time y_b
$I_c(\tau, y_b)$	Incidence density function for birth cohort y_b . The normalization rule for the density is $\int_{\bar{x}_0}^{\infty} I_c(\tau, y_b) d\tau = 1$
$M_c(x, y_b)$	Mortality density function for birth cohort y_b . The normalization rule for the density is $\int_{\bar{x}_0}^{\infty} M_c(x, y_b) dx = 1$
$\pi_s(\tau, y_d)$	The probability to detect stage at diagnosis, s . The normalization rule for the probability is $\sum_s \pi_s(\tau, y_d) = 1$
$S_s(x - \tau, \tau, y_d) f_s(x - \tau, \tau, y_d) = - (S_s(x - \tau, \tau, y_d))'_x$	Stage-specific survival function and respective density function of a patient group diagnosed at age τ and time y_d and survived to x (i.e., living $x - \tau$ years after diagnosis)
$S_t(x - \tau, \tau, y_d)$ and $\mu(x, y)$	Survival function and respective mortality hazard function in the general population for the cohort formed at age τ and y_d i.e., $S_t(0, \tau, y_d) = 1$.

Survival analysis functions for age-adjusted prevalence and mortality	
$P(x, y)$	Prevalence at age x and time y
$I(x, y)$	Incidence hazard function at age x and time y
$M(x, y)$	Incidence based mortality hazard function at age x and time y
$P(y)$	Age-adjusted prevalence at time y
$M(y)$	Age-adjusted incidence-based mortality hazard function at time y
$S^r(x - \tau, \tau, y_d) f^r(x - \tau, \tau, y_d) = - S^{r'}_x(x - \tau, \tau, y_d)$	Relative survival and respective density function of a patient group diagnosed at age τ and year y_d and reached age x (i.e., living $x - \tau$ years after diagnosis and cohort forming)
$p(x)$	The density of age distribution in a standard year. The normalization rule for the density is $\int_{x_0}^{\infty} p(x) dx = 1$
$T \dots(y)$ and $\hat{T} \dots(y)$	Specific contributions to the time trends of age-adjusted disease prevalence and incidence-based mortality (discussed in detail in subsection 2.3)