Theoretical Population Biology 114 (2017) 117–127

Contents lists available at ScienceDirect

Theoretical Population Biology

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

Theory of partitioning of disease prevalence and mortality in observational data

I. Akushevich [a,*], A.P. Yashkin [a], J. Kravchenko [b], F. Fang [a], K. Arbeev [a], F. Sloan [c], A.I. Yashin [a]

[a] Biodemography of Aging Research Unit, Center for Population Health and Aging, Duke University, Durham, NC, United States

[b] Department of Surgery, Duke University School of Medicine, Durham, NC, United States

[c] Department of Economics, Duke University, Durham, NC, United States

ARTICLE INFO

Article history:
Received 23 July 2016
Available online 24 January 2017

Keywords:
Time trend
Partitioning
Incidence
Prevalence
Mortality
Diabetes

ABSTRACT

In this study, we present a new theory of partitioning of disease prevalence and incidence-based mortality and demonstrate how this theory practically works for analyses of Medicare data. In the theory, the prevalence of a disease and incidence-based mortality are modeled in terms of disease incidence and survival after diagnosis supplemented by information on disease prevalence at the initial age and year available in a dataset. Partitioning of the trends of prevalence and mortality is calculated with minimal assumptions. The resulting expressions for the components of the trends are given by continuous functions of data. The estimator is consistent and stable. The developed methodology is applied for data on type 2 diabetes using individual records from a nationally representative 5% sample of Medicare beneficiaries age 65+. Numerical estimates show excellent concordance between empirical estimates and theoretical predictions. Evaluated partitioning model showed that both prevalence and mortality increase with time. The primary driving factors of the observed prevalence increase are improved survival and increased prevalence at age 65. The increase in diabetes-related mortality is driven by increased prevalence and unobserved trends in time-periods and age-groups outside of the range of the data used in the study. Finally, the properties of the new estimator, possible statistical and systematical uncertainties, and future practical applications of this methodology in epidemiology, demography, public health and health forecasting are discussed.

1. Introduction

Prevalence is an epidemiologic characteristic which is easily measured using survey data or medical records. Analyses of prevalence trends play an influential role in health policy planning and are widely used to assess the extent to which a given health problem affects the population. However, conclusions about the relative success or failure of a health policy change cannot be made directly from trends of disease prevalence because temporal changes in age-adjusted prevalence rates are the result of two simultaneously occurring competing processes: (i) changes in incidence and (ii) changes in survival. Health interventions and disease treatment guidelines are usually aimed at decreasing the incidence and increasing the survival rate for a disease. If successful, these measures will push the observed prevalence in different directions. A related quantity of interest is the mortality rate by cause or more generally, the mortality for individuals after the onset of a specific disease. This is also known as the incidence-based mortality rate (Chu et al., 1994). The time trend of incidence-based mortality (Mozaffarian et al., 2016; Smith et al., 2013; Thun et al., 2013) is defined by the same factors that define the time trends in the disease prevalence rate, as well as trends in mortality in the general population. In contrast to disease prevalence, improvements in incidence and survival push the observed incidence-based mortality for a specific disease in the same direction, because improved incidence reduces the total number of people with the disease and improved survival further reduces the number of deaths associated with the disease.

In this paper, we develop a new methodological approach for the decomposition of trends in disease prevalence and incidence-based mortality into their constituent components (such as trends in incidence, survival, and prevalence prior to observation) and for

* Correspondence to: Biodemography of Aging Research Unit, Social Science Research Institute, Duke University, 2024 W. Main Str., United States.
E-mail address: igor.akushevich@duke.edu (I. Akushevich).

http://dx.doi.org/10.1016/j.tpb.2017.01.003
0040-5809/

the evaluation of the strength and the direction of the contribution of each respective component. The methodology described in this study offers a number of distinct strengths: (i) computation of disease prevalence and incidence-based mortality as well as their partitioning through a set of exact formulas without making simplifying assumptions, (ii) evaluation of the individual contributions of each component to the total time trend by direct calculation using exact formulas applied to real data, and (iii) a set of natural generalizations including applications to medical costs, complications of a specific disease, the incorporation of disease risk factors and the use of the historical trends of each of the model components beyond the region directly measured in data.

The only previously existing methodological approach of this type was developed by Tunstall-Pedoe for the partitioning of mortality trends through the use of an approximate formula for the simple decomposition of the annual percent change (APC) for mortality as a sum of APC's of cardiovascular disease incidence and case fatality (percentage of 28-day fatalities) (Tunstall-Pedoe et al., 1999). This approximation is valid only for events (disease onset and death) occurring within a short time of each other and requires that the APC be small and the disease of interest be the primary cause of death. Other methods of decomposition used in demography and epidemiology (see Canudas-Romo, 2003; Horiuchi et al., 2008; Vaupel and Romo, 2003 for a comprehensive review) are not related to the decomposition of prevalence into its constituent components.

Although the primary focus of this paper is to introduce the methodology and describe the mathematics involved in its execution, an example involving type 2 diabetes mellitus is also considered. The application of the methodology to disease prevalence and mortality is intended to address an aspect of a current Public Health problem—with some notable exceptions such as cardiovascular disease (Will et al., 2014), the prevalence rate of many chronic diseases including diabetes has been increasing with time (Akinbami et al., 2012; Bauer et al., 2014; Coresh et al., 2007; Egan et al., 2010). Understanding the contribution each individual component makes to the overall effect on disease prevalence and mortality and how these contributions have changed over time in response to changes in health policy, population age-structure and epidemiologic characteristics could be of great use in identifying likely targets for pro-active policy interventions.

2. Theory

2.1. Mathematical formalism

Data collected in an observational study represent information on eligible individuals over given periods of age and time. In this study, we use a nationally representative 5% sample of the US Medicare population provided for research as restricted access public use files by the Centers for Medicare and Medicaid Services. This database provides individual heath related information on US Medicare beneficiaries after age 65 from 1991 to 2013. The long time period and level of detail provided by such data allow us to calculate disease prevalence and mortality at any point after a certain look-back period (12 months is used in this study) necessary to collect individual information for evaluation of disease presence. Fig. 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. The bound is defined by an initial year (y_{00})

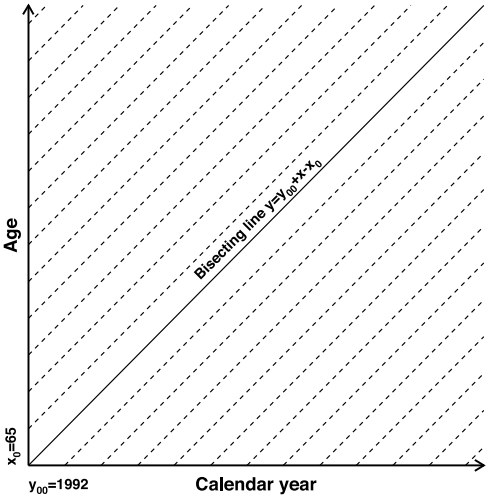

Fig. 1. The two dimensional diagram (Lexis diagram) to show the age–time area in which data are available and represent events (such as disease onset or deaths) that occur to individuals belonging to different cohorts. Calendar time is represented on the horizontal axis, while age is represented on the vertical axis. Dashed lines show time/age points for specific cohorts. Information about a cohort is available starting from bounding lines, i.e., either $y_{00} = 1992$ or $x_0 = 65$. Calculation of age-adjusted rates for a specific time requires integration over all ages starting from x_0, so regions both below and above bisecting line contribute to the integral for any $y > y_{00}$.

or minimal age (x_0) observed in the data. These two subareas are separated by the bisecting line defined as $y = y_{00} + x - x_0$. Above the bisecting line, the starting point is defined by the initial conditions $y = y_{00}$ with various ages while below the line the initial point is defined by boundary condition $x = x_0$ with various years. The cohort-specific bounding point is defined as $\bar{x}_0 = \max(x_0, y_{00} - y_b)$ and $\bar{y}_0 = y_b + \bar{x}_0$. Definitions of ages and times as well as functions of survival analyses used in the paper are collected in Table 1.

The idea for the representation of the formulas for prevalence is based on that the probability of being prevalent $P_c(x, y_b)$ at age x in cohort c with birth time y_b requires either

(i) being prevalent (represented by initial prevalence $P_c(\bar{x}_0, y_b)$) in the initial age $\bar{x}_0$ (and year $\bar{y}_0$) for the cohort and surviving to age x (represented by the survival probability $\bar{S}(x - \bar{x}_0, \bar{x}_0, \bar{y}_0)$ of a patient diagnosed no later than $\bar{x}_0$), or
(ii) being incident at an earlier age τ, $\bar{x}_0 < \tau \le x$ (represented by incidence density function $I_c(\tau, y_b)$) and having survival longer than $x - \tau$ (represented by survival probability $S(x - \tau, \tau, y_d)$ of a patient diagnosed at age τ and year y_d).

Therefore

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

where we integrate over all possible ages at diagnosis. Similarly, for mortality (we consider incidence-based mortality, i.e., mortality after disease onset) 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 $(\bar{x}_0, \bar{y}_0)$ for this cohort or being incident at an earlier age $x - \tau$. Death is represented by a respective density function $M_c(x, y_b)$ such that

$$M_c(x, y_b) = P_c(\bar{x}_0, y_b)\bar{f}_c(x - \bar{x}_0, \bar{x}_0, \bar{y}_0) + \int_{\bar{x}_0}^{x} I_c(\tau, y_b)f_c(x - \tau, \tau, y_d)d\tau. \quad (2)$$

The densities $\bar{f}_c(x - \bar{x}_0, \bar{x}_0, \bar{y}_0)$ and $f_c(x - \tau, \tau, y_d)$ in (2) are related to respective survival functions in (1): $\bar{f}_c() = -\bar{S}'_c()$ and $f_c() =$