Changing Patterns of Diabetes Prevalence in the Aging U.S. Population

I.Akushevich¹, A.Yashkin¹, J.Kravchenko², F. Fang¹, K. Arbeev¹, F.Sloan³, AI 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.
³Department of Economics, Duke University, Durham, NC, United States

Abstract

Diabetes mellitus, Type II (T2D), is a chronic disorder increasingly prevalent in the U.S. population and associated with serious complications affecting cardio-, cerebrovascular, and renal systems. This study uses Medicare data i) to identify age patterns and time trends of prevalence of diabetes and major complications and ii) to develop a methodology for the partitioning of the observed trends of diabetes prevalence into their main contributing components—trends of diabetes incidence, prevalence at 65 years, and patients survival. Specific findings of empirical analyses included: i) the age at which the prevalence of diabetes reaches its maximum is 75 years old and ii) prevalence of diabetes increases over time. Application of the partitioning approach to T2D prevalence showed that the prevalence of T2D at age 65 (approximately 55% contribution to overall prevalence trend) and survival among individuals with T2D (65% contribution) has increased while the post-65 incidence of T2D (-20% contribution) has decreased over the study time period. This approach can be generalized for additional applications to cause-specific death and for inclusion of recovery/long-term remission into the analysis.

1. Introduction

The growing proportion of older adults in the U.S. makes answering the questions of how to improve the provision of health services, individual patient outcomes, and healthy lifespan a problem of the highest importance. Although some substantial successes such as improvements in cardio- and cerebrovascular survival have been observed (Tunstall-Pedoe, et al., 1999), the effectiveness of health services for other chronic diseases and the overall length of life in the U.S. is in need of further improvements. Diabetes Mellitus, Type II (T2D), is a chronic disorder increasingly prevalent in the U.S. elderly population. T2D is associated with serious complications of the cardiovascular, cerebrovascular, renal and ocular systems (Akushevich et al., 2013a, Yashkin et al., 2015) as well as Alzheimer's disease (Haan, 2006, Myint et al., 2013) and a number of cancers (Larsson et al, 2007, Shikata et al, 2013). Older adults, whose proportional share of the overall U.S. population structure has been increasing, are at special risk. This study used two nationally representative administrative datasets to evaluate and model age patterns and time trends of diabetes prevalence, incidence, survival and the rate of long-term remission after diabetes onset with a high level of precision. Based on the results of the models, a partitioning of the current trends in prevalence in terms of the time trends of its components (incidence, prevalence at 65 years, and survival) was performed. This allowed for the identification of the macro-level determinants of previously observed and current trends in T2D among U.S. adults aged 65+, as well as to create an algorithm aimed at increasing the quality of predictive modeling for T2D. This algorithm can serve as an additional tool for the improvement in the quality of health policy planning and more efficient use of the current funds available to the Medicare program, thus potentially increasing the cost effectiveness and by extension the durability of the program.

The paper is organized as follows. Data are briefly described in Section 2. Methods of dealing with Medicare claims data are discussed in Section 3. Empiric evaluations of prevalence of T2D and several other diseases are given in Section 4. Section 5 is the central to our paper. It is devoted to the development and application of the methodology for partitioning of trends in disease prevalence. This section contains several subsections (5.1-5.3) to describe the procedure of parameter estimates of specific submodels. Numeric results for T2D are given in another subsection 5.4. The formalism and technical details on the derivation of the formula for
partitioning prevalence trends are given in the Appendix. Discussion and conclusion are presented in Section 6.

2. Data

Two nationally representative Medicare datasets will be used in this study – the 5% Medicare Claims Dataset (1991-2012), and the SEER-Medicare (1991-2005). Both Medicare-based datasets provide information on services paid by either Medicare Part A (facility-based services, e.g., in hospitals) or Medicare Part B (professional services, e.g., in physician practices) as well as demographic and enrollment information for the beneficiaries.

The expanded SEER registry covers approximately 26% of the U.S. population. The SEER-Medicare includes data on two groups of individuals: those diagnosed with skin melanoma (n=101,123), breast (n=353,285), colon (n=222,659), lung (n=342,961) and prostate (n=448,410) cancers, and a random 5% sample of Medicare beneficiaries residing in SEER areas who had none of the above-mentioned diseases. The total SEER-Medicare sample consisted of 2,154,598 individuals. The SEER-Medicare data is representative of the population of SEER areas only; therefore, SEER represents the U.S. general population only approximately. Even so, the age and sex distribution of the total SEER population is similar to non-SEER areas, though SEER areas have less whites, more urban residents, and less areas with low socio-economic status compared to non-SEER areas (Warren, et al., 2002).

5%-Medicare Claims Dataset is a nationally representative database of a 5% sample of the total US Medicare population, and is provided by the US Centers for Medicare and Medicaid Services (CMS) as a restricted-access public use file. Although limited in the amount of demographic information it provides (sex, race, age, and zip code of residence), this dataset provides generalizability at the national level, large sample size of approximately 5 million patients, relatively long period of observation (1991 to 2012), and the ability to identify trends even for less frequent diseases. 5%-Medicare data will be used for the partitioning analysis of diabetes prevalence.

3. Reconstruction of individual medical histories

Reconstructing the history of a disease from onset to recovery/long-term remission and/or death from Medicare records requires the creation of specific criteria indicating that an individual has a disease at any time \( t \) of his/her follow-up. Figure 1 illustrates our approach for reconstructing individual medical histories for each studied disease from the Medicare files combining all records with their respective ICD-9 codes. First, the optimal disease-specific look-back periods \( \tau_j \) need to be evaluated. Second, since an individual history contains up to seven (4 base - inpatient (IP), outpatient (OP), skilled nursing facility (SNF), and physician (PHY); 3 supplementary - hospice (HOS), home health agency (HHA) and durable medical equipment (DME)) sources of Medicare information, a decision needs to be made as to which sources are to be used. Third, since each diagnosis in the Medicare records can be either primary or secondary, a decision on the use of secondary diagnosis also has to be made. In line with prior research (Akushevich et. al. 2012), this study we will use both primary and secondary diagnoses drawn from the four base sources of Medicare information. A number of alternative approaches will be used for sensitivity analysis.

Figure 1. Individual disease trajectories based on Medicare records.

Figure 2. Prevalence rates for T2D and related conditions using SEER-Medicare data.
4. Prevalence: empiric evaluation

The age patterns of the prevalence probabilities of T2D and associated diseases using a 6-month look-back period, (Figure 2) were similar, showing an initial increase, peaking at age(s) 75-90, followed by a rapid decline. Month-specific time patterns of the prevalence probabilities (Figure 3) show a more varied picture: the rates for many chronic diseases are increasing, however some acute health conditions (e.g. Myocardial Infarction) known to lead to marked decreases in individual health status, have been going down. Surprisingly, many of the studied diseases show strong, systematic seasonal fluctuations, although expected in some diseases (e.g. COPD, pneumonia) this effect was also observed for many conditions which, logically, are not expected to exhibit a seasonal pattern (e.g. myocardial infarction, heart failure, or pancreatic cancer). On further investigation of this phenomenon it was found that i) the pattern is relatively stable among different geographic areas and different definitions of prevalence, ii) the amplitude of fluctuations is less pronounced in areas with hot-dry and marine climate than in areas with cold, mix-humid, and hot-humid climate, and iii) the prevalence probability for the majority of chronic diseases correlates with annual patterns in the frequency of doctor visits. Figure 4 shows the age adjusted prevalence of T2D using both 5%-Medicare (solid lines) and SEER-Medicare (dashed lines) data. To illustrate the sensitivity of the process to changes in the settings of the medical history algorithm, we also show the age-adjusted prevalence of T2D using a 12-month look-back period (red curves).

5. Partitioning analysis

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 rate in different directions. Therefore it is vitally important to know the contribution of each to the overall effect on disease prevalence. In this paper we propose an approach that allows for the partitioning of the time trends of disease prevalence into their constituent components (e.g., trends in incidence, survival, and prevalence of diabetes before 65). A similar approach was applied to the decomposition of mortality trends using an approximated 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 (Tunstall-Pedoe et al., 1999). The approximation is reasonable only for events (disease onset and death) occurring within a short period of time and requires that the APC be small and the disease of interest be the primary cause of death. In this paper we intend to overcome these limitations, develop a new approach for estimating partitioning of trends, and apply the updated method to the analyses of prevalence of T2D.

The idea of our approach is based on the explicit representation of prevalence at ages 65+ with no simplifying assumptions (equation 1). Details on the derivation of the formulas for the decomposition of the time trends are presented in Appendix. The resulted formula for age adjusted prevalence for ages 65 and above is:
\[
P(y) = \int_{-\infty}^{\infty} \left( P(x_0, y_0) S'(x-x_0, x_0, y_0) + \int_{x_0}^{x} I(\tau, y_d) S'_d(x-\tau, \tau, y_d) d\tau \right) p(x)dx
\]

where \( P(y) \) is the age-adjusted prevalence of a disease at time \( y \), \( x \) is age, \( p(x) \) is the density of age distribution in a standard year, \( y_d = y - x + \tau \) is time of T2D diagnosis, \( y_0 = y - x + x_0 \) is the year when an individual reached age 65 (i.e., initial age \( x_0 = 65 \)). Four functions in (1) are i) \( P(x_0, y_0) \) is prevalence at the beginning of an observation, ii) \( S'(x-x_0, x_0, y_0) \) is the relative survival probability of individuals who had diabetes at 65, iii) \( I(\tau, y_d) \) is the incidence rate at age \( \tau \), and iv) \( S'_d(x-\tau, \tau, y_d) \) is the relative survival probability of individuals diagnosed at \( y_d \). These four functions are terms representing the potential sources of the time trend in \( P(y) \). Formally, the measure of the time trend is the slope defined as \( b(y) = \frac{1}{P(y)} \frac{dP(y)}{dy} \). In the Appendix we demonstrated that the slope can be represented in terms of four respective terms as \( b(y) = T_{p_0}(y) + T_S(y) + T_{inc}(y) + T_S(y) \) explicitly representing the time trends as a sum of specific terms due to trends in initial prevalence (i.e., prevalence at \( x_0 \) \( (T_{p_0}) \), incidence rates \( T_{inc} \), and relative survivals after diabetes onset \( (T_S) \) and in patients with the disease at \( x_0 \) \( (T_S) \). Explicit expressions for these terms are given in eq. (A.11). Thus, to complete the analysis we need to create and estimate the following models: i) model for prevalence at \( x_0 \) (i.e., at the beginning age of observation), ii) model for probability of relative survival of prevalent individuals at \( x_0 \), iii) model for incidence rate, and iv) model for probability of relative survival after cancer diagnosis. Each model needs to reconstruct the two-dimensional functions for each respective measure. We use the Weibull model for relative survival and an empirically-based model for incidence rates. Time dependence in these models is represented by the linear model, i.e., the coefficient determining age and survival time dependences linearly depend on time. Linear model for prevalence at 65 contradicts empirically observed patterns, and therefore, this dependence is modeled by B-splines.

5.1. Model for prevalence at \( x_0 \)

We estimate prevalence at 65 for each year and then apply B-splines to fit its time pattern. An important feature of B-splines necessary for our study is that they allow for the calculation of derivatives explicitly and without additional simplifying assumptions. Specifically, \( y_0 \)-dependence of model parameters are modeled as

\[
P(x_0, y_0) = \sum_i \alpha_i B_{i,n}(y_0),
\]

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. The first derivative of \( P(x_0, y_0) \) can then be explicitly calculated because \( B'_{i,n}(y) \) is represented in terms of B-splines of a lower degree. Note also that the approach gives the derivative of \( P(x_0, y_0) \) with respect to \( y_0 \), however since \( y_0 = y - x + x_0 \) it is equal to the derivative with respect to \( y \):

\[
\frac{dP(x_0, y_0)}{dy_0} = \frac{dP(x_0, y_0)}{dy}.
\]

The results of modeling the empiric rates are presented in Figure 5.

5.2. Model for incidence rate

Approaches to reconstructing incidence rates using Medicare data were developed and applied for age-patterns (Akushevich et al., 2012, 2013b), time trends (Akushevich et al. 2013c), and interdependence among disease risks using two Medicare-based datasets: SEER-Medicare and NLTCS-Medicare (Akushevich et al., 2013a). In this study we updated our prior analyses of age-patterns and time trends of incidence rates to evaluate two-dimensional (i.e., over age and

![Figure 5. Prevalence at 65 of diabetes: empiric estimates (dots) and B-spline model (solid line)
time) functions of incidence rates for T2D using 5%-Medicare data. Analysis of the T2D incidence rates calculated empirically using 5%-Medicare data (Figure 6) as well as results of earlier studies (Akushevich et al., 2012, 2013), shows that linear model is applicable in the region of a ges 70-100, i.e.,

$$ I(y_a, y_d) = (\tau - 70)(a_0 + a_1 y_a) + b_0 + b_1 y_d $$

In this model, the parameters describing linear dependence of the incidence rate on age are in turn linearly dependent on the age at diagnosis. Derivative with respect to $y$ are calculated explicitly (recall, $y_d = y - x + \tau$):

$$ I'(y_a, y_d) = (\tau - 70)a_1 + b_1. $$

The model parameters were estimated as $b_0=0.019$, $b_1=-0.00013$ year$^{-1}$, $a_0=-0.00011$ year$^{-1}$, $a_1=1.4E-6$ year$^{-2}$. These estimates show that interaction between age and time is not significant and that incidence decreases with time. However, the detected decrease is due to years close to the boundary of the observed period: the incidence rate increases with time in time period from 1995 to 2008. How this uncertainty influences the final estimates of the partitioning of T2D prevalence is the subject of a future sensitivity study. One way to avoid this uncertainty is to estimate model parameters for each year and to apply B-spline models to model time patterns of estimated parameters. Also note the occurrence of peaks in incidence rates in 2004 and 2008. Future studies will need to explain this effect.

5.3.Models for relative survival

The parameters for the relative survival model for a specific year are estimated using maximizing likelihood. Using the standard likelihood for total survival $L = \prod_i h_{it}^d S(x_{it}, x_i)$ and the definitions of relative survival, $S(x_{it}, x_i) = S_{pop}(x_{it}, x_i)S_{rel}(x_{it}, x_i)$, or for respective hazard functions $h(x) = h_{pop}(x) + h_{rel}(x, \beta)$, we construct the log likelihood as (Dickman et al., 2004):

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

Here $d_i$ is the death indicator and $\beta$ is the set of parameters of interest. The first term does not depend on $\beta$ and therefore can be omitted. The only thing we need to know about the general population is the population hazards at the age of death for all individuals in datasets. 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 $S_{rel} = \exp(-\alpha x^{-1}_t)$ and $h_{rel} = \gamma x^{-1}_t$, $(x_t = x - \tau$ is the survival time) where $\gamma$ is the Weibull shape parameter, and $\alpha = \exp(-\gamma \mu) = \lambda^{-1}$ (i.e., $\lambda = \exp(\mu)$) and $\lambda$ is the Weibull scale parameter. Parameter estimates in age groups show that parameters can have quadratic dependence on age. Only two parameters among $\alpha$, $\gamma$, $\mu$, and $\sigma = 1/\gamma$ are independent and the likelihood ratio test shows that the largest likelihood is for $\sigma$ and $\mu$. The survival function in these parameters is: $S_{rel} = \exp(-\exp(\sigma^{-1}(\log x_t - \mu)))$.

Thus the effects of age (or age at diagnosis) are modeled as

$$ \mu = \mu_0(1 + \mu_0) + \mu_1(1 + \mu_1) + \mu_2(1 + \mu_2), \quad \sigma = \sigma_0(1 + \sigma_0) + \sigma_1(1 + \sigma_1) + \sigma_2(1 + \sigma_2) $$
Thus \( \beta = \{ \mu_{0-2}, \mu_{0-2}, \sigma_{0-2}, \sigma_{10-2} \} \).

The first derivative of \( S_{rel} \) with respect to \( y \) is (subscript ‘rel’ is dropped) defined by the expressions:

\[
S'_y(x, y) = S'_\mu(x, y) \mu' + S'_\sigma(x, y) \sigma',
\]

\[
\mu'_y = \mu_0 \mu_1 + \tau \mu_2, \quad \sigma'_y = \sigma_0 \sigma_1 + \tau \sigma_2 \sigma_1 + \tau^2 \sigma_2 \sigma_2,
\]

\[
S'_\mu(x, y) = \gamma \alpha x^\gamma S_{rel}(x, y),
\]

\[
S'_\sigma(x, y) = (\log(y) - \mu) \gamma \alpha x^\gamma S_{rel}(x, y).
\]

The parameters \( \beta = \{ \mu_{0-2}, \mu_{0-2}, \sigma_{0-2}, \sigma_{10-2} \} \) are estimated separately for the models of relative survival after disease onset \( (S'_y(x - \tau, y)) \) and of relative survival of prevalent individuals at \( x_0 \) (i.e., for \( \tilde{S}'(x - x_0, y_0) \)).

Time and age patterns of relative survival functions are presented in Figure 7. Note that empiric estimates of relative survival are not necessary for our modeling so they are not presented in Figure 7.

5.5. Estimates of partitioning in time trend of diabetes prevalence

Application of estimated models to the 5%-Medicare data resulted in predicted prevalence according eq. (1). Its pattern is presented in Figure 8 in comparison with the pattern empirically estimated using 5%-Medicare (green curve) and SEER-Medicare (red curve) for the same data for 12-month look-back period (both green and red curves are presented in Figure 4). Small difference between our model and empiric evaluation is because all individuals with diabetes diagnosis contribute to the curve representing our model, and only individuals with actual visits during last 12 months give a contribution to the empirical points given by red and green curves. Right plot presents the decomposition (i.e., partitioning) of the trends according its contribution, i.e., in accordance with the formula

\[
b(y) = T_{pop}(y) + T_{S}(y) + T_{inc}(y) + T_{S}(y). \]

We see that prevalence increases with time because of i) increasing prevalence at 65, ii) improving survival at 65, iii) improving survival after disease onset at \( \tau \geq 66 \) years, and iv) decreased incidence. The contributions are mentioned in the order of decreasing sizes of their contributions to the total trend of disease prevalence. These contributions can be evaluated as percent’s of total time trends. For example, time trend in 2000 (i.e., \( b(2000) \)) is explained by time trend in prevalence rates at 65 (53.4%), and by improved survival for individuals diagnosed after (45.4%) and before (13.0%) age 65. All these contributions increased the T2D prevalence. Sum of these contributions exceed 100% because of last contribution, diabetes incidence, whose contribution is negative (-11.8%). The contributions for other years are similar, e.g., for 2010 they are 56.8% (prevalence at 65), 43.9% (survival), 26.2% (survival at 65), and -25.9% (incidence). As we see, prevalence increases because of improved survival and decreases because of decreased incidence. These tendencies reflect improvement of health care both at the level of preventive medicine and at the level of actual treatment. Increased fraction of prevalence at 65 allows us to hypothesize
that a critical moment for T2D prevalence occurs before age 65. However, these results could be also explained by reduction in the average severity of T2D over time.

6. Discussion and Conclusion.

The approaches developed and used in this study provide high precision information about the time-trends of the health related characteristics of T2D and associated complications that are used in demography, epidemiology, and public health. These approaches have the potential to overcome the shortcomings of traditional forecasting methods through the efficient use of informative large Medicare-based datasets, as well as innovative approaches that benefit from: quantifying two-dimensional patterns of prevalence and incidence, including survival and long-term remission after disease onset, and taking possible dependencies among age-related chronic conditions into account. The extrapolations of the time trends identified in our study provide a simple forecasting model. The primary focus of our study, however, is the partitioning of the identified current trends in disease-specific mortality and prevalence into the time trends of their components (incidence, prevalence at 65 years, recovery, and case fatality/survival). This idea is inspired by the body of work centered on the MONICA study and attempting to explain the reasons for a decrease in mortality from cardiovascular disease identified in the 1970s U.S. and the seminal paper of Tunstall-Pedoe et al. (1999) who used an approximate formula for simple decomposition of the annual percent change (APC) for mortality as a sum of APCs of incidence and case fatality. The initial Tunstall-Pedoe partitioning approach has limitations: it is valid only for events occurring within a short time period of each other; it requires that the APC be small, and the disease of interest must be the primary cause of death. In contrast, our study was able to perform the partitioning with greater detail focusing on prevalence of T2D (and potentially on a broader spectrum of diseases, e.g., diabetes complications), and evaluate the partitioning with much better accuracy. These aspects represent major avenues for developing more accurate forecasts of the future health of the U.S. elderly population. Incorporation of all these aspects into analyses of time trends and further forecasting models allowed us to essentially decrease the stochastic components of forecasts by integrating estimated effects of health and survival with known mechanisms of aging related changes on health and survival.

Our extension of the Tunstall-Pedoe partitioning approach and its practical application is made possible by the availability of large longitudinal datasets (e.g., 5%-Medicare) which allow for the evaluation of the prevalence components for over long periods of time (currently up to 23 years are available). The size of the data as well as the need for explicit calculation (as opposed to estimation based on a series of assumptions) require several methodological innovations allowing for the evaluation of disease prevalence and its time trends in terms of the first derivative of the prevalence probability over calendar time. Such methodological innovations include using the available parametric representation of disease incidence and survival and using B-splines that can evaluate the parameters responsible for time evolution of incidence and survival and, most importantly, provide an analytic expression for the first derivative needed for time trend estimates. Obtained formula for the time trend of disease prevalence is decomposed over the contributions of incidence, survival, and prevalence at initial age of observation (i.e., 65) allowing us to evaluate their contribution to the total time trend (i.e., perform partitioning analysis) for all years involved in analysis.

The approach developed in this paper provides a predictive model for disease prevalence over age and time. These predictions can be compared to usual approaches designed for evaluating prevalence from observational data. One such an approach uses 5%-Medicare data and evaluates the prevalence as a fraction of individuals having records with a disease code during a certain period of time before current date. Such analyses require a certain definitions of individual disease presence, and therefore the exact formalism presented in this paper allows us to judge which formalism for evaluating prevalence rates from administrative data (e.g., Medicare data) would be optimal. In this paper we compare these two approaches and discuss how the approaches based on Medicare data can be tuned to better represent the disease prevalence.

In this study we focused on the partitioning of disease prevalence. The explicit formula (1) allows us to determine the origin of the trend in prevalence probability and evaluate the contributions from the components including prevalence at 65, incidence rate, and survival after diabetes onset. One generalization involves incorporating the phenomenon of recovery and its effect of survival (Yashin et al, 2010, Akushevich et al., 2013d). Two conclusions about recovery after T2D diagnosis were made: i) recovered individuals have lower death rates than non-recovered patients, and therefore, patients who stopped visiting doctors are a healthier subcohort and ii) recovered individuals have higher death rates than in the general population for all considered diseases, and therefore, the complete recovery does not occur. The model (1) can be naturally updated by involving recovery (or long-term remission) from disease:
\[ P(x, y) = P(x_0, y_0)S^\tau(x - x_0, x_0, y_0)S^\nu (x - x_0, y_0) + \int_{x_0}^x I(\tau, y_\nu)S^\nu (x - \tau, \tau, y_\nu)S^\nu (x - \tau, \tau, y_\nu)d\tau. \] (A.6)

where \( S^\nu \) denotes not-yet-recovery probability.

We used our definition of incidence rates and considered alternative definitions in sensitivity studies. Different definitions of incidence rate in use in clinical practice (Akushevich et al., 2012) (e.g. inpatient incidence, total incidence) could result in rates differing by several times (up to 10 and more) (Akushevich et al, 2013a,e). Our definition of incidence rates was chosen in such a way as to provide rates that would be maximally close to empirical studies in literature (examples of such comparison of different definitions can be found in Akushevich et al., 2012, 2013a,e).

Finally, we note that a similar approach could be applied for partitioning time trends of mortality by cause (or in the other words by incidence-based mortality rate, Chu et al., 1994) using the same methods. That is, we can model the mortality by cause in the style of eq. (1): the probability of dying in the age interval requires being incident at an earlier age \( x - \tau \) and having death survival in the interval \((x, x + dx)\) thus resulting in the formula similar to our starting formula (A.1):

\[ M_d(x) = \int_0^x I(u) p_d(x - u)du, \]

where \( p_d \) is the p.d.f. of disease survival. In this equation we used cohort functions and kept only the first argument of the functions. Similar formulas could be obtained for period functions following the formalism described in Appendix.

Our analysis shows that although the prevalence of T2D continues to grow in the U.S., its increase became less prominent during last 10 years. This is in agreement with recent reports on T2D trends of prevalence and incidence (Geiss et al, 2014). While the role of incidence and survival in observed trends of T2D prevalence have been widely studied (Lipscombe, Hux, 2007 Carstensen et al, 2008; Menke et al, 2015 ), we have conducted the first quantitative investigation of the structure of the components contributing to the prevalence trend using the partitioning analysis approach. Specifically, we demonstrated that the time trend of T2D prevalence is due to contribution of the following factors: i) approximately 10-25% of prevalence reduction is due to the changes in the post-65 incidence of T2D; this percentage has been increasing over the recent decade, ii) about 45% of the prevalence increase is due to survival of patients who have been diagnosed at age 65 and older; the contribution of this component remained relatively stable over time, iii) approximately 15-25% of prevalence increase is due to the role of the survival of all patients with T2D including those who were diagnosed at age younger than 65; and, finally, iv) almost 50% of the increase to the total prevalence is from the prevalence of T2D at age 65; the role of this component has been slightly decreasing with time but still remains the leading component contributing to the overall T2D prevalence. Therefore, it could be speculated that more attention to improvement in survival of patients with T2D who were diagnosed at an age younger than 65 and programs that target the prevention of diabetic complications at this age group may help to improve the overall disease burden more effectively. Further analysis of contributing components that help to explain the dynamics of mortality and survival of patients with T2D, as well as partitioning analysis for the components of incidence rates can provide detailed information for public health specialist that can be used in optimization of existing screening programs and guidelines for T2D disease management.

Appendix

In this Appendix we demonstrate how the formulae for partitioning can be derived.

Analysis and modeling of the prevalence of diabetes requires information about incidence and survival. Rigorously, the probability of being prevalent \( P_c(x) \) at age \( x \) requires being incident at an earlier age \( \tau, \tau \leq x \) (represented by incidence rate \( I_c(\tau) \)) and having survival larger than \( x - \tau \) (represented by survival probability \( S(x - \tau, \tau) \) of patient diagnosed a age \( \tau \)). Thus

\[ P_c(x) = \int_0^x I_c(\tau)S(x - \tau, \tau)d\tau \] (A.1)

where we integrate over all possible ages of incidence. The eq. (A.1) is valid for cohort functions i.e., when the denominator is the number of individuals at the time of the cohort forming. This formula can be understood in
terms of the numbers of individuals: if $N_0$ is the size of a birth cohort and $N_d(\tau) \Delta \tau$ is the number of individuals with disease onset at age period $\Delta \tau$. 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 $\tau$ over all age periods, i.e.,

$$N_d(x) = \sum_{n} N_d(\tau_n)(\Delta \tau_n)S(x-\tau_n, \tau_n)$$ (A.2)

where $S(x-\tau_n, \tau_n)$ is the survival function of individuals diagnosed at age period $\Delta \tau_n$ who survived to age $x$. Formula (A.1) is obtained when we consider infinitely small age periods (i.e., $\Delta \tau \to 0$) and define $P_c(x) = N_d(x) / N_0$ and $I_c(\tau) = N_d(\tau) / N_0$.

Usually we start to observe individuals at a specific age (e.g. 65) after which we begin to observe diseases onset. This is especially applicable to Medicare claims data. In this case, similar arguments (the number of individuals sick at age $x$ consists of two groups of survivors: those who had a diseases at $x_0$ and those who was diagnosed at $x > x_0$) result in the expression:

$$P_c(x) = P_c(x_0)S(x-x_0, x_0) + \int_{x_0}^{x} I_c(\tau)S(x-\tau, x)d\tau$$ (A.3)

where $S(x-x_0, x_0)$ is the survival probability of individuals who have the disease at age $x_0$ independent of age of diagnosis $\tau$. The eq. (A.3) can be obtained by representing eq. (A.1) in the form

$$P_c(x) = \int_{0}^{\infty} I_c(\tau)S(x-\tau, x)d\tau + \int_{x_0}^{\infty} I_c(\tau)S(x-\tau, x)d\tau = P_c(x_0)S(x-x_0, x_0) + \int_{x_0}^{\infty} I_c(\tau)S(x-\tau, x)d\tau$$

Exactly the definition of $S(x-x_0, x_0)$ is:

$$S(x-x_0, x_0) = \frac{\int_{x_0}^{\infty} I_c(\tau)S(x-\tau, x)d\tau}{\int_{0}^{\infty} I_c(\tau)S(x_0-\tau, x_0)d\tau}.$$

Incidence and survival after disease onset can be cohort dependent (i.e., dependent on the birth year of a cohort $y_b$). In this case, we use

$$P_c(x, y_b) = P_c(x_0, y_b)S(x-x_0, x_0, y_b) + \int_{x_0}^{\infty} I_c(\tau, y_b)S_c(x-\tau, x, y_b)d\tau$$ (A.4)

The exact definition of $P_c(x, y_b)$ is the fraction of individuals living with the disease aged $x$ and born in year $y_b$ to the total number of individuals born in year $y_b$. Similarly $I_c(\tau, y_b)$ is the cohort incidence defined as 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 calendar year (denoted as $y = y_b + x$) or calendar year of diagnosis (denoted as $y_d = y_b + \tau$). Therefore, we avoid dealing with cohort prevalence and incidence and use their definitions involving the population at risk rather than birth cohort size. Within these definitions the cohort prevalence and incidence are expressed through accepted definitions of prevalence and incidence: $P_c(x, y_b) = P(x, y)S_c(x, y_b)$ and $I_c(x, y_b) = I(x, y_d)S_c(x, y_b)$, where $S_c(x, y_b)$ is the survival function of the cohort born at year $y_b$. The eq. (A.4) within these definitions are rewritten as ( $y_0$ is the year when individuals born at $y_b$ reach age $x_0$ ),

$$P(x, y) = P(x_0, y_0)\frac{S(x_0, y_b)S(x-x_0, x_0, y_b)}{S_c(x, y_b)} + \int_{x_0}^{\infty} I(\tau, y_d)\frac{S_c(\tau, y_b)S(x-\tau, x, y_b)}{S_c(x, y_b)}d\tau$$ (A.5)

$$= P(x_0, y_0)S'(x-x_0, x_0, y_b) + \int_{x_0}^{\infty} I(\tau, y_d)S'(x-\tau, x, y_d)d\tau.$$
Since \( S^{-1}_t(x, y_b) S_t(r, y_b) \) is the reciprocal of survival probability at \( x \) for the cohort of patients formed at \( r \) (and similarly for \( x_0 \)), the two factors containing all three survival functions in (A.5) are the relative survivals for patients with disease at \( x_0 \) and diagnosed at \( r \) respectively:

\[
\tilde{S}'(x - x_0, x_0, y_0) = \frac{S_t(x_0, y_b)}{S_t(x, y_b)} \tilde{S}(x_0 - r, r, y_b); \quad \tilde{S}'(x - r, r, y_d) = \frac{S_t(r, y_b)}{S_t(x, y_b)} \tilde{S}(x - r, r, y_b)
\] (A.6)

Note, we can use \( y_0 \) (or \( y_d \)) instead of \( y_b \) in the third argument of \( S' \) because relative survival probability does not change \( (x, \tau, \tau, y_b) \rightarrow (x, \tau, \tau, y_0 + \tau) \) : in both case this is the survival probability that is formed at a certain time. This time can be fixed by age at diagnosis \( \tau \) and additionally, by i) birth cohort \( y_b \) or alternatively by calendar year at diagnosis \( y_d = y_b + \tau \).

The quantity of interest is the time trend of age adjusted prevalence (over the age region \((x_0, x_{max})\)) and its partitioning. Age-adjusted prevalence based on (A.5) is:

\[
P(y) = \int_{x_0}^{x_{max}} \left( P(x_0, y_0) \tilde{S}'(x - x_0, x_0, y_0) + \int_{x_0}^{x} I(\tau, y_d) S'_d(x - \tau, \tau, y_d) d\tau \right) p(x) dx
\] (A.7)

where \( p(x) \) is the density of age distribution in a standard year. Recall, \( y_d \) and \( y_0 \) are functions of \( y \) and integration variables: \( y_d = y - x + \tau \) and \( y_0 = y - x + x_0 \). We see that time trend in \( P(y) \) can be due to trends in initial prevalence (i.e., prevalence at \( x_0 \)), incidence rates, relative survival after diabetes onset and in patients with the disease at \( x_0 \). Formally the time trend is defined as,

\[
\frac{1}{P(y)} \frac{dP(y)}{dy}
\] (A.8)

Explicit calculation of this derivative results in

\[
\frac{1}{P(y)} \frac{dP(y)}{dy} = \int_{x_0}^{x_{max}} \left( \frac{dP(x_0, y_0)}{dy} \tilde{S}'(x - x_0, x_0, y_0) + P(x_0, y_0) \frac{d\tilde{S}'(x - x_0, x_0, y_0)}{dy} \right.
\]

\[
+ \left. \int_{x_0}^{x} \left( \frac{dI(\tau, y_d)}{dy} S'_d(x - \tau, \tau, y_d) + I(\tau, y_d) \frac{dS'_d(x - \tau, \tau, y_d)}{dy} \right) d\tau \right) p(x) dx
\] (A.9)

Thus, the time trend is determined through four terms in brackets of eq. (A.9) representing the partitioning of time trends for prevalence rates:

\[
\frac{1}{P(y)} \frac{dP(y)}{dy} = T_{p0}(y) + T_{S'}(y) + T_{inc}(y) + T_{inc}(y)
\] (A.10)

where

\[
T_{p0}(y) = \frac{1}{P(y)} \int_{x_0}^{x_{max}} \frac{dP(x_0, y_0)}{dy} \tilde{S}'(x - x_0, x_0, y_0) p(x) dx,
\]

\[
T_{S'}(y) = \frac{1}{P(y)} \int_{x_0}^{x} P(x_0, y_0) \frac{d\tilde{S}'(x - x_0, x_0, y_0)}{dy} p(x) dx,
\]

\[
T_{inc}(y) = \frac{1}{P(y)} \int_{x_0}^{x} \int_{x_0}^{x} \frac{dI(\tau, y_d)}{dy} S'_d(x - \tau, \tau, y_d) p(x) d\tau dx,
\]

\[
T_{inc}(y) = \frac{1}{P(y)} \int_{x_0}^{x} \int_{x_0}^{x} I(\tau, y_d) \frac{dS'_d(x - \tau, \tau, y_d)}{dy} p(x) d\tau dx.
\] (A.11)
Acknowledgement: This study has been supported by the National Institute on Aging (grants R01-AG017437, R01-AG046860, P01-AG043352).

References


