A note on the decomposition of number of life years lost according to causes of death

Per Kragh Andersen

Research Report 12/02

Department of Biostatistics
University of Copenhagen
A note on the decomposition of number of life years lost according to causes of death

Per Kragh Andersen, Department of Biostatistics, University of Copenhagen, Ø. Farimagsgade 5 PB2099, DK-1014 Copenhagen K, Denmark.
pka@biostat.ku.dk

Abstract

The standard competing risks model is studied and we show that the cause $j$ cumulative incidence function integrated from 0 to $\tau$ has a natural interpretation as the expected number of life years lost due to cause $j$ before time $\tau$. This is analogous to the $\tau$-restricted mean life time which is the survival function integrated from 0 to $\tau$. The large sample properties of a non-parametric estimator are outlined, and the method is exemplified using a standard data set on survival with malignant melanoma. It is discussed how the number of years lost may be related to subject-specific explanatory variables in a regression model based on pseudo-observations. The method is contrasted to cause-specific measures of life years lost used in demography.

Key words: demography; life expectancy; Survival analysis.
Introduction

In demography the expected life time is a useful summary of the age-specific mortality experience (e.g., Preston et al., 2001). If $\alpha(t)$ is the age-specific mortality rate (hazard function) then

$$S(t) = \exp\left(-\int_0^t \alpha(u) du\right)$$

is the corresponding survival function, and the life expectancy at birth is

$$e(0, \infty) = \int_0^\infty S(t) dt.$$

In practice, $\infty$ is represented by the “maximally attainable life length”, say $\omega$, and age-specific mortality rates between ages 0 and $\omega$ are needed to estimate the expected life time.

When death can be caused by a number, $k$, of exclusive and exhaustive causes $j = 1, \ldots, k$, the survival function is

$$S(t) = \exp\left(-\sum_{j=1}^k \int_0^t \alpha_j(u) du\right)$$

where $\alpha_j(\cdot)$ is the cause-$j$-specific hazard function. Thereby, the life expectancy at birth can be written as

$$e(0, \infty) = \int_0^\infty S_1(t) \cdots S_k(t) dt$$

with $S_j(t) = \exp\left(-\int_0^t \alpha_j(u) du\right)$. So-called “cause-$j$-deleted life-tables” are
then defined via the survival function

\[ S(t)/S_j(t) = \exp\left(-\sum_{\ell \neq j} \int_0^t \alpha_\ell(u)du\right) \]

and under “independence” of causes of death, this is the survival function in a population where cause \( j \) is no longer operating and where (by “independence”) the rate of death from cause \( \ell \neq j \) is still given by the cause-specific hazard \( \alpha_\ell(t) \). Following those lines, “life expectancy without cause \( j \)”

\[ e_{-j}(0, \infty) = \int_0^\infty S(t)/S_j(t)dt \]

and number of years lost “due to cause \( j \)”

\[ e_{-j}(0, \infty) - e(0, \infty) \]

can be defined (e.g., Beltrán-Sánchez et al., 2008).

In biostatistics, the assumption of independent competing risks has to a large extent been abandoned by now because of the recognition of the fact that data from “this world” (where all \( k \) causes of death are operating) provide little information on how the mortality experience would be in a hypothetical world where cause \( j \) has been eliminated (e.g., Andersen and Keiding, 2012). It is therefore of interest to develop alternative measures of the impact on life expectancy of given causes of death which do not rely on such extrapolations. This is what we will do in this note and, as we shall see, a very simple approach to the question can be taken.
Restricted mean life time and number of life years lost

In biostatistics the expected life time can rarely be estimated reliably because of right-censoring and as an alternative summary of the mortality experience, the $\tau$-restricted mean life time is sometimes used (e.g., Irwin, 1949; Karrison, 1987; Andersen et al., 2004). This is defined as

$$ e(0, \tau) = \int_{0}^{\tau} S(t)dt, \quad (1) $$

thus for $\tau = \omega = \infty$, this is simply the expected life time at birth. The restricted mean life time is the expectation $E(T \wedge \tau)$ of the random variable $T \wedge \tau$ where $T$ is the (unrestricted) life time. Therefore, the difference

$$ L(0, \tau) = \tau - E(T \wedge \tau) \quad (2) $$

is the expected number of years lost before time $\tau$.

In a competing risks situation, the cause $j$ cumulative incidence is the probability

$$ F_j(t) = P(T \leq t, D = j) = \int_{0}^{t} S(u-)\alpha_j(u)du $$

of failing from cause $j$ before time $t$ (where the random variable $D$ denotes the cause of failure with possible values $j = 1, \ldots, k$). For all values of $t$ we have

$$ S(t) + \sum_{j=1}^{k} F_j(t) = 1 $$
and, therefore, the expected number of years lost before time $\tau$ can be written as a sum

$$L(0, \tau) = \tau - \int_0^\tau S(t)dt = \sum_{j=1}^k \int_0^\tau F_j(t)dt$$

with a term for each of the failure causes. We will now show that each term

$$L_j(0, \tau) = \int_0^\tau F_j(t)dt$$

has a quite simple interpretation.

The competing risks model can conveniently be represented by a multistate model for a stochastic process, $X(t)$, with states “0: alive” (the initial state) and “$j$: dead by cause $j$”, $j = 1, \ldots, k$ (the final, “absorbing” states) (e.g., Andersen et al., 1993, Section IV.4). The $0 \to j$ transition intensity is the cause-specific hazard $\alpha_j(t)$ and the state occupation probabilities are $P(X(t) = 0) = S(t)$ and $P(X(t) = j) = F_j(t), j = 1, \ldots, k$. The failure time is the (proper) random variable, $T = \inf_t\{X(t) \neq 0\}$, denoting the time of exit from the initial state 0 whereas the times $T(j) = \inf_t\{X(t) = j\}$ of entry into the final state $j = 1, \ldots, k$ are all improper random variables because $P(T(j) = \infty) = 1 - F_j(\infty) > 0$. However, for any $\tau$, the random variable $T(j) \land \tau$ is proper with an expected value given by

$$E(T(j) \land \tau) = \int_0^\tau tS(t)\alpha_j(t)dt + \tau(1 - F_j(\tau))$$
which by partial integration equals $\tau - \int_0^\tau F_j(t)\,dt$. This leads to the result

$$L_j(0, \tau) = \int_0^\tau F_j(t)\,dt = E(\tau - T_j \wedge \tau). \quad (3)$$

Thereby, the total number of years lost before time $\tau$ can be decomposed into a sum over failure causes $j$ of terms which can be interpreted as the number of years lost “due to that cause”.

Restricting attention to subjects still alive at, say $t_0$, is simply handled by considering the conditional survival function $S(t \mid t_0) = S(t)/S(t_0)$ and the conditional cumulative incidences $F_j(t \mid t_0) = F_j(t)/S(t_0)$ given survival till $t_0$. Thereby, for those still alive at time $t_0$, the expected life time before $\tau > t_0$

$$e(t_0, \tau) = \int_{t_0}^\tau S(t \mid t_0)\,dt$$

and the expected number of years lost before $\tau$ “due to cause $j$”

$$L_j(t_0, \tau) = \int_{t_0}^\tau F_j(t \mid t_0)\,dt$$

can be defined.

Graphically, $e(0, \tau)$ and $L_j(0, \tau), j = 1, \ldots, k$ can be represented, as follows. Suppose for simplicity that $k = 2$ causes are considered. The cumulative incidences may be presented on a “stacked” plot, i.e., $F_1(t)$ and $F_1(t) + F_2(t)$ are plotted against $t$ and on such a plot, $L_1(0, \tau)$ is the area under the lower curve between 0 and $\tau$, and $L_2(0, \tau)$ the area between the two curves between 0 and $\tau$. Finally, $e(0, \tau)$ is the area between the upper
curve and 1 between 0 and $\tau$ (see Figure 1 for an example).

Non-parametric estimation of the survival and cumulative incidence functions based on right-censored (and possibly, left-truncated) survival data is straightforward using the Kaplan-Meier and Aalen-Johansen estimators. Thereby, also the quantities $e(t_0, \tau)$ and $L_j(t_0, \tau)$ are easily estimable. For the former, asymptotic properties as the number, $n$ of subjects increases were presented by Andersen et al. (1993, Example IV.3.8). For the latter, asymptotic properties are outlined in the Appendix.

In the case of no censoring, i.e. all failures before time $\tau$ are observed, both the Kaplan-Meier and the Aalen-Johansen estimators are simple relative frequencies, thus in this case the estimators of (1) and (3) are simply

$$\hat{e}(0, \tau) = \frac{1}{n} \sum_i (T_i \wedge \tau)$$

and

$$\hat{L}_j(0, \tau) = \frac{1}{n} \sum_i I(D_i = j)(\tau - T_i \wedge \tau).$$

Here, $T_i$ is the time of failure for subject $i = 1, \ldots, n$ and $D_i$ is the cause of failure indicator for subject $i = 1, \ldots, n$ with possible values $j = 1, \ldots, k$.

If interest focuses on how the parameters $e(\cdot, \tau)$ and $L_j(\cdot, \tau)$ relate to individual explanatory variables (such as gender, ethnic groups, occupational groups etc.) then regression analysis is straightforward in the case of no censoring. In the more realistic case where some observations are right-censored (before time $\tau$), one possible approach would be to use so-called
pseudo-observations (e.g., Andersen et al., 2003, 2005). If \( \theta \) denotes one of these parameters and \( \hat{\theta} \) the estimator based on the full sample then the \( i \)th pseudo-observation for \( \theta \) is

\[
\hat{\theta}_i = n\hat{\theta} - (n - 1)\hat{\theta}^{(-i)}, \quad i = 1, \ldots, n. \tag{5}
\]

Here, \( \hat{\theta}^{(-i)} \) is the estimator applied to the sample (of size \( n - 1 \)) obtained by taking subject \( i \) out of the full sample. In this way the association between the parameter and explanatory variables, \( Z_i \) can be ascertained via generalized estimating equations with \( \hat{\theta}_i \) as the outcome variable (Andersen et al., 2003). The method is implemented as the function `pseudoyl` in the R-package `pseudo` available at CRAN.

**Example: Survival with malignant melanoma**

Andersen et al. (1993, Example I.3.1) presented a study of survival with skin cancer (malignant melanoma). Here, 205 patients had radical surgery at Odense University Hospital, Denmark between 1962 and 1977. At the closing date of the study, 57 patients had died from the disease (cause \( j = 1 \)), 14 had died from other causes (cause \( j = 2 \)) and 134 were still alive. Figure 1 shows the stacked plot of the cumulative incidences of dying from the two causes in question. For \( \tau = 10 \) years the restricted mean life time is \( \hat{c}(0, 10) = 7.510 \) (SD=0.245) while \( \hat{L}_1(0, 10) = 2.005 \) (SD=0.232) years were lost due to deaths from the disease and \( \hat{L}_2(0, 10) = 0.485 \) (SD=0.133) years were lost due to deaths from other causes. The corresponding areas on the stacked plot can
be seen in Figure 1.

In regression models using pseudo-observations it was found that men on average lost 1.21 (SD=0.488) more years than women because of the disease and 0.211 (SD=0.284) more years than women due to other causes. The estimated intercepts in these models (estimated numbers of years lost for women) were 1.538 for deaths from the disease and 0.404 for deaths from other causes. Since $\frac{79}{205}=38.5\%$ of the patients were men, these results are well in line with the overall numbers of years lost (since $1.538 + 0.385 \times 1.21 \approx 2.005$ and $0.404 + 0.385 \times 0.211 \approx 0.485$). The effect of tumor thickness on the number of years lost due to the disease was 0.361 years (SD=0.082) for each mm and the effect of age on the number of years lost due to other causes was 0.269 years (SD=0.082) for every 10 years of age.

**Discussion. Accounting for standard mortality**

In demography, number of life years lost is usually related to life expectancy based on standard mortality tables (e.g., Vaupel and Canudas-Romo, 2003, Aragón et al., 2008, Shkolnikov et al., 2011). Thus, for uncensored data the expected number of life years lost is defined as

$$\tilde{L}(0, \omega) = \frac{\sum_{i=1}^{n} D_i e^{*}(T_i, \omega)}{\sum_{i} D_i}$$

where $\sum D_i = n$ and $e^{*}(T_i, \omega)$ is the mean residual life time at the time (age) $T_i$ of death, calculated based on standard mortality rates. For cause
\( j = 1, \ldots, k \) the number of life years lost due to that cause is similarly defined as

\[
\tilde{L}_j(0, \omega) = \sum_{i=1}^{n} I(D_i = j) e^*(T_i, \omega) / \sum_i I(D_i = j).
\]

Thus, the mean residual life times at time (age) of death \textit{from cause} \( j \) are averaged over those \textit{who died from that cause}. In the parlance of Gardner and Sanborn (1990) this is a measure of “Years of potential life lost”. Replacing the mean residual life time in these expressions by the difference \( \max(\tau - T_i, 0) \), a measure of “Premature (to \( \tau \)) years of potential life lost” is obtained. This is seen to resemble (4) (for uncensored data), however, we divide by \( n \) when the cause-specific measure is calculated while, in demography, division by the number \( \sum_i I(D_i = j) \) of deaths from cause \( j \) (before \( \tau \)) is done. As we have seen, our measure of the expected number of life years lost due to cause \( j \) before \( \tau \) has an interpretation as the parameter \( \mathbb{E}(\tau - T_j \wedge \tau) \) from the population where all \( k \) causes of death are operating and it does not involve standard mortality tables.

In biostatistics a framework where it is feasible to account for standard mortality is that of \textit{relative survival} or \textit{excess hazard} models. Here, the mortality rate, \( \alpha(t) \) for a subject is written as a sum

\[
\alpha(t) = \alpha_0(t) + \alpha^*(t)
\]

of a standard mortality rate, \( \alpha^*(t) \) (assumed known) and an excess mortality, \( \alpha_0(t) \). Often, this model is used in the absence of cause of death information, however, formally this is the hazard rate in a competing risks model with two causes of failure in which case \( \alpha_0(t) \) is interpreted as the mortality rate.
from the disease while $\alpha^*(t)$ is the mortality rate from other causes. Further, if cause of death information is available and if cause-specific standard mortality rates, $\alpha^*_j(t), j = 1, \ldots, k$ are given with

$$\alpha^*(t) = \sum_j \alpha^*_j(t)$$

then an excess mortality competing risks model for the cause-specific hazard

$$\alpha_j(t) = \alpha_{j0}(t) + \alpha^*_j(t)$$

can be studied. This can be seen as a competing risks model with $k + 1$ cause-specific hazards: $\alpha^*(t)$ for dying from disease-unrelated causes and $\alpha_{j0}(t) = \alpha_j(t) - \alpha^*(t)$ for dying from cause $j, j = 1, \ldots, k$ “due to the disease”.

For this competing risks model we have the relation

$$S(t) + \sum_j F_{j0}(t) + F^*(t) = 1$$

(with obvious notation) and the expected total number of life years lost before time $\tau$

$$L(0, \tau) = \tau - \int_0^\tau S(t) dt$$

can be decomposed into

$$L(0, \tau) = \sum_j \int_0^\tau F_{j0}(t) dt + \int_0^\tau F^*(t) dt.$$ 

Here, the interpretation is that $\int_0^\tau F^*(t) dt$ is the unavoidable expected number of life years lost due to the standard mortality while $\int_0^\tau F_{j0}(t) dt$ represents
the number of years lost from cause $j$ due to the disease.

Note that, in $L(0, \tau)$ the quantity

$$L^*(0, \tau) = \int_0^\tau F^*(t)\,dt = \int_0^\tau \int_0^t S(u)\alpha^*(u)\,du\,dt$$

appeared. This represents the standard number of life years lost when the excess mortality rates $\alpha_{0j}(t)$, $j = 1, \ldots, k$ are also operating and replaces the quantity

$$\int_0^\tau \int_0^t S^*(u)\alpha^*(u)\,du\,dt = \tau - \int_0^\tau S^*(t)\,dt$$

used in demography. The latter would be the number of life years lost before $\tau$ in a population where $\alpha^*(t)$ is the only hazard rate operating, a quantity which does not play a role here since all our computations relate to this world where all causes of death are present.

**Acknowledgements**

The research was supported by the Danish Natural Science Research Council [grant number 272-06-0442 “Point process modelling and statistical inference”]. Comments from Niels Keiding, Department of Biostatistics, University of Copenhagen are greatly appreciated. We thank Maja Pohar Perme, Institute of Biostatistics and Medical Informatics, University of Ljubljana for updating the pseudo package.
References


Appendix: Asymptotics

\[ \int_0^\tau (\hat{F}_j(t) - F_j(t)) dt = \int_0^\tau \int_0^t \left( (\hat{S}(u) - S(u)) d\hat{A}_j(u) + S(u)(d\hat{A}_j(u) - dA_j(u)) \right) dt \]

\[ = \int_0^\tau \int_0^t \left( \left( \frac{\hat{S}(u)}{S(u)} - 1 \right) S(u) d\hat{A}_j(u) + S(u)(d\hat{A}_j(u) - dA_j(u)) \right) dt \]

\[ \approx \int_0^\tau \int_0^t \left( - \int_u^t \frac{J(x)}{Y(x)} dM(x) S(u) d\hat{A}_j(u) + S(u) \frac{J(u)}{Y(u)} dM_j(u) \right) dt \]

(Duhamel’s equation and consistency of \( \hat{S} \) and \( \hat{A}_j \))

\[ = \int_0^\tau \int_0^t \left( - \int_u^t f_j(x) \frac{J(u)}{Y(u)} dM(u) + S(u) \frac{J(u)}{Y(u)} dM_j(u) \right) dt \]

(changing order of integration)

\[ = \int_0^\tau \int_0^t \left( (S(u) - F_j(t) + F_j(u)) \frac{J(u)}{Y(u)} dM_j(u) - (F_j(t) - F_j(u)) \frac{J(u)}{Y(u)} dM_{-j}(u) \right) dt. \]

In large samples, a process of the form \( \sqrt{n} \int_0^t f_t(u) J(u) dM(u)/Y(u) \) with a deterministic \( f_t(\cdot) \) behaves like a mean zero Gaussian process \( U(t) \) with a covariance \( \sigma(s,t) \) which may be estimated by

\[ \hat{\sigma}(s,t) = \int_0^{s\wedge t} \frac{f_s(u) f_t(u)}{Y^2(u)} dN(u). \]

Therefore, we have in large samples

\[ \sqrt{n} \int_0^\tau (\hat{F}_j(t) - F_j(t)) dt \approx \int_0^\tau U_j(t) dt + \int_0^\tau U_{-j}(t) dt. \]
where \( U_j \) and \( U_{-j} \) are independent. It follows that \( \hat{L}_j(0, \tau) \) is asymptotically normal and that the variance may be estimated by

\[
\int_0^\tau \int_0^\tau (\hat{\sigma}_j(s, t) + \hat{\sigma}_{-j}(s, t)) \, ds \, dt
\]

with \( f^+_j(u) = S(u) - F_j(t) + F_j(u) \) and \( f^-_j(u) = -F_j(t) + F_j(u) \). Thus,

\[
\hat{\sigma}_j(s, t) = \int_{s \wedge t}^{s \wedge t} \frac{(\hat{S}(u) - \hat{F}_j(s) + \hat{F}_j(u))(\hat{S}(u) - \hat{F}_j(t) + \hat{F}_j(u))}{Y(u)^2} \, dN_j(u)
\]

and

\[
\hat{\sigma}_{-j}(s, t) = \int_{s \wedge t}^{s \wedge t} \frac{(\hat{F}_j(u) - \hat{F}_j(s))(\hat{F}_j(u) - \hat{F}_j(t))}{Y(u)^2} \, d \sum_{r \neq j} N_r(u),
\]

where, \( N_r(\cdot) \) counts failures from cause \( r \).
Figure 1: Malignant melanoma study: Stacked plot of cumulative incidences of dying from the disease (dashed curve) and from other causes (difference between solid and dashed curves).
10/1  Andersen, P.K. & Skrondal, A. “Biological” interaction from a statistical point of view.


10/3  Parner, E.T. & Andersen, P.K. Regression analysis of censored data using pseudo-observations.

10/4  Nielsen, T. & Kreiner, S. Course Evaluation and Development: What can Learning Styles Contribute?

10/5  Lange, T. & Hansen, J.V. Direct and Indirect Effects in a Survival Context.

10/6  Andersen, P.K. & Keiding, N. Interpretability and importance of functionals in competing risks and multi-state models.

10/7  Gerds, T.A., Kattan, M.W., Schumacher, M. & Yu, C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring.


11/1  Kreiner, S. Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment.

11/2  Andersen, P.K. Competing risks in epidemiology: Possibilities and pitfalls.

11/3  Holst, K.K. Model diagnostics based on cumulative residuals: The R-package gof.


11/5  Holst, K.K., Budtz-Jørgensen, E. & Knudsen, G.M. A latent variable model with mixed binary and continuous response variables.


12/01 Keiding, N., Hansen, O.K.H., Sørensen, D.N. & Slama, R. The current duration approach to estimating time to pregnancy.

12/02 Andersen, P.K. A note on the decomposition of number of life years lost according to causes of death