Austrian Journal of Statistics
March 2016, Volume 45, 53–67.
http://www.ajs.or.at/
doi:10.17713/ajs.v45i2.121

Life Expectancy Comparison between a Study Cohort and a Reference Population

Georg Zimmermann
Department of Neurology,Christian Doppler Klinik, Salzburg
Paracelsus Medical University, Salzburg
Paris Lodron University of Salzburg

Abstract

Questions of life expectancy are highly relevant in clinical practice, in particular if one is interested in the life expectancy loss due to a certain disease in comparison with the general population. Therefore, we propose a methodology which enables the researcher to compare more easily life expectancies between a study cohort and a reference population. Firstly, we take the formula commonly used in official statistics and adapt it to a survival analytic setting, thus establishing a link between official statistics and survival analysis. Further, we discuss two commonly encountered sources of potentially severe bias, and how to remedy them. We hope that the proposed approach facilitates the use of life tables, particularly in clinical studies with mortality as primary endpoint.

Keywords: life expectancy, life table, life expectancy comparison, survival analysis, expected remaining life, Larynx dataset.

1. Introduction

In medical studies, the survival experience of a patient cohort is often expressed in terms of some risk measure, for example, the risk of dying. However, sometimes, this type of quantities is not completely satisfying: A patient suffering from a certain disease is probably more interested in his or her loss of remaining lifetime than the mere information that he or she has a risk of dying elevated by some factor compared to healthy persons, because the first quantity is much easier and more naturally interpretable than the latter one. Consequently, the doctors should also be able to provide their patients with information about life expectancies. But interestingly, it seems as if the number of studies merely examining some risk measures such as standardized mortality ratios or hazard ratios is much higher than the number of life expectancy analyses. In particular, life expectancy comparisons between a patient group and some reference population are hardly available, although this kind of question is a very natural and highly relevant one.
There are basically two measures of life expectancy (for another concept of "life expectancy", see Bradshaw, Stobie, Knuiman, Briffa, and Hobbs 2015): The so-called *mean residual life* is used in survival analysis, a certain branch of mathematical statistics (Kalbfleisch and Pren-

tice 2002), whereas the *average expectation of life* or *life expectancy at exact age* x is popular in the context of life table calculations in offical statistics (Chiang 1979). The latter measure can also be used in analyses of study cohorts, which is a quite popular method in life expectancy analyses (Nusselder, Slockers, Krol, Slockers, Looman, and van Beeck 2013; DuGoff, Canudas-Romo, Buttorff, Leff, and Anderson 2014; Guaraldi, Cossarizza, Franceschi, Roverato, Vaccher, Tambussi, Garlassi, Menozzi, Mussini, and D'Arminio Monforte 2014; Nagai, Kuriyama, Kakizaki, Ohmori-Matsuda, Sone, Hozawa, Kawado, Hashimoto, and Tsuji 2011). On the other hand, survival analytic quantities such as Kaplan meier estimators are sometimes used, too (Chang, Lu, Lee, Hwang, Cheng, and Wang 2015; Luangasanatip, Hongsuwan, Lubell, Limmathurotsakul, Teparrukkul, Chaowarat, Day, Graves, and Cooper 2013). Moreover, there are also "mixtures" between these two concepts insofar as, at first, a survival analytic model is fitted to the data and, then, life expectancies are calculated based on the life table methodology (Li, Hüsing, and Kaaks 2014; Strauss, DeVivo, Paculdo, and Shavelle 2006; Gaitatzis, Johnson, Chadwick, Shorvon, and Sander 2004). However, in some of these studies (Li *et al.* 2014; Nagai *et al.* 2011; Strauss *et al.* 2006), life expectancy comparisons are made only between different subgroups of the study cohort, but not between the patients and some reference population. Even in those cases where comparisons are carried out (Luangasanatip *et al.* 2013; Gaitatzis *et al.* 2004), some mathematical and conceptual issues are not completely clear. Therefore, clarification is needed as to how a study cohort and some reference population can be compared with regard to their life expectancy. Such clarification should rest on both statistically and conceptually rigorous arguments. These theoretical clarifications will hopefully lead to a methodology which can facilitate understanding by statistics practitioners, too. Furthermore, we will establish a link between official statistics and survival analysis, which is, in our opinion, a desirable goal not only for scientific, but also for practical reasons (see the closing part of this paper).

2. Basic quantities measuring life expectancy

To set the stage, let's assume that we want to define the life expectancy for a person of exact age x years. As it would hardly make any sense to take an unrestricted range for x, we assume $x \in \{0, 1, ..., x_{max}\}$, where x_{max} is chosen appropriately, for example, $x_{max} = 95$ or $x_{max} = 100$. Furthermore, let X denote a nonnegative continuous random variable measuring the time from birth to death of this person. For sake of simplicity, we do not take gender into account here.
Then, in the context of life table calculations, the following definitions are popular (Hanika and Trimmel 2005):

(i) The *death probability in the age interval* $[x, x+1)$ is defined as

$$q_x := P(x \leq X < x+1 | X \geq x).$$

(ii) Let $l_0 \in \mathbb{N}$. The *number of survivors at age* x is defined as

$$l_x := \begin{cases} l_0 & x = 0 \\ l_{x-1} \cdot (1 - q_{x-1}) & x \geq 1 \end{cases}$$

(iii) The *number of years lived in the age interval* $[x, x+1)$ is defined as

$$L_x := \frac{1}{2}\left(l_x + l_{x+1}\right).$$

Based on these quantities, we define the *life table life expectancy (LTLE) at age* x as

$$E_x := \frac{1}{l_x} \sum_{y=x}^{x_{max}} L_y. \tag{1}$$

Note that this quantity is usually referred to as *life expectancy at exact age* x. However, we choose this terminology to stress the contrast to the life expectancies calculated for the study cohort (see below). Next, we turn to the survival analytic setting, which requires slightly different assumptions. We consider a situation typically encountered in clinical studies, namely that we have a study cohort which is followed over a certain time period. Let T be a non-negative continuous random variable measuring the time from the starting point of the study. Then, the *mean residual life at time* $t \geq 0$ is defined as

$$r(t) := E[T - t | T \geq t].$$

If $E[T] < \infty$, it can be shown that for all $t \geq 0$, we have

$$r(t) = \frac{1}{S(t)} \int_t^\infty S(u) \mathrm{d}u,$$

where $S(t) = P(T > t), t \geq 0$, is the so-called *survivor function*. The survivor function can be estimated either by some step function (e.g., the Kaplan-Meier estimator) or by specifying a certain parametric model such as the Weibull model.
Since this model will be used for illustrative purposes in the theoretical considerations below, we shall state now how it is specified. A Weibull model is characterized by setting

$$S(t) = exp(-\lambda t^p),$$

where $\lambda, p > 0$. According to the impact on the survival function, λ is usually called *scale parameter*, whereas p is referred to as *shape parameter*. As in other parametric models used in survival analysis, covariates can be easily incorporated by an additional term of the form $exp(\beta' \mathbf{x})$: The survivor function is then given as

$$S(t, \mathbf{x}) = exp(-\lambda t^p exp(\beta' \mathbf{x})).$$

The parameters of such a model can be estimated by means of maximum likelihood estimation, which eventually yields an estimator $\hat{S}$ of the survivor function S. For details concerning these pieces of survival analytic theory, we refer to Kalbfleisch and Prentice (2002).

3. Comparing a study cohort to some reference population

Now, the question arises if a link between these two life expectancy measures can be established in order to compare the study cohort with a reference population. In general, we can't just calculate the differences between the mean residual lifetimes (originating from the cohort data) and the corresponding LTLEs: For example, in the context of life tables, the deaths are assumed to be uniformly distributed in each age interval $[x, x + 1)$, whereas this is not the case if we, for example, fit a Weibull model to the study cohort data. However, if we take a closer look at the LTLE formula, it turns out that there is a possibility for the survivor function S to enter the stage. We will see in the following sections that for subsequent years after start of follow-up, we are thus able to calculate life expectancies for the study cohort.

3.1. Mathematical justification

To begin with, note that for all $x \in \{0, 1, ..., x_{max}\}$, we have

$$E_x = \frac{1}{2} + \frac{\tilde{S}(x_{max} + 1)}{2\tilde{S}(x)} + \frac{1}{\tilde{S}(x)} \sum_{k=x+1}^{x_{max}} \tilde{S}(k), \tag{2}$$

where $\tilde{S}(x) = P(X > x)$.
For a proof of this equality, see Appendix A.

Now, based on the LTLE formula (2) just derived, we want to calculate the life expectancy at time $t \geq 0$ after start of follow-up for a person from the study cohort. The only thing left is to examine the relation between $S(t) = P(T > t)$ and $\tilde{S}(x) = P(X > x)$. Recall that X measures the time from birth to death, whereas T takes start of follow-up as the time origin. Therefore, if we let a_E denote the exact age of that person at start of follow-up, we have $T = X - a_E$. Thus, we get that for all $x \in \{a_E, a_{E+1}, ..., x_{max}\}$ (note that $x < a_E$ wouldn't make any sense since we are only interested in time points after start of follow-up!), the following equations hold:

$$\tilde{S}(x) = P(X > x) = P(X - a_E > x - a_E) = P(T > x - a_E) = S(x - a_E).$$

So, we can write (2) as

$$E_x = \frac{1}{2} + \frac{S(x_{max} - a_E + 1)}{2S(x - a_E)} + \frac{1}{S(x - a_E)} \sum_{k=x+1}^{x_{max}} S(k - a_E),$$

for $x \in \{a_E, a_{E+1}, ..., x_{max}\}$. If we set $t := x - a_E$ and do some re-indexing, we can define the *study cohort life expectancy* as

$$SCLE(t) := \frac{1}{2} + \frac{S(x_{max} - a_E + 1)}{2S(t)} + \frac{1}{S(t)} \sum_{k=t+1}^{x_{max} - a_E} S(k), \tag{3}$$

where $t \in \{0, 1, ..., x_{max} - a_E\}$.

Now, we are ready to compare life expectancies between the study cohort and a reference population: At first, we have to pick a certain estimator $\hat{S}$ for the survivor function S of the study cohort (e.g., the survivor function of a regression model fitted to the data). Then, we calculate $SCLE(t)$ for some values $t \in \{0, 1, ..., x_{max} - a_E\}$ of interest. Due to the derivation carried out above, the corresponding value from the life tables which should be taken for comparison is E_{t+a_E}.

3.2. Conceptual considerations

So far, we have established a method where we use the LTLE formula not only for calculations in the context of life tables, but also for quantities based on a survival analytic model. However, although the structure of the formula is the same now, we may need further refinements in certain situations to make sure that we control several sources of bias.

Bias due to "infinite life" models

To begin with, the follow-up time in clinical studies is usually restricted to one or several years. As a consequence, we expect that in general, some of the subjects will be still alive at the end of the study period. For these persons, we don't have the exact survival times. This phenomenon itself, known as right-censoring, which is a key issue in survival analytic methodology, doesn't cause any serious problems. But, especially when we want to estimate life expectancies, we have to be aware of the fact that we often can't avoid making extrapolations to some extent. For example, to calculate the life expectancy of a subject aged 20 at study entry using formula (3), we need values $S_{20}(j)$ up to $j = x_{max} - 19$. By the subscript 20, we indicate that we take the age at entry as a covariate into account here (e.g., by specifying some regression model as mentioned above). Most likely, $x_{max} - 19$ will be greater than the follow-up time, which means that we have to extrapolate beyond the range of our data. In other words, it may not be possible to judge whether, for example, the estimate of $S_{20}(50)$ is reliable or not because most likely, the follow-up time won't be 50 years. Observe that although we may well have data of patients aged 70 at study entry, we must not assume that $S_{70}(0) = S_{20}(50)$

holds in general. Thus, we see that looking at the values of the survivor functions of the patients who were fairly old at time of entry doesn't help solving the problems concerning extrapolation. Therefore, especially for the patients who were quite young at study entry, the estimation of at least some values of the survivor function can be crucial. This point is very important to keep in mind, especially when using models such as the Weibull model which yields values $S(t)$ for all $t \geq 0$: Although it seems as if we could immediately get all we need for the calculation of life expectancies from such a model, we must not use these values without examining their appropriatness. For example, if we consider a Weibull model, it can be shown that the so-called hazard function, which measures the "instantaneous risk of dying", is given as

$$h(t) = \lambda p t^{p-1},$$

which is either decreasing ($p < 1$), increasing ($p > 1$) or constant ($p = 1$) (Kalbfleisch and Prentice 2002). So, if our data yields a maximum likelihood estimate $\hat{p}$ of p which is smaller than 1, this suggests that the risk of dying decreases over time. Whereas this could be reasonable to some extent (for example, think of a risk reduction due to protective effects of clinical monitoring etc.), this is most likely wrong for large values of t: At some time point, the "force of death" will be stronger than any protective effects. If we don't take this issue into account, we will get life expectancy estimators which may be extremely biased since the behaviour of the hazard discussed above translates to a biased survivor function.

To deal with this problem, it may help to pick a model which allows for more flexibility (e.g., more complicated shapes of the hazard function). Although such questions of model appropriatness are important and should not be neglected, it can well be that in some cases, the values of $S(t)$ remain unrealistic: For example, if we have a cohort of relatively young subjects and the follow-up time is, say, 5 years, we can't see a substantial increase in the risk of dying at age 50 simply because we don't observe persons at this age!

Consequently, we suggest defining some kind of "breakpoint" value t_B from which on we don't "trust" the model any more, or at least put some restrictions on the use of model death probabilities (i.e., the q_j's in formula (6) in the appendix, based on the survivor function of the model), which are defined as

$$q(t) := 1 - \frac{S(t+1)}{S(t)}, t \in \{0, 1, ..., x_{max} - a_E\}.$$

For example, it would make sense to take the maximum of the model and the life table death probability q_{t+a_E} for time points t greater than t_B. From a medical point of view, this means that from a certain time point after diagnosis onwards, you assume that the patients are at a risk of dying which is at least as high as for the reference population. To illustrate this idea, assume that we have decided to set $t_B = 5y$. This means that for the first 5 years following study entry, we use $q(0), q(1), q(2), ..., q(5)$ as annual death probabilities for our calculations. But, for all time points $t = 6, 7, ..., x_{max} - a_E$, we take $max\{q(t), q_{t+a_E}\}$ as annual death probability. Of course, if $x_{max} - a_E < 6$, that is, the patient is fairly old at time of diagnosis, this modification doesn't have any effect on the life expectancies because we only need annual death probabilities up to, for example, 3 years after study entry.

Besides, putting a restriction like the one above on the death probabilities for the entire time range would not make sense (unless you have good reasons for such a decision) since we would thus wipe out any protective effects due to, for example, regular visits at the hospital. Note that at least in some studies, it's indeed reasonable to account for such a sort of effect: For example, if we consider cancer patients, we can rightfully assume that they are scheduled for regular examinations, which may imply that some other diseases are detected earlier than in healthy subjects. Note that this does not mean the results are biased: Of course, if we, for example, examined the effect of a certain therapy on the survival of patients suffering from brain tumors, it's obvious that the effect measure would indeed be biased when the control group contains far more subjects with cardiovascular disease than the treatment group. But, note that in this case, the occurence of brain tumors is not supposed to be

related to cardiovascular disease. By contrast, in the example with the regular medical checks from above, we don't have confounding because as we stated before, the main point is that these examinations are caused by the fact that the patient suffers from cancer. Therefore, the possibly beneficial effect actually turns out to be a part of the disease effect.
Moreover, protective disease-related effects may also be caused by the fact that a patient is sometimes forced to change his or her lifestyle: For example, if you have epilepsy, you are usually not allowed to drive a car any more. Likewise, you won't do any dangerous activities such as parachuting or climbing. Thus, you most likely reduce your risk of dying substantially compared to the overall population.
Summing up, we want to point out that although at first sight, it seems as if protective effects introduce bias to the results, we see that in fact, the contrary is true - if we wiped out these effects in our analysis, we would actually eliminate a part of the disease effect! Of course, it's theoretically possible to eliminate the impact of the protective effects mentioned above in order to arrive at some, let's say, "cleaned" effect estimate which, loosely speaking, measures the effect of the disease "itself" (i.e., the loss of life due to medical reasons only). However, what's the practical relevance of such an estimate? For example, is there any use in calculating the effect of epilepsy "itself" (i.e., the impact of the fact that the patients don't drive cars is somehow eliminated) if there isn't anyone who has epilepsy and is allowed to drive a car? To cut it short, we would then have calculated an effect for a person which actually doesn't exist!
To turn back to the breakpoint issue, we must admit that the choice of such a breakpoint value is a crucial task. We suggest taking a look at carefully collected and thoroughly analysed data such as the most recent life tables for the population the patients come from. For example, the Austrian life tables from 2000/02 show a monotonic increase of the death probabilities from age 30 onwards (Hanika and Trimmel 2005). Alternatively, medical experts may also provide useful information concerning this issue. Moreover, an appropriate breakpoint value may be suggested by the design of the study, for example, if at some time point, the members of the study cohort are not scheduled for regular visits at the hospital and thus don't take advantage of protective effects any more. However, keep in mind that for example, there can be protective effects which may influence the survival experience for the entire time range (e.g., if the patients aren't allowed to drive cars any more).
To sum things up, it may be quite difficult to find an appropriate breakpoint value. We suggest taking not only one of the approaches outlined above, but various different considerations into account. For example, medical experts may have quite a good idea of the amount of risk decrease due to protective effects. On the other hand, a look at some life tables may indicate at which age the "force of death" is most likely stronger than any protective effects. Of course, we must admit that this way of solving the problem might not be satisfying at first sight. But, on the other hand, this means of correction is quite easy to carry out and helps to avoid an amount of bias that can be quite large. To cut it short, having a breakpoint which is roughly appropriate is definitely better than introducing no breakpoint value at all: Note that if we have an extremely unrealistic assumption such as a decreasing risk of dying even for large values of t, we sum up this bias when calculating life expectancies and thus get results which are actually useless. So, with our method, we can substantially reduce this kind of bias, with the only drawback that we don't have a formally justified method at hand to determine the breakpoint value t_B.
By the way, recall that the extrapolation issue was the main point we started from. If we are indeed interested in calculating life expectancies, we suggest following the "breakpoint" idea discussed in detail above. But, instead of calculating this quantity (i.e., the expected number of years the subject has left to live), we may also think of restricting ourselves to looking only at some time span such as, for example, 5 years after study entry. All we have to do is to choose x_{max} appropriately. To stay with our example, we just set $x_{max} = a_E + 5$. Note that due to the concept underlying the life table calculations (see section 2), we then get the expected number of years a subject aged a_E has left to live during the following 5 years. Especially in studies with short follow-up time, extrapolation is quite a serious

concern. To circumvent this problem, the idea outlined above may be a good alternative to life expectancy calculations. Likewise, when analyzing data from patients who are at a very high risk of dying in a moreorless short time interval after study entry (e.g., if some surgical procedure is defined as the start of follow-up), there may be no need to calculate life expectancies - instead, something like a "average 5-year lifetime" as proposed above would perhaps be a more useful measure.

Bias due to changes in life expectancy over time

Now, let's turn to a second important issue, namely the information underlying the data from the study cohort and the life table, respectively. What we haven't mentioned so far is the fact that in a clinical study, the subjects often enter the study at different time points, for example, when the diagnosis of a certain disease is defined as the starting point. As far as the calculation of estimates within the study cohort is concerned, this does not cause problems at all: If the follow-up period is quite short, the difference in entry times is negligible, and even if the subjects are followed over several years, we can for instance fit a regression model which includes a covariate "year of entry" or something like that. To judge if one should account for the time point of entry, we suggest using variable selection methods. In addition to that, a look at the life tables for the population of the area the subjects (mainly) belong to can be also useful to get an idea of the amount of change during the study period. However, if we turn to comparisons with life table quantities, we have to be more careful. Again, if the recruitment as well as the follow-up period are quite short, say, for example, one year, it suffices to take one single reference life table for comparisons. This is also appropriate if hardly any changes in the survival experience of the general population can be observed over the study period. But what if the recruitments and/or follow-ups stretch over several years **and** we observe substantial changes in the general population? At first sight, the solution to this problem is easy: If we want to compare the life expectancy of a patient aged a_E diagnosed in, say, 2000, at time of diagnosis (i.e., in the notation from above, $t = 0$) to the corresponding value of the reference population, we just take the number $LTLE(a_E)$ from the life table of 2000 and calculate the difference to $SCLE(0)$. Analogously, if a comparison of life expectancies 5 years after diagnosis is desired, we calculate $SCLE(5)$ and compare this quantity with $LTLE(a_E + 5)$ from the life table of 2005.
However, note that a period life table is always based on the survival experience of a population in a very certain year (or a quite short time period, say, for example, three years, see Hanika and Trimmel 2005). For the calculation of life expectancies, the annual death probabilities for future years are estimated by the corresponding values of the reporting period, which means that the status quo is carried forward (Hanika and Trimmel 2005). Although this basically makes sense, we get into troubles when using the LTLEs for comparison purposes: To turn back to the example from above, let's see what happens if we use $LTLE(a_E)$ from the life table of 2000 as comparison value for $SCLE(0)$. We assume that the follow-up time is 10 years, and, as stated above, a substantial change in life expectancies can be seen in the life tables published during this time span. Then, the key point is that the subject on study has been followed over these years and, thus, the changes in survival experience are somehow implicitely accounted for, whereas the comparison value from the life table of 2000 is based on information of that certain year only and therefore "ignores" the changes observed from 2000 to 2010. To bridge this information discrepancy, we propose the following approach: Suppose a patient enters the study at age a_E in the year y, where $y \in \{y_S, y_S + 1, ..., y_E\}$. By y_S and y_E, we denote the year of study start and end, respectively. For sake of simplicity, we don't assume separate recruiting and follow-up periods, although the method outlined below can be carried out analogously for this case. Although a_E can basically take values from 0 to x_{max}, for the following algorithm, we exclude x_{max} because for a person aged x_{max}, we only need the values $l_{x_{max}}$ and $l_{x_{max}+1}$ to calculate his or her life expectancy, so the quantities we have in the table of the year y are already the ones we need. Secondly, our study ends in y_E, so the life table of this year doesn't require an update since if we assumed any knowledge

concerning survival experience in $y_E + 1, y_E + 2$ and so forth, we would again have different degrees of information in the study cohort and the population.
Recall that for LTLE calculations (i.e., for life expectancy calculations based on formula (1) in section 2), we need the numbers of survivors for all ages from a_E to $x_{max} + 1$, which are denoted by $l_{a_E}, \ldots, l_{x_{max}+1}$. The algorithm used for this purpose can be described as follows:

$$l_{a_E} := l_{a_E,y}, \quad l_{a_E+1} := l_{a_E+1,y}$$
$$l_{a_E+k} := \begin{cases} l_{a_E+k-1} \cdot (1 - q_{a_E+k-1,y+k-1}) & k \in \{2, 3, ..., \min(y_E - y + 1, x_{max} - a_E + 1)\} \\ l_{a_E+k-1} \cdot (1 - q_{a_E+k-1,y_E}) & k \in \{y_E - y + 2, y_E - y + 3, ..., x_{max} - a_E + 1\} \end{cases}$$

The first line means that for a_E and $a_E + 1$, we simply take the values from the life table of the year of entry y. Then, the numbers of survivors are calculated according to the following scheme: To get l_{a_E+2}, we multiply l_{a_E+1} with the annual survival probability for the age $a_E + 1$ from the life table of the year $y + 1$. In this way, we proceed until we reach either $x_{max} - a_E + 1$ or $y_E - y + 1$. Now, if $x_{max} - a_E \leq y_E - y$, we are done since this means that we already have all numbers of survivors up to $l_{x_{max}+1}$. Otherwise, the remaining l_{a_E+k}'s are calculated by applying the "standard" method of life table construction, using the annual survival probabilities of the life table y_E.
So, to sum things up, we basically follow the usual formula for the number of survivors, but with the important modification that for the entire study period, we take all available information on the population's survival experience into account because our calculations are based on year-wise survival probabilities. In other words, we thus get "dynamic" life expectancies for the general population which are, in terms of the information used, indeed comparable to the SCLEs based on the study cohort's data.

To close this section, we briefly discuss an issue of practical importance. As already mentioned at the beginning of this paper, life expectancy is a measure which can be quite easily understood not only by researchers, but also by people not involved in statistics or medicine. Therefore, it's desirable to use the results from a life expectancy analysis as a nice interpretable piece of information the doctors can communicate to their patients. However, we should be aware of the fact that possible changes in life expectancy are crucial for answering the question if we can indeed take the values we've calculated as estimates for the losses or gains in lifetime of future patients. For example, if the study was terminated at the end of 2014, it's most likely fine if a doctor communicates these values to patients who are currently at his clinic. But, if the follow-up ended in 2000, you should be more cautious of course: Although the "dynamic" population life expectancies can quite easily be re-calculated by additionally taking the life tables from 2001 to 2015 into account, the life expectancies for the patient cohort can't be updated since the study was terminated in 2000. However, the patients' life expectancies may also substantially change over time, partly due to the changes in the population, but also as a consequence of therapeutic advances, etc. Of course, if we fitted a regression model with year of entry to the data, we could theoretically plug in the current year as covariate value and finally get a life expectancy estimate. But, obviously, the corresponding regression coefficient was estimated using data from patients who entered the study before 2000, so we would, again, extrapolate beyond the range of our data.
So, to sum things up, before carrying over the results of studies with limited follow-up time to future patients, one should carefully think about changes in life expectancies which may have occurred since the end of the study. For this purpose, information on medical as well as demographic developments and trends are needed. Besides, it should be mentioned that sometimes, the follow-up time can be considered as unlimited: For example, if you actually have the original data (i.e., date of birth, etc.) of the patients at hand and link it with a death registry (e.g., with some probabilistic record linkage method, see Oberaigner and Stühlinger (2005)), the survival experience of the patient cohort can indeed by updated at any time. Based on this data, the survival analytic model is fitted again, and life expectancies are

calculated afterwards. Thus, in this case, the estimates are reliable even for a person being at the doctor's office right now.

4. Data example

To illustrate the methods outlined above, let's take a look at a dataset which was reported by Kardaun (1983). The cohort consists of 90 males diagnosed with laryngeal cancer between 1970 and 1978 at a Dutch hospital. The dataset contains their survival times (in years), that is, the time from entry to either death or Jan 1, 1983, as well as the year of diagnosis, the age at study entry and the stage of cancer on a scale from 1 to 4. Some descriptives of this data are contained in Table 1. For details concerning this dataset, we refer to Kardaun (1983) and Klein and Moeschberger (2003). This dataset is available in the survival package (Therneau 2014) in R (R Core Team 2015).

Table 1: Some descriptives for the Larynx dataset.

number of patients	90
person years of follow-up	377.8
number of deceased (%)	50 (55.6)
median age (range)	65 (41-86)
median diagnosis year (range)	1974 (1970-1978)
stage 1 (%)	33 (36.7)
stage 2 (%)	17 (18.9)
stage 3 (%)	27 (30.0)
stage 4 (%)	13 (14.4)

Following the considerations of the previous chapters, we now carry out the following steps using R:

1. We fit a Weibull regression model with stage, age at diagnosis and year of diagnosis to the data by using the packages survival and SurvRegCensCov (Hubeaux and Rufibach 2014). Note that the 4 stages are coded as dummy variables in the following way: We introduce 3 dummy variables *dummyA, dummyB* and *dummyC* such that stage 1 corresponds to $dummyA = dummyB = dummyC = 0$. The other stages are coded as $100, 010$ and 001, respectively. The results of the model fit are displayed in the table below.

Table 2: Results of Weibull model fit for the Larynx dataset.

	estimate	standard error
scale (λ)	0.094	0.525
shape (p)	1.121	0.141
dummyA	0.181	0.463
dummyB	0.665	0.356
dummyC	1.780	0.431
age	0.019	0.014
year	-0.022	0.073

Recall that the survival function of a Weibull regression model is given as $S(t, \mathbf{x}) = \exp(-\lambda \mathbf{t}^{\mathbf{p}} \exp(\boldsymbol{\beta}' \mathbf{x}))$. As indicated by the estimate of the shape parameter, the probability of surviving is decreasing moderately fast. To turn to the impact of the different stages on survival, the estimates of the dummy variables show what we expect for medical reasons: The higher the stage, the lower the value of the survival function is. Likewise, it's small suprise that the probability of surviving decreases with age. Interestingly, patients who were diagnosed later seem to have a better chance of surviving. However, all the interpretations I've mentioned have to be taken with caution since the estimated standard errors are quite large in most cases.

2. To deal with the "infinite life" problem discussed in section 3.2., we take a look at the age-specific death probabilities of the annual Dutch life tables from 1970 to 1982 which are available upon request at the website of Statistics Netherlands (Statistics Netherlands 2015). Obviously, there is a remarkable increase from age 40 onwards. As the patients' ages at diagnosis range from 41 to 86 years, it's reasonable to assume that the whole study cohort is exposed to a relatively high, increasing risk of dying. Therefore, the breakpoint should be set to a fairly small value (of course, the choice of the breakpoint can also be based on medical considerations, as mentioned above). We decided to take 0, 3 and 5 years as breakpoints and compare the results. This means that from 1, 4 and 6 years after diagnosis onwards, respectively, we take the maximum of the model and the life table death probabilities instead of merely using the model death probabilities, as described in section 3.2. In step 3, these calculations will be illustrated by an example.

3. For sake of simplicity, we only take one certain covariate combination, namely a person aged 65 which enters the study in 1974 with the diagnosis of stage 2 cancer. We are interested in this person's life expectancy at time of entry (i.e., t = 0) and 2 years after entry (t = 2), respectively. To calculate the life expectancies based on the model, we at first take $\hat{S}$, the estimate of the Weibull regression model CDF from step 1, and plug in the given covariate values for stage, age and year as well as $t = 0$ (or $t = 2$). Then, we calculate the annual death probabilities $1 - \hat{S}(1)/\hat{S}(0), 1 - \hat{S}(2)/\hat{S}(1)$ and so forth, as stated in section 3.2. and in the proof of (2) in the appendix. As the Dutch life tables contain values up to an age of 99 years, we need annual death probabilities up to $1 - \hat{S}(t_{max} + 1)/\hat{S}(t_{max})$, where $t_{max} = 99 - 65 = 34$. To explain the latter formula, recall that we want to calculate the life expectancy for a person aged 65 at time of diagnosis. To make sure that there is a correspondence between the highest age x_{max} in the life table and the greatest time point t_{max} after diagnosis, we choose the latter value such that if we "follow" the patient aged 65 at time of diagnosis for t_{max} years, we arrive exactly at the highest life table age x_{max}. Once we have the annual death probabilities, we immediately get all the other life table quantities, especially the life expectancies, by using the definitions given in section 2. Keep in mind that the value of the breakpoint is crucial for the actual usage of the probabilities just calculated: For example, if we set $t_B = 0$, we at first use $1-\hat{S}(1)/\hat{S}(0)$. But then, we take the maximum of $1 - \hat{S}(2)/\hat{S}(1)$ and the annual death probability for a 66-year old from the life table of 1975, the maximum of $1 - \hat{S}(3)/\hat{S}(2)$ and the annual death probability for a 67-year old from the life table of 1976, and so on, as described in section 3.2.
   It should be mentioned that for the life expectancy calculations, we can take formula (3) as well, which is more straightforward than the method outlined above. However, it's sometimes good to have the annual death probabilities at hand, too. Therefore, we used the life table formulas in our calculations.

4. To assess the variability of the estimates calculated in the previous step, we calculate 95% confidence intervals using the bias-corrected and accelerated (BCa) bootstrap method (Carpenter and Bithell 2000). More to the point, we draw a sample of size $n = 90$ with replacement from the dataset, carry out the steps 1-3 2000 times and save the results (i.e., the life expectancy estimates) in an array. Finally, we use the bootBCa function in the rms package (Harrell 2015) to get the desired confidence intervals.

5. To turn to the population life tables, we observe life expectancy changes up to 1.6 years in the period from 1970 to 1982. Especially for ages between 0 and 50, the life expectancy increase exceeds 1 year. Therefore, we should use "dynamic" life expectancies: Based on the Dutch life tables, we create a new array containing the "dynamic" life expectancies by implementing the algorithm proposed in section 3.2.

6. Finally, we take the "dynamic" life expectancies for a 65-year old in 1974 and a 67-year old in 1976 and subtract them from SCLE(0) and SCLE(2) calculated in step

3, respectively. In the same manner, we get confidence intervals for the resulting life expectancy differences: As we consider the population life expectancies as fixed values, we can simply take the confidence intervals from step 4 and subtract the population values from the interval endpoints.

The results of the procedure outlined above are displayed in Table 2 and Table 3. Recall that a negative value indicates a decreased life expectancy of the patient, whereas a positive number corresponds to an increased life expectancy of the cohort member compared to a healthy person. So, obviously, a laryngeal cancer patient with the characteristics mentioned above tends to have decreased life expectancies compared to the population. However, the variability of the estimates is fairly high (recall that the Weibull parameter estimates come with a relatively high standard error, see Table 2). Consequently, it is hard to tell if the life expectancy loss is substantial or not. Likewise, when taking only the point estimates for 0 and 2 years after the start, there seems to be a slight time trend. However, the corresponding confidence intervals look quite similar, so actually, the data does not allow a clear statement about any time trends in the life expectancy differences.

Table 3: Life expectancies for a patient aged 65 at time of entry in 1974 and corresponding values for the reference population. SCLE = study cohort life expectancy (i.e., life expectancy of the patient with the covariate values mentioned above), LTLEd = dynamic life table life expectancy. SCLEs are given with 95% confidence intervals in brackets.

	$SCLE(0)$	$SCLE(2)$	$LTLEd(65,74)$	$LTLEd(67,76)$
$t_B = 0$	8.86 (5.36, 14.17)	8.36 (4.79, 12.94)	14.16	12.94
$t_B = 3$	8.86 (5.27, 14.04)	8.36 (4.80, 12.97)	14.16	12.94
$t_B = 5$	8.86 (5.18, 14.01)	8.36 (4.62, 12.85)	14.16	12.94

Table 4: Life expectancy differences with 95% confidence intervals for a patient aged 65 at time of entry in 1974. SCLE = study cohort life expectancy (i.e., life expectancy of the patient with the covariate values mentioned above), LTLEd = dynamic life table life expectancy

	$SCLE(0) - LTLEd(65,74)$	$SCLE(2) - LTLEd(67,76)$
$t_B = 0$	-5.29 (-8.80,0.01)	-4.58 (-8.14, 0.00)
$t_B = 3$	-5.29 (-8.89, -0.11)	-4.58 (-8.14, 0.03)
$t_B = 5$	-5.29 (-8.98, -0.14)	-4.58 (-8.32, -0.08)

Moreover, the point estimates of the life expectancy differences seem to be a bit surprising at first sight because they are -5.29 years for $t = 0$ and -4.58 years for $t = 2$, no matter which of the three breakpoint values we choose! When having a look at the annual death probabilities based on the model, we see that right from t = 0 onwards, these values are much higher than the corresponding quantities for the population, except for time points which are quite far away from the time of entry (in the example above, the annual survival probabilities of the patients exceed the corresponding values for the population from 20 years after start onwards). Consequently, it's small suprise that it does not matter if we set t_B to 0, 3 or 5 - the model probabilities will be higher anyway.
However, the choice of t_B may be crucial if we consider other covariate combinations, for example, if the value of age at entry is large. Moreover, the impact of the breakpoint heavily depends on the parameters of the Weibull CDF, especially on the scale λ and the shape p. To take a hypothetical example, we set $\lambda = 0.01$, instead of $\lambda = 0.09$ as in our model fit above (actually, this choice might not be that hypothetical because the standard error of the maximum likelihood estimate of λ is quite large!). Then, the values of $SCLE(0)$ are 0.26, 1.01 and 1.71 for $t_B = 0, 3, 5$, respectively. Thus, we see that different choices of t_B can indeed affect the results! In other words, one should be aware of the fact that introducing a breakpoint is not a question of how pedantic you are: There may be situations where this idea helps avoiding a substantial amount of bias. As discussed in section 3.2., the survival probabilities of the patients may be at least partially unrealistic for time points not covered

by the follow-up time (i.e., time points which are "far away" from the start). To turn back to the hypothetical example from above, the change in the life expectancy differences indicates that the survival probabilities of the cohort are higher than the corresponding values for the population soon after the start. However, it is sometimes crucial to decide which breakpoint is the most appropriate one. For several possibilities of solving this problem, we refer to the discussion in section 3.2.

5. Summary and closing remarks

In this paper, we have described an approach to life expectancy comparisons which is based on a re-formulation of the formula used in the context of life table calculations. Thus, we have made sure that the study cohort and the population quantities are structurally the same. Then, we have discussed two common major sources of bias and how to remedy them. We'd like to emphasize once more that this is not a question of only peripheral interest: If those sources of bias aren't controlled, the results will be unreliable or, in fact, even useless.
It should be mentioned that the methodology proposed above can be applied to quite a broad variety of settings: If we want to set up a more complicated model for the study cohort and take several covariates such as type of disease, health status, etc. into account, all we have to do is to fit an appropriate survival analytic model to the study cohort data. Then, we can proceed in the way outlined above and make life expectancy comparisons with the age- and gender-matched "dynamic" values from the life tables. Furthermore, although our work is motivated by questions arising in a medical context (see the introductory section), the proposed method can be used in other scientific branches (e.g., the Social Sciences), too. In general, we hope that our method is reasonably well to understand not only for statisticians, but also for researches who want to use it for their data analyses.
Finally, we'd like to stress the importance of thinking about links between applied mathematical statistics and official statistics. Apart from the scientific progress gained, there is also a practical reason for such considerations: As to the life expectancy comparisons, we could theoretically consider a case-control design and set up a matched control group recruited from the population as well. However, this would require substantial (monetary) efforts especially in long-term studies or in studies where quite large sample sizes are desired. So, it is a better idea to take high-quality and easy-to-access data such as the period population life tables for comparison purposes.

Acknowledgements

The author thanks Yvonne Höller, Claudia A. Granbichler and Eugen Trinka (Department of Neurology, Christian Doppler Klinik, Salzburg) as well as Arne Bathke (Department of Mathematics, Paris Lodron University, Salzburg) for many constructive discussions about this paper's content. Moreover, the author is grateful to the editor and the reviewer for their valuable suggestions, which led to an improvement of several parts of this paper.

Appendix

Recall that in section 3.1., we stated that for all $x \in \{0, 1, ..., x_{max}\}$, we have

$$E_x = \frac{1}{2} + \frac{\tilde{S}(x_{max}+1)}{2\tilde{S}(x)} + \frac{1}{\tilde{S}(x)} \sum_{k=x+1}^{x_{max}} \tilde{S}(k), \tag{4}$$

where $\tilde{S}(x) = P(X > x)$.

To prove this, we first expand E_x: According to the definitions given above, we have

$$\begin{aligned} E_x &= \frac{1}{l_x}\sum_{k=x}^{x_{max}} L_k = \frac{1}{2l_x}\sum_{k=x}^{x_{max}}(l_k + l_{k+1}) \\ &= \frac{1}{2l_x}\left(\sum_{k=x}^{x_{max}} l_k + \sum_{k=x+1}^{x_{max}+1} l_k\right) \\ &= \frac{1}{2l_x}\left(l_x + l_{x_{max}+1} + 2\sum_{k=x+1}^{x_{max}} l_k\right). \end{aligned} \tag{5}$$

Now, we use the recursive definition of l_k for $k \geq 1$ to obtain

$$l_k = l_x \prod_{j=x}^{k-1}(1-q_j), k \in \{x+1, x+2, ..., x_{max}+1\}.$$

When applying this to (5), l_x cancels out, and thus we get

$$E_x = \frac{1}{2} + \frac{1}{2}\prod_{j=x}^{x_{max}}(1-q_j) + \sum_{k=x+1}^{x_{max}}\prod_{j=x}^{k-1}(1-q_j). \tag{6}$$

Now, according to the definition of q_j, we have

$$P(X > j+1 | X \geq j) = 1 - P(j \leq X < j+1 | X \geq j) = 1 - q_j$$

for $j = 0, 1, ..., x_{max}$. Applying some basic probability theory leads to the equation

$$1 - q_j = P(X > j+1 | X \geq j) = \frac{P(X > j+1)}{P(X \geq j)} = \frac{\tilde{S}(j+1)}{\tilde{S}(j)}, j = 0, 1, ..., x_{max}.$$

Combining this result with (6) yields

$$E_x = \frac{1}{2} + \frac{1}{2}\prod_{j=x}^{x_{max}}\frac{\tilde{S}(j+1)}{\tilde{S}(j)} + \sum_{k=x+1}^{x_{max}}\prod_{j=x}^{k-1}\frac{\tilde{S}(j+1)}{\tilde{S}(j)} = \frac{1}{2} + \frac{\tilde{S}(x_{max}+1)}{2\tilde{S}(x)} + \frac{1}{\tilde{S}(x)}\sum_{k=x+1}^{x_{max}}\tilde{S}(k),$$

which completes the proof.

References

Bradshaw P, Stobie P, Knuiman M, Briffa T, Hobbs M (2015). "Life Expectancy After Implantation of a First Cardiac Permanent Pacemaker (1995–2008): A Population-Based Study." *Int J Cardiol.*, **190**, 42–46.

Carpenter J, Bithell J (2000). "Bootstrap Confidence Intervals: When, Which, What? A Practical Guide for Medical Statisticians." *Statist. Med.*, **19**, 1141–1164.

Chang K, Lu T, Lee K, Hwang J, Cheng C, Wang J (2015). "Estimation of Life Expectancy and the Expected Years of Life Lost Among Heroin Users in the Era of Opioid Substitution Treatment (OST) in Taiwan." *Drug Alcohol Depend.*

Chiang C (1979). *Life Table and Mortality Analysis*. World Health Organization. URL http://apps.who.int/iris/bitstream/10665/62916/1/15736_eng.pdf?ua=1.

DuGoff E, Canudas-Romo V, Buttorff C, Leff B, Anderson G (2014). "Multiple Chronic Conditions and Life Expectancy: a Life Table Analysis." *Med Care*, **52**(8), 688–694.

Gaitatzis A, Johnson A, Chadwick D, Shorvon S, Sander J (2004). "Life Expectancy in People with Newly Diagnosed Epilepsy." *Brain*, **127**, 2427–2432.

Guaraldi G, Cossarizza A, Franceschi C, Roverato A, Vaccher E, Tambussi G, Garlassi E, Menozzi M, Mussini C, D'Arminio Monforte A (2014). "Life Expectancy in the Immune Recovery Era: the Evolving Scenario of the HIV Epidemic in Northern Italy." *J Acquir Immune Defic Syndr.*, **65**(2), 175–181.

Hanika A, Trimmel H (2005). "Sterbetafel 2000/02 für Österreich." *Stat Nachr Osterr Stat Zent Amt*, **2**, 121–131. URL http://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bevoelkerung/sterbetafeln/index.html.

Harrell F (2015). *rms: Regression Modeling Strategies.* R package version 4.3-1., URL http://CRAN.R-project.org/package=rms.

Hubeaux S, Rufibach K (2014). *SurvRegCensCov: Weibull Regression for a Right-Censored Endpoint with Interval-Censored Covariate.* R package version 1.3, URL http://CRAN.R-project.org/package=SurvRegCensCov.

Kalbfleisch J, Prentice R (2002). *The Statistical Analysis of Failure Time Data.* John Wiley & Sons, Inc. ISBN 9780471363576.

Kardaun O (1983). "Statistical Analysis of Male Larynx-Cancer Patients–A Case Study." *Statistical Nederlandica*, **37**, 103–126.

Klein J, Moeschberger M (2003). *Survival Analysis. Techniques for Censored and Truncated Data.* Springer.

Li K, Hüsing A, Kaaks R (2014). "Lifestyle Risk Factors and Residual Life Expectancy at Age 40: a German Cohort Study." *BMC Med.*, **12**(59). URL http://www.biomedcentral.com/1741-7015/12/59.

Luangasanatip N, Hongsuwan M, Lubell Y, Limmathurotsakul D, Teparrukkul P, Chaowarat S, Day N, Graves N, Cooper B (2013). "Long-Term Survival After Intensive Care Unit Discharge in Thailand: a Retrospective Study." *Crit Care*, **17**(5). URL http://www.ccforum.com/content/17/5/R219.

Nagai M, Kuriyama S, Kakizaki M, Ohmori-Matsuda K, Sone T, Hozawa A, Kawado M, Hashimoto S, Tsuji I (2011). "Impact of Walking on Life Expectancy and Lifetime Medical Expenditure: the Ohsaki Cohort Study." *BMC Open*, **1**(2). URL http://bmjopen.bmj.com/content/1/2/bmjopen-2011-000240.full.

Nusselder W, Slockers M, Krol L, Slockers C, Looman C, van Beeck E (2013). "Mortality and Life Expectancy in Homeless Men and Women in Rotterdam: 2001âĂŞ2010." *PLoS One*, **8**(10). URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073979.

Oberaigner W, Stühlinger W (2005). "Record Linkage in the Cancer Registry of Tyrol, Austria." *Methods Inf Med*, **44**, 1–5.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Statistics Netherlands (2015). Accessed: 2015-07-31, URL http://www.cbs.nl/en-GB/menu/home/default.htm.

Strauss D, DeVivo M, Paculdo D, Shavelle R (2006). "Trends in Life Expectancy after Spinal Cord Injury." *Arch Phys Med Rehabil.*, **87**(8), 1079–1085.

Therneau T (2014). *A Package for Survival Analysis in S*. R package version 2.37-7, URL http://CRAN.R-project.org/package=survival.

Affiliation:

Georg Zimmermann
Department of Neurology, Christian Doppler Klinik
A-5020 Salzburg, Austria
Spinal Cord Injury and Tissue Regeneration Center, Paracelsus Medical University,
A-5020 Salzburg, Austria
Department of Mathematics, Paris Lodron University
A-5020 Salzburg, Austria
E-mail: georg.zimmermann@pmu.ac.at

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 45
March 2016

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2015-08-04
Accepted: 2015-11-26