# A review of methods for meta-analysis of aggregated survival data

Combescure, Christophe

UNIVERSITÉ DE GENÈVE
FACULTÉ DE MÉDECINE

HUG
Hôpitaux Universitaires de Genève

**Clinical Medicine Section**

***Department of health and community medicine***

Division of clinical epidemiology

# "A review of methods for meta-analysis of aggregated survival data"

**Thesis submitted to the Medical School of the University of Geneva**

for the degree of Privat-Docent

by

Christophe COMBESCURE

Geneva

*2014*

# Table des matières

# Abstract

In most of meta-analyses of aggregated survival data, a pooled measure of the intervention's effect is obtained by combining reported hazard ratios. Advanced statistical methods have been proposed for a better characterization of the effect. With these methods, published survival curves can be synthesized in a single summary survival curve and a variation over follow-up time of the effect can be tested. Besides the pooled intervention's effect, the synthesis of survival in each arm is helpful for the appraisal of the intervention's benefits. Moreover, some meta-analyses aim to evaluate survival in a single group. In meta-analysis of comparative studies, testing a variation of the effect over time can avoid simplistic conclusions. These methods have been published in journals of statistics and may not be accessible to a large audience. The purpose of this article is to review novel methods for meta-analysis of aggregated survival data. Distribution-free and regression methods are presented for the assessment of summary survival curves, as well as methods to obtain a pooled hazard ratio. Most methods are based on survival estimates collected at various points in time. The principle of each method, and its advantages and limitations are explained.



**Keywords**: meta-analysis, aggregated survival data, summary survival curve, hazard ratio

# Introduction

The purpose of a meta-analysis is to summarize the statistics reported in a set of studies. The pooled statistic depends on the nature of the outcome analyzed. When the outcome is a time-to-event and in presence of censored data, the analysis of individual data requires specific statistical methods for the assessment of the risk of the event or of the intervention's effect over the whole follow-up. Because of the nature of the reported statistics, meta-analyses of survival studies are often more difficult than meta-analysis of studies with no follow-up.

The most frequently reported statistics in survival studies are survival estimates, represented by survival curves to capture the variation of risk of event over the follow-up, and the hazard ratio (HR) which is a relative measure of the difference in survival between compared groups [1]. HRs from various studies can be combined easily by using the method of the inverse of variance, on condition that HRs are reported with their 95% confidence intervals or standard errors [2, 3]. Additional methods have been proposed to assess the pooled HR directly from the survival estimates. These methods are useful when HRs are not reported in all studies [2-5]. Moreover, methods based on survival estimates collected in each study at various points in time offer interesting possibilities in data analysis. For instance, the pooled HR can be assessed for various periods of follow-up and a variation over time of the intervention's effect can be tested [5, 6]. In addition to the pooled HR, the level of risk in both arms is helpful in the appraisal of the benefits of a new intervention. Of note, the CONSORT statement recommends reporting of both relative and additive measures of effect [7]. Methods have been proposed to obtain summary survival curves using a distribution-free approach [8] or regression models [9-12]. These methods can also be used to synthesize studies evaluating survival in a single group.

A review of methods for combining survival data was done in 2000 [13], but several methodologic developments have been published since. The aim of this paper is to propose a review of new statistical methods for meta-analysis of aggregated survival data. The principle of methods and data needed for their application are described. We explain how the heterogeneity is measured and how random effects are introduced in the models. The advantages and limitations are presented.

# 1) Survival estimates pooled at a single point in time

## Principle

Reporting the survival estimate at a clinically relevant point in time, denoted *t*, is a common way to present results of a survival analysis. A pooled estimate of the survival estimates can be obtained, assuming a fixed effect model, by using the method of the inverse of the variance [14]: the survival estimates are averaged and each study is weighted by the inverse of the variance of the estimate. A condition for the application of this method is the normality of the estimator. The normality is achieved when sample size is large. Transformations (logarithm, arc-sine, complementary log-log and logit transformations) can also be applied to the survival estimates to help meet this requirement (Appendix A). With the arcsine transformation, the variance is only related to the sample size [15].

## Data

The survival estimates and their standard errors are needed to apply this approach. Standard errors are rarely reported but they can be extrapolated using the formula of the standard error for a proportion by replacing the number of observations by the effective number of at-risk patients during the interval of time [0,*t*] (Appendix A). The effective number of at-risk patients is lower than the sample size because of censoring [2, 3, 5]. Details on calculation of the effective number of at-risk patients are given in Appendix B.

## Heterogeneity & Random effects

The between-study variability is measured by $I^2$ and *H* statistics and the Cochran Q test can be applied to test the null hypothesis that the survival is equal in all studies [16, 17]. The exploration of the source of heterogeneity can be performed by testing the heterogeneity between sub-groups and by using meta-regression [14, 18]. The DerSimonian and Laird's approach is applied for accounting for between-studies variability (model with random effects) [19].

## Comments

This approach is suitable when the selected point in time is clinically relevant and when the variation of risk over time is not of interest. However, in most of survival studies, data analyses aim to capture the variation of risk over time and pooling survival estimates at a single point in time appears as a poor approach. The procedure for combining survival estimates can be repeated at various points in time to obtain a pseudo-summary survival curve but several concerns arise. First, the correlations between survival estimates at various points in time are omitted. It has been shown that ignoring between-study correlations can alter the results of the meta-analysis [20]. Second, studies with a follow-up ending prior the

selected point in time are omitted with a loss of precision and a potential bias as consequences. Especially, the pseudo-summary survival curve can increase if the survival in studies with long follow-up is better than in studies with short follow-up. Characteristics of this approach are summarized in Table 1.

# 2) Distribution-free summary survival curve

## Principle

Recently, a distribution-free approach for assessing a summary survival curve has been proposed [8]. With this approach, the follow-up is partitioned in intervals of time [$t_{i-1}$, $t_i$], $i$=1 to $I$. For each interval, the arc-sine transformation is applied to the estimates of the conditional survival probabilities and the transformed estimates are combined using the method of the inverse of the variance. The pooled estimate of the survival at any time $t_i$, $i$=1,…$I$, is obtained by the product of the pooled estimates of the conditional survival until $t_i$ (denoted $\hat{p}_i^{pooled}$):

$$S^{pooled}(t_i) = \prod_{h=1}^{i} \hat{p}_h^{pooled}$$

This approach is an extension of the product-limit estimator of survival to aggregated data. Since the pooled survival estimates are a product of estimated probabilities, the obtained survival function is decreasing over time, as it should be.

## Data

The survival estimates are read off at a set of points in time ($t_0$,$t_1$,…,$t_J$) from the survival curves reported in the studies. The conditional survival estimates for the interval of time [$t_{i-1}$, $t_i$] are obtained in each study by dividing the survival estimate at time $t_i$ by the survival estimate at time $t_{i-1}$. The effective number of at-risk patients during the interval [$t_{i-1}$, $t_i$] can be derived from the numbers of at-risk patients at times $t_{i-1}$ and $t_i$, (Appendix B). If numbers of at-risk patients are not reported or are reported at points in time different from $t_1$,…,$t_I$, the effective numbers can still be extrapolated on certain conditions [2, 3, 5].

## Heterogeneity & Random effects

$I^2$ and $H$ statistics measure the between-study variability of the arcsine transformed conditional survival estimates [21]. In contrast to meta-analysis of survival at a single point in time, the homogeneity assumption is that the conditional survival probabilities are equal in the studies for any time $t_i$. The Cochran Q test can also be applied to test the null hypothesis of homogeneity. The rejection of the null hypothesis means that the conditional survival probabilities are different across studies at least for one point in time. Potential heterogeneity factor can be explored by testing the between-strata heterogeneity, but meta-regression methods are not applicable. When a model with random effects is assumed, the between-study covariance between the estimates of conditional survival has to be accounted for.

Therefore, the extension of DerSimonian and Laird's approach to multiple outcomes is applied to obtain the pooled conditional survival estimates [22].

### Comments

This approach allows the assessment of a summary survival curve by pooling the survival estimates at various points in time. The mean (or restricted mean when survival curve are incomplete) and the median survival times can be obtained from the summary curve. The caveat is the selection of points in time at which survival estimates are collected: the points in time should be abundant enough to capture the shape of the survival curve but they should be spaced enough to avoid conditional survival probabilities close to 1 because a bias can occur in this situation [8]. With this method, the advantage is that the summary survival curve is distribution-free and decreasing over time. Moreover, studies with various lengths of follow-up can be combined. The estimation of the pooled survival probability at time *t* also involves all the studies ending before *t* because the latter contribute to the estimation of conditional survival probabilities for intervals of time prior to *t*. Characteristics of this approach are summarized in Table 1.

## 3) Regression models

### Principle

The use of regression models to estimate simultaneously pooled survival at various points in time, accounting for the within-study correlations, has been proposed initially by Dear et al [23]. This approach has been expanded by Arends et al to account for between-study variability (mixed effects models) and to ensure that the pooled survival probabilities fall between 0 and 1 [9]. In the latter approach, a transformation is applied to the survival estimates and the transformed survival estimates are linearly modeled with a set of predictors including follow-up time. With the complementary log-log transformation, the regression model is equivalent to a Weibull survival model:

$$\text{Ln}[-\ln(S_k)]=\alpha+\beta \ln(t)+\varepsilon_k,\ \ \varepsilon_k \sim N(0,V_k)$$

where $S_k$ is the vector of the survival estimates at all points in time and $V_k$ the within-study correlation matrix of the survival estimates in study *k*.

### Data

The approach proposed by Arends et al requires estimates of the survival probabilities at various points in time and their standard errors [9].

### Heterogeneity & Random effects

Between-study variability is introduced in the regression model by adding random effects:

$$\text{Ln}[-\ln(S_k)]=\alpha+\beta \ln(t)+a+b \ln(t)+\varepsilon_k,\ \ \varepsilon_k \sim N(0,V_k)\ \text{ and } (a,b)\sim N(0,D) \qquad (1)$$

where $a$ and $b$ are the random effects with expectation 0 and $D$ is the between-study covariance matrix. The amount of heterogeneity has to be interpreted from the size of the random effects. Potential heterogeneity factors, both continuous and categorical, can be added in the model to be tested.

### Comments

This approach allows the estimation of a summary survival curve and the combination of studies with various lengths of follow-up. Its advantage is the possibility to conduct multivariate meta-regression, by adding study-related fixed effects into the model above (regression models are not limited to the model presented above). Other transformations for the survival estimates can be applied and any function of follow-up time can be introduced as a predictor. However, with complex models, there is a risk of overfitting and of obtaining a summary survival curve that is locally increasing. For instance, using splines for the follow-up time, the right-tail of the summary survival curve can be increasing if survival in studies with long follow-up is higher than in studies with short follow-up. Moreover, because the number of studies is usually low in meta-analyses, verification of the goodness-of-it may be difficult. Characteristics of this approach are summarized in Table 1.

## 4) Combining hazard ratios reported with their 95% confidence intervals

### Principle

When hazard ratios (HRs) are reported with their 95% confidence interval in all studies included in the meta-analysis, they can be combined using the method inverse of variance [2, 3]. A pooled estimate of the logarithm of HR is obtained by the averaged logarithm of estimated HR weighted by the inverse of the variance of the estimates. The variance is deduced from the lower and upper bounds of the 95% confidence intervals, for instance by dividing the range from the logarithm of the lower bound to the logarithm of the upper bound by 2*1.96 and squaring the result.

### Data

Estimated HRs and their 95% confidence intervals from each study.

### Heterogeneity & Random effects

As for meta-analysis of survival estimates at a single point in time, the amount of heterogeneity is investigated with $I^2$ and $H$ statistics and the Cochran Q test [16, 17]. The sources of heterogeneity are explored using sub-groups comparison and meta-regression

[14, 18]. The DerSimonian and Laird's approach is applied for models with random effects [19].

### Comments

With this approach, the advantage is that censored data within each study are already taken into account in the estimate of HRs and the 95% confidence intervals. The extraction of survival estimates and calculation of effective numbers of at-risk patients are not needed. This approach does not allow an exploration of a possible variation over the follow-up of the intervention's effect. Characteristics of this approach are summarized in Table 2.

## 5) Combining hazard ratios using survival estimates at a single time point

### Principle

When a HR is not reported, it is still possible to obtain an estimate from survival estimates at a single point in time since the ratio of the logarithm of survival, denoted RLS, is a natural estimator of the HR [24]:

$$RLS(t) = \frac{\ln\ S_{Int}(t)}{\ln\ S_{Con}(t)}$$

where $S_{Int}(.)$ and $S_{Con}(.)$ are the survival functions in intervention and control arms respectively. Since the logarithm of the estimator of the RLS follows asymptotically a normal distribution, the method of the inverse of the variance can be used to combined RLS($t$) from various studies [4, 25]. The variance of the logarithm of RLS($t$) is:

$$Var[lnRLS(t)] = \frac{S_{Con}(t)}{N_{Con}S_{Con}(t)[lnS_C(t)]^2} + \frac{S_{Int}(t)}{N_{Int}S_{Int}(t)[lnS_{Int}(t)]^2}$$

where $N_{Con}$ and $N_{Int}$ are the effective numbers of at-risk patients during the interval of time [0,$t$] in both arms (control and intervention).

### Data

The survival estimates and the effective numbers of at-risk patients during the interval of time [0,$t$] are needed to apply this approach (Appendix B).

### Heterogeneity & Random effects

As for meta-analysis of combination of HRs reported with their 95% confidence intervals, the amount of heterogeneity is investigated with $I^2$ and $H$ statistics and the Cochran Q test [16, 17]. The sources of heterogeneity are explored using sub-groups comparison and meta-

regression [14, 18]. The DerSimonian and Laird's approach is applied for models with random effects [19].

### Comments

With this approach, hazards are assumed proportional. Since HRs are not reported in studies, this assumption was likely not checked by the authors. The log cumulative hazard, obtained from survival estimates extracted at various points in time from published survival curve, can be plotted versus time to verify the proportionality of hazards for each study [5]. Even if hazards are approximately proportional, the estimate of RLS(*t*) can vary according to the point in time because of a lack of regularity in survival curves, especially when the size of steps of the curves is great. Characteristics of this approach are summarized in Table 2.

## 6) Combining hazard ratios using survival estimates at multiple points in time (1)

### Principle

Williamson et al proposed a method to test whether an intervention's effect is constant over time from aggregated data [5]. The principle is to split the follow-up into several intervals of time and to extract the HR for each interval and for each study, denoted $HR_{k,i}$ for the study *k* and the interval of time $[t_{i-1},t_i]$. By applying the method of the inverse of the variance to HRs, pooled estimates of the intervention's are obtained for all intervals of time, denoted $HR_{pooled,i}$. The hypothesis of homogeneity of the intervention's effect across intervals of time is tested by a Cochran Q test. The tested null hypothesis is $H_0$:$\log(HR_{pooled,1})= \log(HR_{pooled,2})=..= \log(HR_{pooled,I})=\theta$ where *I* is the number of intervals and $\theta$ is the logarithm of the overall pooled HR. Under this null hypothesis, the intervention's effect is constant over time and an overall HR is assessed for each study *k* by combining the $HR_{k,i}$ of the various intervals of time *i*=1,..,*I* (here also using the method of the inverse of the variance), denoted $HR_{k,pooled}$. The pooled overall HR, denoted $\theta$, is then the combination of $HR_{k,pooled}$ from all studies and the statistic *Q* of the Cochrane test is:

$$Q = \sum_{i=1}^{I} w_i [lnHR_{pooled,i} - \theta]^2$$

where $w_i$ is the inverse of the variance of ln $HR_{pooled,i}$. Under the null hypothesis, the statistic *Q* follows a Chi-squared distribution with *I*-1 degrees of freedom.

### Data

Data needed for this approach are the estimates of HR for each interval of times and theirs standard errors. Williamson et al explained how to approximate the estimates of HRs from published survival curves [5]: the effective numbers of at-risk patients and the numbers of

events during intervals of time are derived from the published survival curves and reported number of patients still followed at various points in time (see Appendix B). Alternatively, the standard errors can be extrapolated from the p-value of a log-rank test [2].

### Heterogeneity & Random effects

The between-study variability is measured within each interval of time by $I^2$ or $H$ statistics and the between-intervals variability of pooled HRs is tested by a Cochrane Q test. With this approach, models with fixed effects are assumed and the between-study heterogeneity is ignored.

### Comments

This approach provides an easy way to test a variation of the intervention's effect over time. If not accounted for, such variation can be a source of between-study heterogeneity when studies have different lengths of follow-up. When the effect is constant over time, the estimate of the pooled overall HR is more accurate than the combination of RLSs because the HR for each study, which is an average of HRs of various intervals of time, is less sensitive to the selection of points in time than the RLS, which assessed at a single point in time. The approach proposed by Williamson et al is suitable when there is little between-study variability because models with fixed effect are assumed. The pooled HRs for the various intervals of time are independent when models with fixed effect are assumed, but not when models with random effects are assumed: random effects can be correlated within a study. For instance, a study with a greater intervention's effect than other studies will produce HRs higher than those of other studies for any interval of time. If between-study variability is observed, models for multiple outcomes assuming random effects should be used [21, 22]. Characteristics of this approach are summarized in Table 2.

## 7) Combining hazard ratios using survival estimates at multiple points in time (2)

### Principle

This approach is an extension of the combination of RLS($t$) at single point in time to several points in time [6]. The logarithm of RLS collected at various points in time $t_i$, $i$=1,…,$I$, are modeled as a linear function of follow-up time, for instance the logarithm of time:

$$\text{Ln RLS}_k = \alpha + \beta \ln(t_k) + \varepsilon_k,\ \varepsilon_k \sim N(0, V_k)$$

where $RLS_k$ is the vector of RLSs in a study $k$ for all times $t_i$ and $V_k$ is the within-study covariance matrix of $RLS_k$. Alternatively, the predictor can be a piecewise constant function of time:

$$Ln\ RLS_k = \sum_{i=1}^{I} \alpha_i I_{t\in[t_0,t_i]} + \varepsilon_k$$

In both models, the hypothesis that the RLS is constant over time can be tested, and if necessary, rejected. If the RLS is assumed constant over time, it is modeled only by the intercept.

### Data

The survival estimates and the effective numbers of at-risk patients during the intervals of time $[t_{i-1},t_i]$ are needed to apply this approach.

### Heterogeneity & Random effects

Potential heterogeneity factors can be added as predictors in the models and tested. Assuming a fixed effect model, the estimates of regression coefficients are obtained with the generalized least squared method. A goodness-of-fit test, based on residuals, has been proposed by Combescure et al [6]. Because discrepancies between observed and predicted RLSs are caused not only by the between-study variability but also by an incorrectly specified time-effect, this test cannot be interpreted as a test for detection of heterogeneity. Models with random effects can also be used: a between-study covariance matrix is introduced and the regression coefficients are estimated by maximum likelihood method [26, 27].

### Comments

Like the previous approach, this approach provides a test for detecting a variation of the intervention's effect over time. When the effect varies over time, the pooled estimate of RLS at the point in time $t$ is a value of the HR averaged over $[0,t]$ [4, 6, 28]. With this approach, it is possible to introduce potential heterogeneity factors as predictors in the models. When models with fixed effect are used, studies with various lengths of follow-up can be combined. In contrast, for models with random effects, the latest point in time $t_l$ must be the same in all studies. Studies with short follow-up have to be excluded from the analysis or the right tail of survival curves may be ignored in some studies. A code for S-plus (S-plus 8.0 for Windows, Insightful Corp., Seattle, USA) is available from the first author of Combescure et al [6]. Characteristics of this approach are summarized in Table 2.

## 8) Regression models for intervention's effect

### Principle

The intervention's effect is assessed in the regression model described in Equation (1) by adding the arm as covariate (variable taking 0 for reference arm and 1 for intervention arm):

$$\text{Ln}[-\ln(S_k)]=\alpha+\beta \ln(t)+\gamma \text{Arm}+\varepsilon_k,\ \ \varepsilon_k \sim N(0,V_k)$$

The within-study covariance matrix $V_k$ is block diagonal (one block per arm). In this model, the exponential of regression coefficients $\gamma$ is an estimate of the HR because the model corresponds to a Weibull survival model. An interaction term between logarithm of time and arm can be tested to explore a variation of the intervention's effect over time. Alternatively, other functions of time can be introduced as predictor (for instance a piecewise constant function) and other transformations can be applied to the survival estimates.

### Data

Needed data are the same than for regression models with single arm studies.

### Heterogeneity & Random effects

The modification of the intervention's effect caused by a factor can be tested by adding an interaction term between this factor and the arm. Between-study variability is introduced in the regression model by adding random effects:

$$\text{Ln}[-\ln(S_k)]=\alpha+\beta \ln(t)+\gamma\text{Arm}+a+b \ln(t)+c \text{ Arm}+\varepsilon_k,\ \ \varepsilon_k \sim N(0,V_k) \text{ and } (a,b,c)\sim N(0,D)$$

where $a$, $b$ and $c$ are the random effects and $D$ is the between-study covariance matrix.

### Comments

The estimation of regression coefficients provides not only an estimate of the intervention's effect but also summary survival curves in both arms. The variation of the effect over time can be tested. As in the previous approach, additional predictors can be introduced in the regression model. The main difference compared to the previous approach is that the dependent variable is the survival and not the intervention's effect. Therefore, an assumption on the shape of survival has to be made (for instance a Weibull survival model). Moreover, study-level factors modifying the survival, but not the intervention's effect, should not be ignored. Characteristics of this approach are summarized in Table 2.

## Discussion

Various methods have been proposed to combine aggregated data from survival studies. Each method for combining aggregated data has advantages and limitations (Tables 1 and 2). The choice of the method by the analyst depends also on the statistics reported in the studies. A review of randomized clinical trials with survival endpoints in oncology showed that hazard ratios were reported in 65% of articles [29]. These findings suggest that combining only the HRs reported with their 95% confidence intervals can be a source of bias because of a selection of studies that can be pooled. The methods to assess a pooled HR from survival estimates circumvent this concern. Some of these methods allow testing a variation of the

intervention's effect over time. Besides the pooled HR, which is a relative measure of the intervention's effect, summary survival curves provide information about the risk of the event in both arms. The risk in the reference arm is useful to appraise the benefit of a new intervention.

A major concern in the meta-analysis of survival studies is to take censored data into consideration because the weight allocated to a study depends on the standard error of the estimate to be pooled, which in turn depends on the number of participants that is analyzed over time. Ignoring censorship (for instance by taking the initial sample size instead of the effective number of at-risk patients) leads to an underestimation of the standard error of the pooled estimate. When methods that require the number of at-risk patients are used, suitable approaches have to be applied to infer the effective numbers of at-risk patients at different points in time [2, 3, 5, 30]. User-friendly spreadsheets have been proposed for this purpose [3, 31].

The complexity of some methods discussed in this paper limits their use in practice. In particular, statistical software dedicated to meta-analyses do not offer the possibility of analysis of aggregated survival data by regression models and, more generally, of analysis of survival estimates collected at various points in time. However, a R package (MetaSurv) is available to assess a distribution-free summary survival curve and some codes are available upon request from the authors the methods [6, 8, 9].

The proposed review of methods is not exhaustive. Additional regression models have been proposed. Models proposed by Ouwens et al [12] are an extension of the regression models proposed by Arends et al [9] to network meta-analyses and are not reviewed in this paper. In a different approach, Fiocco et al modeled the counts of events in various intervals of time by a Poisson process [10, 11]. This approach, that requires advanced skills in statistics and computation, is not reviewed here. For readers interested by regression models, the performance of several models has been compared by Fiocco et al [32]. Guyot et al proposed an algorithm to extrapolate individual survival data from published survival curves [33] to which they applied methods for meta-analysis of individual survival data [34]. These methods specific to individual data have already been reviewed by Katashian et al [34]. In our review, we focused on methods for assessing summary survival curves and pooled HRs from survival estimates reported in studies. Some of the methods presented in the review may also be used to combine odds ratios or risk ratios as an alternative to HR when hazards are not proportional [6, 9]. Methods for combining the median and the ratio of median survival times have also been proposed [35, 36], but they have not been review because they can be used only when survival is below 50%. A perspective would be to expand meta-analysis methods to other statistics, for instance to the restricted *C* index when two arms are compared [37] or to time-dependent ROC curves when the predictor is continuous [38, 39].

The recently developed statistical methods for meta-analysis of survival data are valuable tools for an accurate synthesis of combined studies. Authors of meta-analyses should be aware of these new methods. Our review aims to present the principle, advantages and limitations of methods in a way that is accessible to non statisticians and to promote their uses.

# Appendix A: Transformations applicable to proportions

| Transformation F(.) | F(S($t$)) | Asymptotic variance of F(S($t$)) |
|---|---|---|
| Untransformed | S($t$) | S($t$)[1- S($t$)]/ $N_{eff}$ |
| Arc sine | Arcsin[$\sqrt{S(t)}$] | 1/(4 $N_{eff}$) |
| Logarithm | Ln S($t$) | [1- S($t$)]/[S($t$)$N_{eff}$] |
| Complementary log-log | Ln[-ln S($t$)] | [1- S($t$)]/[$N_{eff}$ $\ln^2$(S($t$))] |
| Logit p | Ln[S($t$)/(1- S($t$))] | 1/[S($t$) $N_{eff}$]+1/[$N_{eff}$ - S($t$)$N_{eff}$] |

S($t$) is the survival at time $t$ and $N_{eff}$ is the effective number of at-risk patients during the interval of time [0,$t$] (see Appendix B).

# Appendix B: Extraction of data from published survival curves

The survival estimates are not systematically reported but they can be read off from published survival curves using software (for instance R package Digitize or DigitizeIt, a commercial software). With the R package Digitize, the digitalized picture of the survival curve (format .jpg) is loaded in R. Then, points of reference for the scale of the follow-up time (horizontal axis) and for survival (vertical axis) are selected by clicking on the picture. In the last step, the user clicks on points of the survival curve and their coordinates (time and survival) are given with respect to the scales indicated by the user.

Most of methods presented in the review need not only survival estimates but also the effective numbers of at-risk patients because the variance of the estimates is needed to determine the weight of each study. The effective numbers of at-risk are not the number reported on survival curves, rather the number of patients at-risk during a time interval. Williamson et al proposed formula to obtain this effective number for an interval of time $[t_{i-1},t_i]$, denoted $N_{eff,i}$ when the number of patients still followed at $t_{i-1}$ and $t_i$, denoted $N_{i-1}$ and $N_i$ respectively, are reported:

$$N_{eff,i} = (N_{i-1}+N_i)\ S(t_{i-1})/[S(t_{i-1})+S(t_i)]$$

This approach assumes that the censorship is uniform over the interval of time $[t_{i-1},t_i]$. When survival estimates are pooled at a single time point, denoted $t_1$, the previous equation can be simplified because S($t_{i-1}$) equal 1 and $N_{i-1}$ is the sample size:

$$N_{eff} = (N+N_1)/[1+S(t_1)]$$

The transformed survival estimates are combined using the method of the variance or the DerSimonian and Laird's methodology with the variance given in Appendix A. If the numbers of at-risk patients are reported for other times than $t_i$, it is still possible to extrapolate the effective number of at-risk patients. User friendly Excel Spreadsheets are available [3, 31].

# References


1. Spruance SL, Reid JE, Grace M, and Samore M. Hazard ratio in clinical trials. *Antimicrob Agents Chemother* 2004; **48**(8): 2787-2792.
2. Parmar MK, Torri V, and Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine* 1998; **17**(24): 2815-2834.
3. Tierney JF, Stewart LA, Ghersi D, Burdett S, and Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007; **8**(16.
4. Moodie PF, Nelson NA, and Koch GG. A non-parametric procedure for evaluating treatment effect in the meta-analysis of survival data. *Statistics in Medicine* 2004; **23**(7): 1075-1093.
5. Williamson PR, Smith CT, Hutton JL, and Marson AG. Aggregate data meta-analysis with time-to-event outcomes. *Statistics in Medicine* 2002; **21**(22): 3337-3351.
6. Combescure C, Courvoisier D, Haller G, and Perneger T. Meta-analysis of two-arm studies: Modeling the intervention effect from survival probabilities. *Stat Methods Med Res* In press.
7. Schulz KF, Altman DG, and Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med*; **152**(11): 726-732.
8. Combescure C, Foucher Y, and Jackson D. Meta-analysis of single-arm survival studies: a distribution-free approach for estimating summary survival curves with random effects. *Stat Med* In press.
9. Arends LR, Hunink MG, and Stijnen T. Meta-analysis of summary survival curve data. *Statistics in Medicine* 2008; **27**(22): 4381-4396.
10. Fiocco M, Putter H, and van Houwelingen JC. Meta-analysis of pairs of survival curves under heterogeneity: a Poisson correlated gamma-frailty approach. *Statistics in Medicine* 2009; **28**(30): 3782-3797.
11. Fiocco M, Putter H, and Van Houwelingen JC. A new serially correlated gamma-frailty process for longitudinal count data. *Biostatistics* 2009; **10**(2): 245-257.
12. Ouwens MJ, Philips Z, and Jansen JP. Network meta-analysis of parametric survival curves. *Research Synthesis Methods* 2010; **21**(22): 3337-3351.
13. Earle CC, Pham B, and Wells GA. An assessment of methods to combine published survival curves. *Medical Decision Making* 2000; **20**(1): 104-111.
14. Borenstein M, Hedges LV, Higgins JPT, and Rothstein HR, *Introduction to Meta-analysis*. 2008, Chichester (UK): John Wiley & Sons.
15. Rucker G, Schwarzer G, Carpenter J, and Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine* 2009; **28**(5): 721-738.
16. Higgins JP and Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**(11): 1539-1558.
17. Higgins JP, Thompson SG, Deeks JJ, and Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; **327**(7414): 557-560.
18. Thompson SG and Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* 1999; **18**(20): 2693-2708.
19. DerSimonian R and Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3): 177-188.
20. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. *J. R. Statist. Soc. A* 2009; **172**(4): 789-811.
21. Jackson D, White IR, and Riley RD. Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine* 2012; **31**(29): 3805-3820.
22. Jackson D, White IR, and Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine* 2010; **29**(12): 1282-1297.
23. Dear KB. Iterative generalized least squares for meta-analysis of survival data at multiple times. *Biometrics* 1994; **50**(4): 989-1002.

24. Perneger TV. Estimating the relative hazard by the ratio of logarithms of event-free proportions. *Contemp Clin Trials* 2008; **29**(5): 762-766.
25. Combescure C, Courvoisier DS, Haller G, and Perneger TV. Meta-analysis of binary outcomes from two-by-two tables when the length of follow-up varies and hazards are proportional. *Stat Methods Med Res* 2011; **20**(5): 531-540.
26. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, and Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Stat Med* 1998; **17**(22): 2537-2550.
27. Trikalinos TA and Olkin I. A method for the meta-analysis of mutually exclusive binary outcomes. *Stat Med* 2008; **27**(21): 4279-4300.
28. Yang S and Prentice RL. Estimation of the 2-sample hazard ratio function using a semiparametric model. *Biostatistics* 2011; **12**(2): 354-368.
29. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, and Kramar A. Survival end point reporting in randomized cancer clinical trials: a review of major journals. *J Clin Oncol* 2008; **26**(22): 3721-3726.
30. Duchateau L, Collette L, Sylvester R, and Pignon JP. Estimating number of events from the Kaplan-Meier curve for incorporation in a literature-based meta-analysis: what you don't see you can't get! *Biometrics* 2000; **56**(3): 886-892.
31. Hoyle MW and Henley W. Improved curve fits to summary survival data: application to economic evaluation of health technologies. *BMC Med Res Methodol* 2011; **11**(139.
32. Fiocco M, Stijnen T, and Putter H. Meta-analysis of time-to-event outcomes using a hazard-based approach: Comparison with other models, robustness and meta-regression. *Computational Statistics and Data Analysis* 2012; **56**(1028-1037.
33. Guyot P, Ades AE, Ouwens MJ, and Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology* 2012; **12**(9.
34. Katsahian S, Latouche A, Mary JY, Chevret S, and Porcher R. Practical methodology of meta-analysis of individual patient data using a survival outcome. *Contemporary Clinical Trials* 2008; **29**(2): 220-230.
35. Michiels S, Piedbois P, Burdett S, Syz N, Stewart L, and Pignon JP. Meta-analysis when only the median survival times are known: a comparison with individual patient data results. *Int J Technol Assess Health Care* 2005; **21**(1): 119-125.
36. Zang J, Xiang C, and He J. Synthesis of median survival time in meta-analysis. *Epidemiology* 2013; **24**(2): 337-338.
37. Combescure C, Perneger TV, Weber DC, Daures JP, and Foucher Y. Prognostic ROC curves: a method for representing the overall discriminative capacity of binary markers with right-censored time-to-event endpoints. *Epidemiology* 2014; **25**(1): 103-109.
38. Combescure C, Daures J, and Foucher Y. A literature-based approach to evaluate the predictive capacity of a marker using time-dependent summary receiver operating characteristics. *Stat Methods Med Res* In press.
39. Heagerty PJ, Lumley T, and Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**(2): 337-344.

## Tables

Table 1: Comparison of methods for meta-analysis of aggregated survival data from a single arm

| | **Survival estimates pooled at a single point in time *t*** | **Distribution-free summary survival curve** | **Regression models** |
|---|---|---|---|
| ***Summary survival curve*** | No | Yes | Yes |
| ***Assumption on the survival function*** | No | No | Yes |
| ***Combining studies with various length of follow-up*** | Studies ending before t are omitted | Yes | Yes |
| ***Statistics for heterogeneity*** | $I^2$, $H$ | $I^2$, $H$ | Size of the random effects |
| ***Fixed effect*** | Method of the inverse of the variance | Method of the inverse of the variance | Generalized least squared |
| ***Random effects*** | DerSimonian and Laird's methodology | Extension of DerSimonian and Laird's methodology to multiple outcome | Mixed-effects models |
| ***Data needed*** | Survival estimates at time point t and standard errors or effective numbers of at-risk patients during [0,*t*] | Survival estimates at various time points and effective numbers of at-risk patients during each interval of time | Survival estimates at various times and their standard errors |
| ***Transformation applied to survival estimates*** | Arc-sine, but other transformations are applicable (logit, log, log minus | Arc-sine | Log minus log, but other transformations are applicable |

| | | | |
|---|---|---|---|
| | log…) | | |
| ***Within-study correlation*** | | Taken into account | Taken into account |
| ***Software*** | Generic method of the inverse of variance (available in most of software) | R package available MetaSurv | SAS code available from first author of Arends et al upon request |
| ***Pros*** | -Easy to compute<br>-Statistics for heterogeneity | -Distribution-free summary survival curve<br>-Summary survival decreases over time<br>-Statistics for heterogeneity<br>-R package available | -Summary survival curve<br>-Possibility to introduce several potential heterogeneity factors (categorical and continuous) in models<br>-Large panel of models |
| ***Cons*** | -Variation of risk over time not captured<br>-Approach meaningful only of the selected time point is clinically relevant<br>- Repeating the estimation at several points in time is not recommended because within-study correlations are omitted and survival can be increasing over time | -No meta-regression<br>-Approach biased when events are rare and sample sizes are small | -Fit of model difficult to check because number of studies is often low<br>-No clear measure of heterogeneity<br>-Code available upon request<br>-Complex models (using splines for instance) may fit well data but with the risk that the survival increases over time |

Table 2: Comparison of methods for meta-analysis of aggregated survival data from comparative studies

| | **Combining HRs reported with their 95% CIs** | **Combining HRs using survival estimates at a single point in time** | **Combining HRs using survival estimates at multiple points in time (1)** | **Combining HRs using survival estimates at multiple points in time (2)** | **Regression model for intervention's effect** |
|---|---|---|---|---|---|
| ***Summary survival curve*** | No | No | No | No | Yes |
| ***Assumption on the survival functions*** | Proportionality of hazards | Proportionality of hazards | Proportionality of hazards | Proportionality of hazards | Yes |
| ***Combining studies with various length of follow-up*** | Yes | Only if the effect is assumed constant over time | Yes | Only with fixed effect models | Yes |
| ***Statistics for heterogeneity*** | $I^2$, $H$ | $I^2$, $H$ | $I^2$, $H$ | No | Size of the random effects |
| ***Fixed effect*** | Method of the inverse of the variance | Method of the inverse of the variance | Method of the inverse of the variance | Generalized least squared | Generalized least squared |
| ***Random effects*** | DerSimonian and Laird's methodology | DerSimonian and Laird's methodology | No | Maximum likelihood method | Mixed-effects models |
| ***Data needed*** | HRs and 95% confidence intervals | Survival estimates at time $t$ and effective numbers of at-risk patients during [0,$t$] | Survival estimates at various points in time and their standard errors | Survival estimates at time $t$ and effective numbers of at-risk patients during [0,$t$] | Survival estimates at various points in time and their standard errors |

| | | | | | |
|---|---|---|---|---|---|
| ***Transformation applied to survival estimates*** | No | No | No | No | Log minus log, but other transformations are applicable |
| ***Within-study correlation*** | | | No need | Yes | Accounted |
| ***Software*** | Generic method of the inverse of variance (available in most of software) | Generic method of the inverse of variance (available in most of software) | Generic method of the inverse of variance (available in most of software) | S-plus/R code available from first author of Combescure et al upon request | SAS code available from first author of Arends et al upon request |
| ***Pros*** | -Easy to compute<br>-Statistics for heterogeneity<br>-Study-level factors modifying the survival, but not the intervention's effect, can be ignored | -Easy to compute<br>-Statistics for heterogeneity<br>-All studies reporting survival estimates at time t are combined<br>-Study-level factors modifying the survival, but not the intervention's effect, can be ignored | -Easy to compute<br>-Statistics for heterogeneity<br>-Equality of pooled HRs from various intervals of time can be tested<br>-Study-level factors modifying the survival, but not the intervention's effect, can be ignored | -A variation of the effect over time can be tested<br>-Study-level factors modifying the survival, but not the intervention's effect, can be ignored | -Summary survival curves in both arms<br>-Possibility to introduce several potential heterogeneity factors (categorical and continuous) in models<br>-Large panel of models |
| ***Cons*** | -Studies without HR and 95% CI are omitted in | -RLS(t) may be an inaccurate estimate of | - only models with fixed effect | -Studies with various length of follow-up | -Fit of model difficult to check because |

| | | | | | |
|---|---|---|---|---|---|
| | the analysis<br>-A variation over time of effect cannot be tested | HR<br>-A variation of the effect over time cannot be tested | | can be combined only with fixed effect models<br>-No clear measure of heterogeneity<br>-Code available upon request | number of studies is often low<br>-No clear measure of heterogeneity<br>-Code available upon request<br>-Complex models (using splines for instance) may fit well data but with the risk that the survival increases over time<br>- study-level factors modifying the survival, but not the intervention's effect, cannot be ignored |