Issues of the hazard ratio estimate and application of the restricted mean survival time to a non-inferiority study

Miki Horiguchi, Kyongsun Pak, Masashi Mikami, and Masahiro Takeuchi

Department of Clinical Medicine (Biostatistics), School of Pharmacy, Kitasato University

Abstract

In a randomized clinical trial with a right-censored time-to-event outcome, the hazard ratio by the Cox proportional hazards model is conventionally used. When the ratio of the two hazard functions is approximately constant overtime, the proportional hazards assumption is satisfied and the Cox model is useful to quantify the between-group difference. However, when the proportional hazards assumption is not satisfied, the hazard ratio changes over time making assessment of the between-group difference difficult. In addition, clinical interpretation of the hazard ratio is difficult regardless of the adequacy of the proportional hazards assumption. In this paper, we demonstrate one of the critical issues of the hazard ratio estimate by a numerical study. Model-free measures based on the restricted mean survival time (RMST), e.g., the difference of the RMSTs, are practically useful alternatives. We present secondary analysis results of a recent non-inferiority study to compare gefitinib and docetaxel in patients with advanced/metastatic non-small-cell lung cancer (NSCLC) who had failed one or two chemotherapy regimens, using the RMST. RMST-based measures are robust and provide clinically interpretable results. We recommend using RMST-based measures to quantify the between-group difference.

1. Introduction

In a clinical trial comparing two groups, the primary endpoint is often time-to-event such as overall survival or progression free survival. When overall survival is the primary endpoint, we often want to compare the distributions of survival times. We first estimate the two survival curves by the Kaplan-Meier (KM) estimate and then test the null hypothesis that two survival curves are identical by using log-rank test. Finally the hazard ratio by the Cox model is conventionally reported. The hazard ratio is the most famous measure for the between-group difference in survival analysis but it is difficult to clinically interpret. When the proportional hazards assumption for the Cox model is satisfied, the hazard ratio is a valid measure. When the proportional hazards assumption is not satisfied, the hazard ratio changes over time indicating it would not be an appropriate measure quantifying the between-group difference. Since the proportional hazards assumption is unlikely satisfied in practice, using the hazard ratio estimate in a clinical trial with the survival endpoint is questionable. Model-free measures such as measures based on the restricted mean survival time (RMST) have been studied in several papers[1-3]. The difference (or ratio) of the RMSTs can be one of the practically useful alternatives to the hazard ratio.

This paper has two main purposes; the first purpose is to demonstrate one of the critical issues of the hazard ratio estimate (i.e., dependence of underlying study-specific censoring distributions) via a numerical study and the second purpose is to illustrate the usefulness of RMST-based measures as alternatives to the hazard ratio with the data from a previously analyzed clinical trial with survival as the endpoint. In section 2, issues of the hazard ratio estimate with a numerical study and an example a non-inferiority study are shown. Section 3 introduces model-free measures for the between-group difference and section 4 applies RMST-based analysis to the example of the non-inferiority study from section 2. Section 5 is discussion, followed by conclusion (section 6).

We used the R version 3.2.0 and the survRM2 library for the analysis, which is available from the CRAN (http://cran.r-project.org).

2. Hazard ratio estimate: Issues

2.1) No reference value

The hazard ratio is difficult to clinically interpret. When a reported hazard ratio (treatment group over control group) is 0.8, the treatment group is better than the control group since the ratio is smaller than one. However, a reduction of 0.2 cannot be clinically interpreted, since there is not absolute hazard as a reference.

2.2) Wide confidence intervals

When there are few events, the confidence interval of the estimated hazard ratio is wide. For example, a clinical trial enrolled a hundred thousand patients to evaluate the between-group difference, and each of the two groups had only one event. It is intuitively clear that there is no difference between the two groups. On the other hand, the resulting 95% confidence interval for the hazard ratio estimate could be 0.1 to 10, making it difficult to conclude that the two groups are equivalent.

2.3) Dependence of underlying study-specific censoring distributions

The Cox proportional hazards model is a commonly used statistical model to estimate the hazard ratio. In this model, the hazard function for a subject with a k-dimensional covariate vector, $z_i = [z_{i1}\ z_{i2}\ z_{i3}\ \cdots\ z_{ik}]^T$, is modeled as

$$h(t; z_i) = h_0(t) \cdot \exp(\beta^T z_i) = h_0(t) \cdot \exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \cdots + \beta_k z_{ik}), \tag{1}$$

where i $(i = 1, 2, 3, \dots, n)$ denotes the number of observations, $\beta = [\beta_1\ \beta_2\ \beta_3\ \cdots\ \beta_k]^T$ is the parameter to be estimated, and $h_0(t)$ is the baseline hazard function which is the probability of the event when all of the covariates are zero[4]. If we consider two observations z_i and $z_{i'}$, the hazard ratio for these subjects using (1) is

$$\frac{h(t; z_i)}{h(t; z_{i'})} = \frac{h_0(t) \cdot \exp(\beta^T z_i)}{h_0(t) \cdot \exp(\beta^T z_{i'})} = \frac{\exp(\beta^T z_i)}{\exp(\beta^T z_{i'})}.$$

The above is independent of t and is the proportional hazards (PH) assumption. When the PH assumption is not satisfied, the estimated hazard ratio would not be an appropriate measure for the between-group difference since it is not simply an average of the true hazard ratio over time. Furthermore, the parameter depends on underlying study-specific censoring distributions, which will be illustrated via a numerical study below.

Figure 1A (left) shows the survival functions of event time for the two groups using the Weibull distribution with shape and scale parameters 1 and 50 for the treatment group and 1 and 40 for the control group. Since the shape parameters of the two groups are the same, the PH assumption is satisfied. In Figure 1B (left) the survival curves were drawn using the Weibull distribution with shape and scale parameters 2 and 35 for the treatment group and 1 and 30 for the control group. The shape parameters are not the same in this setting thus the PH assumption is not satisfied (Non-PH). On the right-hand side of each figure, the corresponding hazard ratios over time are drawn. For the survival functions of censoring time from the Weibull distribution, three patterns of the shape and scale parameters were considered for the treatment group: $(3, 12)$; $(3, 18)$; and $(3, 24)$ (Figure 1C). For the control group, the shape and scale parameters are constant, $(3, 18)$.

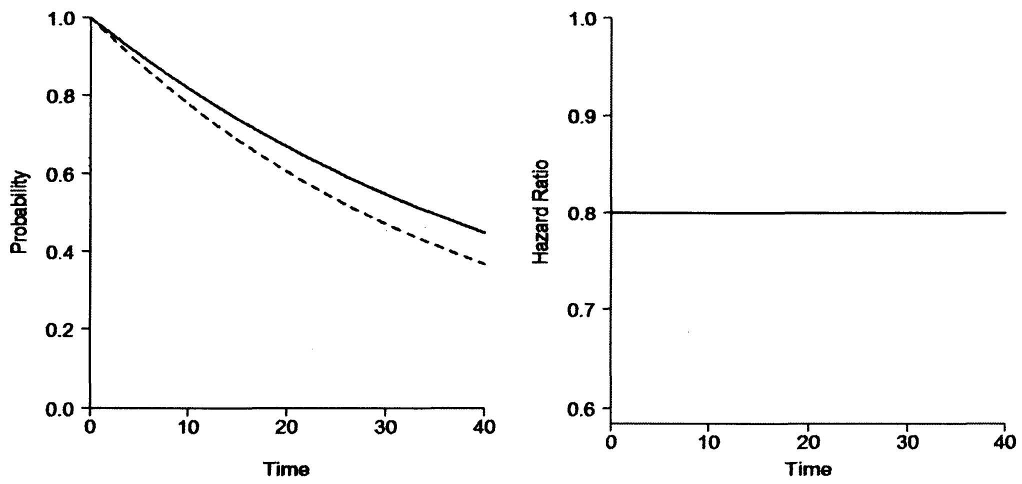

Figure 1A. Survival functions of event time from the Weibull distribution with shape and scale parameters $(1, 50)$ for the treatment group (solid line) and $(1, 40)$ for the control group (dashed line) (left), and the hazard ratio overtime (treatment group over control group) (right) in PH situation.

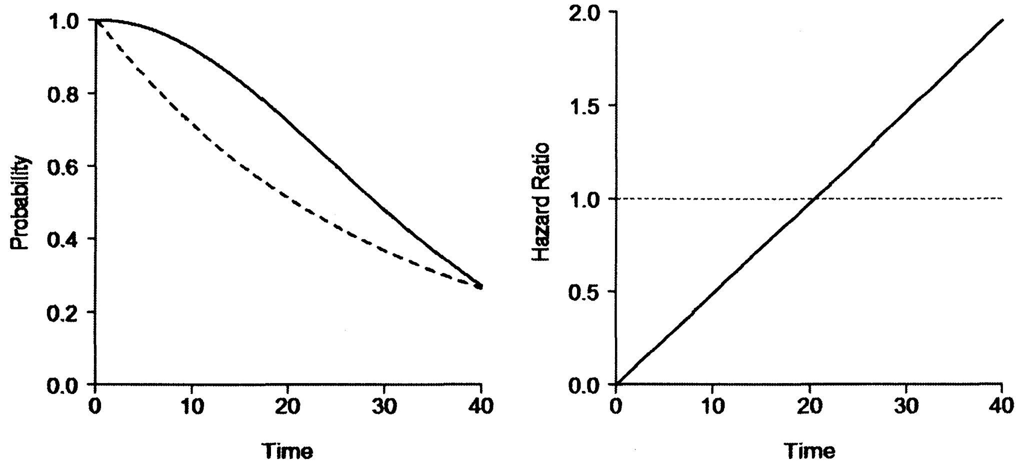

Figure 1B. Survival functions of event time from the Weibull distribution with shape and scale parameters $(2, 35)$ for the treatment group (solid line) and $(1, 30)$ for the control group (dashed line) (left), and the hazard ratio overtime (treatment group over control group) (right) in Non-PH situation.

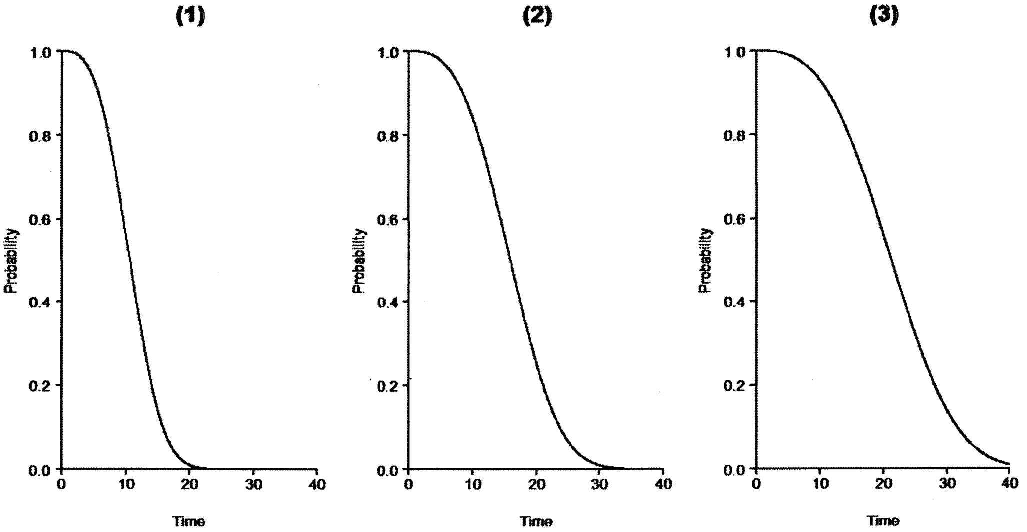

Figure 1C. Three patterns of the survival function of censoring time from the Weibull distribution for the treatment group with the shape and scale parameters: (1) (3, 12); (2) (3, 18); and (3) (3, 24).

One million survival time T and censoring time C were sampled and then we obtained observable survival data (X, δ), where $X = \min(T, C)$ and δ is an event indicator. Using these data, the hazard ratios between treatment and control groups were estimated by the Cox model and summarized in Table 1. In the PH situation, the estimated hazard ratios were always the same regardless of the censoring patterns. However, in the Non-PH situation, estimates were different depending on the censoring patterns. Since we generated one million observable data for each group, we consider these estimates as the true hazard ratios. The true hazard ratio depends on underlying study-specific censoring distributions when the PH assumption is not satisfied, so the parameter being estimated is no longer meaningful. This is one of the critical issues of the hazard ratio estimate.

	PH	Non-PH
Censoring		
(1)	0.80	0.30
(2)	0.80	0.44
(3)	0.80	0.49

Table 1. Estimates of the hazard ratio (treatment group over control group) based on three patterns of censoring distributions.

2.4) Example- Cancer trial V-15-32

Issues of the hazard ratio estimate are illustrated by the cancer trial, V-15-32[5] (Iressa study). The Iressa study was a phase III study comparing gefitinib with docetaxel in patients with advanced/metastatic non-small cell lung cancer (NSCLC) who had failed one or two chemotherapy regimens. The primary objective was to compare overall survival to demonstrate non-inferiority for gefitinib relative to docetaxel. The predefined non-inferiority margin for the hazard ratio (gefitinib over docetaxel) was 1.25, meaning the upper confidence band of the estimated hazard ratio must be lower than 1.25 to achieve non-inferiority. This study enrolled 489 patients (245 were randomly assigned to gefitinib and 244 were randomly assigned to docetaxel).

In this illustration, we reconstructed data from the original paper since the original data were not publically available. We reconstructed the individual patient data from the Kaplan-Meier (KM) curves, the number of patients at risk and the total number of events from published article by using the algorithm proposed by Guyot et al[6]. The KM curves of overall survival (Figure 2 (left)) and the hazard ratio, 1.11 (95%CI, 0.89 to 1.39), for the reconstructed data are identical to those published in Maruyama et al[5].

Figure 2 (right) shows the estimate of the log hazard ratio over time with 95% confidence band for the reconstructed data[7]. As seen in Figure 2 (right), the log hazard ratio was not constant over time. In addition to the visual validation, the standard lack-of-fit test based on Schoenfeld residuals[8] also validated the violation of the PH assumption (p=0.002).

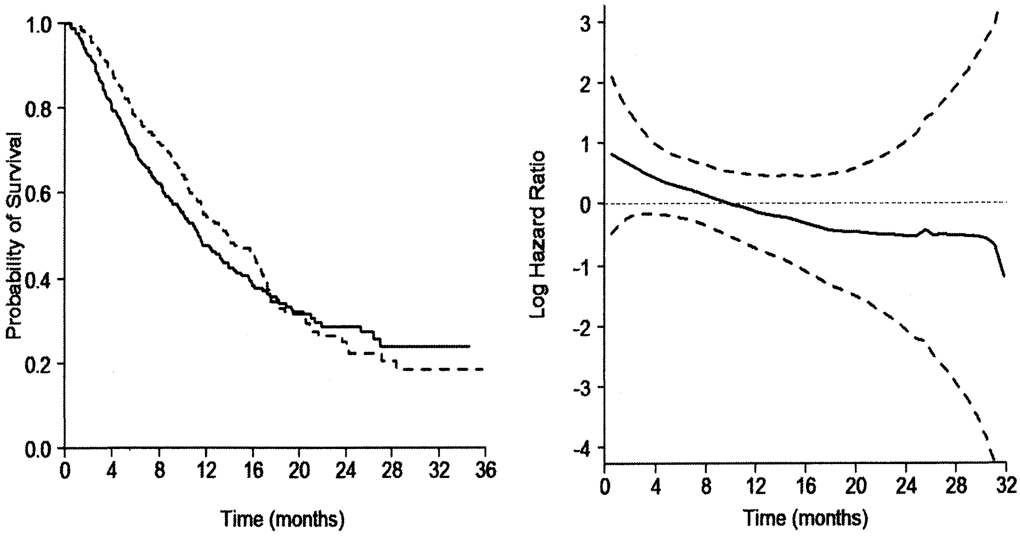

Figure 2. KM curves for gefitinib (solid line) and docetaxel (dashed line) (left) and the log hazard ratio (solid line) with the 95 per cent confidence band (dashed line) (right) for the reconstructed Iressa data.

In the Iressa study, the hazard ratio was 1.12 (95%CI, 0.89 to 1.40) thus non-inferiority for gefitinib relative to docetaxel was not achieved. It is difficult to clinically interpret the hazard ratio of 1.12. In addition, the wide confidence interval of the hazard ratio and violation of the PH assumption do not allow for a clear decision on the outcome of the trial.

The same as the Iressa study, the PH assumption is unlikely satisfied in practice. As mentioned in the paper by Uno et al[1], critical issues such as those in the Iressa study have yet to be acknowledged by the broader community of health science researchers. There is no single metric or parameter that quantifies the between-group difference without any problems. However, population summary measures are crucial for study design and planning. Uno et al[1] suggested using RMST-based measures as alternatives to the hazard ratio. The next section will introduce three alternative model-free measures to the hazard ratio.

3. Alternative model-free measures to the hazard ratio

3.1) Difference (or ratio) of t-year survival probabilities

The difference (or ratio) of the *t*-year survival probabilities is used to compare the survival probabilities between two groups at the time of interest. It is difficult to make a decision by this measure since the optimal time point is difficult to choose especially when two survival curves cross.

3.2) Difference (or ratio) of median survival times

The difference (or ratio) of the median survival times is used to compare the times between two groups where the survival probabilities are 50%. Estimation of this measure is difficult when survival curves do not reach 50% due to the small number of events.

3.3) Difference (or ratio) of restricted mean survival times

If we are interested in the survival probabilities at a fixed time point or in the times for a specific percentile, the corresponding population measure introduced in section 3.1 or 3.2 can be an appropriate measure. Conversely, if we are interested in a global profile of the between-group difference, RMST-based measures are useful.

The restricted mean survival time, $\mu(t^*)$, is defined as the area under the survival function, $S(t)$, up to the restricted time t^* $(< \infty)^{2\text{-}3}$:

$$\mu(t^*) = \int_0^{t^*} S(s)\mathrm{d}s.$$

$\mu(t^*)$ can be estimated regardless of the number of events. When we consider death as the event, we can interpret the RMST as the t^*-year life expectancy. The simple interpretation of the RMST would be "on average, the life expectancy with this treatment over the next t^* is $\mu(t^*)$."

We denote RMST of the treatment group by $\mu_\mathrm{t}(t^*)$ and the control group by $\mu_\mathrm{c}(t^*)$. The difference, $\mu_\mathrm{t}(t^*) - \mu_\mathrm{c}(t^*)$, and the ratio, $\mu_\mathrm{t}(t^*)/\mu_\mathrm{c}(t^*)$, of the two estimated RMSTs can be measures for the between-group difference. These measures are robust and provide clinically interpretable results.

4. Secondary analysis of a non-inferiority study using the RMST

In this section, secondary analysis of the Iressa study using the RMST will be shown. Since the hazard ratio was used for the determination of the non-inferiority margin in the Iressa study, we evaluate the non-inferiority margins by two approaches using RMST-based measures.

4.1) First method to determine RMST-based non-inferiority margin

The first approach is to use the information of the hazard ratio with the following steps:

i. Assume we already have data for the control group (docetaxel) and fit the Weibull regression model to estimate scale and shape parameters for the control group, $\hat{\eta}_\mathrm{c}$ and $\hat{m}$.

ii. The survival function and the RMST for the control group are obtained by the standard integration, $\hat{S}_\mathrm{c}(t) = \exp\left\{-\left(\frac{t}{\hat{\eta}_\mathrm{c}}\right)^{\hat{m}}\right\}$, and $\hat{\mu}_\mathrm{c}(t^*) = \int_0^{t^*} \hat{S}_\mathrm{c}(s)\mathrm{d}s$.

iii. Since the Weibull shape parameters for two groups are the same under the PH assumption, $\hat{m}$ is used also for the treatment group (gefitinib). We denote the scale parameter for the treatment group by $\hat{\eta}_\mathrm{t}$. The non-inferiority margin for the hazard ratio, $\frac{h_\mathrm{t}(t)}{h_\mathrm{c}(t)} = \frac{m \cdot t^{m-1}}{\eta_\mathrm{t}{}^m} / \frac{m \cdot t^{m-1}}{\eta_\mathrm{c}{}^m} = \lambda$, is satisfied, where $h_\mathrm{t}(t)$ and $h_\mathrm{c}(t)$ are the hazard functions for the treatment group and the control group. From this relationship, $\hat{\eta}_\mathrm{t}$ can be derived by $\hat{\eta}_\mathrm{t} = \hat{\eta}_\mathrm{c}/\exp\left(\frac{\log \lambda}{\hat{m}}\right)$.

iv. Under the above calculations, the survival function and the RMST for the treatment group can be estimated by $\hat{S}_\mathrm{t}(t) = \exp\left\{-\left(\frac{t}{\hat{\eta}_\mathrm{t}}\right)^{\hat{m}}\right\}$, and $\hat{\mu}_\mathrm{t}(t^*) = \int_0^{t^*} \hat{S}_\mathrm{t}(s)\mathrm{d}s$.

v. $\Delta_\mathrm{d} = \hat{\mu}_\mathrm{t}(t^*) - \hat{\mu}_\mathrm{c}(t^*)$ and $\Delta_\mathrm{p} = \hat{\mu}_\mathrm{t}(t^*)/\hat{\mu}_\mathrm{c}(t^*)$ can be regarded as the non-inferiority margin for the difference and the ratio of the RMSTs.

4.2) Second method to determine RMST-based non-inferiority margin

When determining the non-inferiority margin for the hazard ratio, a conventional value like 1.25 or 1.33 is likely to be used. However, as previously described, the hazard ratio is difficult to clinically interpret and the PH assumption is unlikely satisfied in practice. Next, we evaluate the non-inferiority margin independently of the hazard ratio. In this approach, we consider the non-inferiority margin as $\alpha \cdot t^*$ years, where α is, for example, 0.02, meaning that the t^*-year life expectancy of the patient in the treatment group is $\alpha \cdot t^*$ years shorter than that of the patient in the control group. α can be specified from previous clinical trials.

4.3) Results

4.3.1) Summary of estimates

If we specified the restricted time as 24 months (t^*= 24 months), the estimated RMST for gefitinib was 13.0 months and for docetaxel was 14.1 months. From these results, we can say “on average, the patient is expected to live 13 months treated with gefitinib, and 14 months with docetaxel over the next 24 months.” The estimated difference of the two RMSTs (gefitinib minus docetaxel) was -1.1 months (95%CI, -2.6 to 0.4), and the ratio (gefitinib over docetaxel) was 0.92 (95%CI, 0.83 to 1.03). According to these measures, we can estimate that the between-group difference is about a month in the next two years. The confidence interval of the hazard ratio, 0.89 to 1.40, is wide, while that of the RMST-based measures are more compact. Table 2 shows the summary of estimates and Figure 3 shows estimates of the RMST up to 12 months and 24 months (t^*= 12 months, 24 months).

	Estimate (95% CI)	
	12 months	24 months
gefitinib (months)	8.66 (8.16 to 9.16)	13.0 (11.9 to 14.1)
docetaxel (months)	9.53 (9.09 to 9.97)	14.1 (13.0 to 15.1)
Difference (months) (gefitinib minus docetaxel)	-0.9 (-1.5 to -0.2)	-1.1 (-2.6 to 0.4)
Ratio (gefitinib over docetaxel)	0.91 (0.84 to 0.98)	0.92 (0.83 to 1.03)

Table 2. Summary of estimates up to 12 months and 24 months for the reconstructed Iressa data.

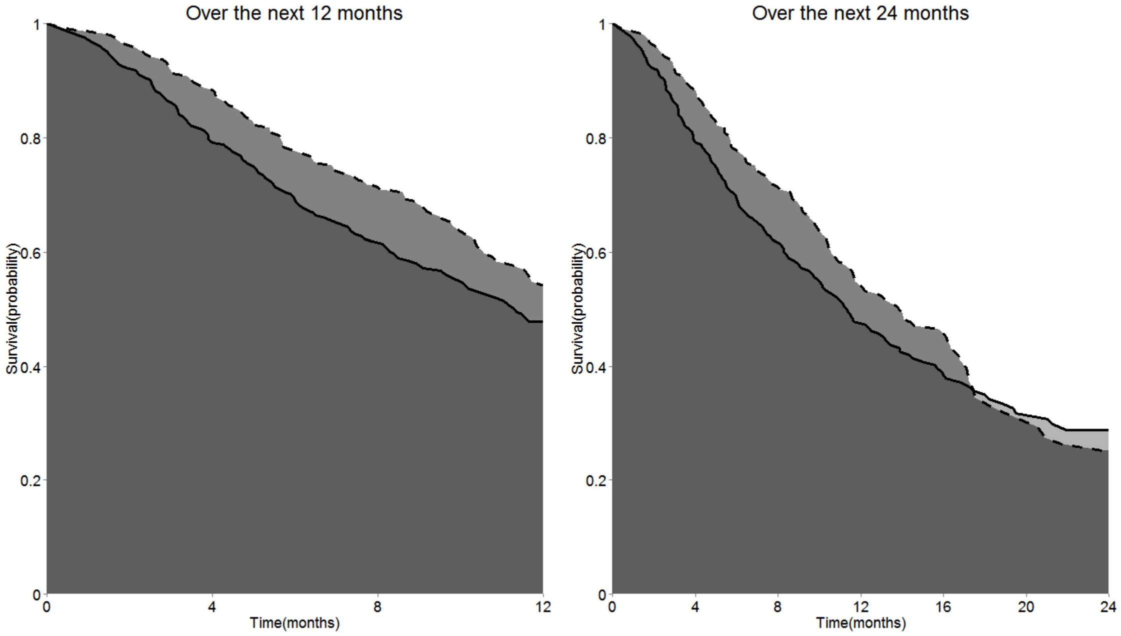

Figure 3. Estimates of RMST up to 12 months (8.7 months for gefitinib, area under the solid line; 9.5 months for docetaxel, area under the dashed line) and up to 24 months (13 months for gefitinib, area under the solid line; 14 months for docetaxel, area under the dashed line).

4.3.2) Result of non-inferiority by the first method

In the first approach to determine the RMST-based non-inferiority margin, Δ_d and Δ_p were -1.4 and 0.90, respectively. Thus non-inferiority for gefitinib relative to docetaxel was not achieved (t*= 24 months).

4.3.3) Result of non-inferiority by the second method

In the second approach, we considered the non-inferiority margin as 0.48 months (α = 0.02, t*= 24 months). Δ_d and Δ_p were -0.48 months and 0.97. Non-inferiority was not achieved.

5. Discussion

The hazard ratio estimate by the Cox model is commonly used in survival analysis. Under appropriate conditions, the hazard ratio is a valid measure for the between-group difference. However, there is one of the critical issues of the hazard ratio estimate such as independence of underlying study-specific censoring distributions. In addition, the clinical interpretation of the hazard ratio is difficult and the confidence interval of the estimated hazard ratio is wide when there are few events. The restricted mean survival time (RMST), the t^*-year life expectancy when death is the event, is easy to understand and RMST-based measures are practically useful alternatives for the between-group difference. Considering other ideal statistical properties shown in this paper, we recommend using RMST-based measures to quantify the between-group difference.

Since t^* is required to estimate the RMST, it is important to choose t^* carefully in advance and to evaluate the sensitivity on the primary result. One may choose t^* as the last follow up time. In case the number of patients at risk is small around the last follow up time, it would be appropriate to set t^* at the time with enough number of patients at risk to estimate the RMST. When we plan RMST-based analysis in a clinical trial, t^* should be prespecified based on the results of previous clinical trials.

In the Iressa study example, we do not set t^* at the end of the study but at 12 months and 24 months. For discussion and sensitivity analysis, we conducted the analyses at every restricted time point ($t^* = 12,13,14, \ldots ,34$ months) by two approaches to determine the RMST-based non-inferiority margins described in section 4.1 and 4.2. Figure 4 shows the results. In order to achieve non-inferiority, the non-inferiority margin should be lower than the lower confidence interval of the difference (or ratio) of the RMSTs. The results showed that non-inferiority for gefitinib relative to docetaxel was not demonstrated at all restricted time points. We recommend conducting this type of sensitivity analysis for a clinical trial with RMST-based measures.

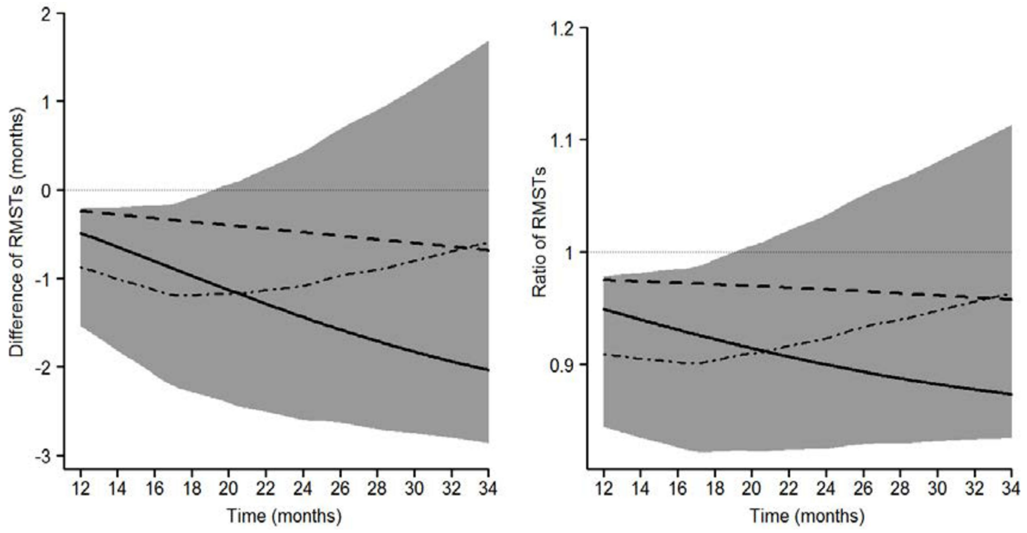

Figure 4. Estimates of RMST with the 95 per cent confidence intervals (dot dashed lines with gray areas), and the non-inferiority margins by two approaches (the first approach described in section 4.1, solid lines; the second approach in section 4.2, dashed lines) for the difference (left) and the ratio (right) of the RMSTs.

6. Conclusion

The hazard ratio estimator depends on underlying study-specific censoring distributions when the proportional hazards assumption is violated, as we demonstrated via the numerical study in this paper. In addition, due to the lack of a baseline reference number, the interpretation of the hazard ratio estimate is quite difficult. Using clinically interpretable model-free measures, such as the difference of the RMSTs, is recommended instead of the hazard ratio unless there is biological justification or strong belief of the proportional hazards assumption.

Acknowledgements

We are deeply grateful to all people who helped us throughout this work, especially to Dr. Hajime Uno for his warm support. Our appreciation also goes to Madoka Takeuchi for correction of English sentences on the manuscript, Dr. Mindy Fang for technical advice for programming, and Dr. Beth Ann Griffin for insightful comments.

References

[1] Uno H, Claggett B, Tian L, et al. Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. J Clin Oncol 2014; 32: 2380-2385.

[2] Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Stat Med 2011; 30: 2409-2421.

[3] Royston P, Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol 2013; 13: 152.

[4] Cox DR. Regression models and life tables. J R Stat Soc B 1972; 34: 187-220.

[5] Maruyama R, Nishiwaki Y, Tamura T, et al. Phase III Study, V-15-32, of Gefitinib Versus Docetaxel in Previously Treated Japanese Patients With Non-Small-Cell Lung Cancer. J Clin Oncol 2008; 26: 4244-4252.

[6] Guyot P, Ades AE, Ouwens MJ, et al. Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol 2012; 12: 9.

[7] Gilbert PB, Wei LJ, Kosorok MR, et al. Simultaneous inferences on the contrast of two hazard functions with censored observations. Biometrics 2002; 58: 773-780.

[8] Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. Biometrika 1980; 67: 145-153.

Department of Clinical Medicine (Biostatistics)
School of Pharmacy, Kitasato University
5-9-1 Shirokane, Minato-ku, Tokyo, 108-8641
JAPAN

北里大学薬学部臨床医学(臨床統計学)
堀口 みき　朴 慶純　三上 剛史　竹内 正弘