Annals of Internal Medicine RESEARCH AND REPORTING METHODS

# Alternatives to Hazard Ratios for Comparing the Efficacy or Safety of Therapies in Noninferiority Studies

**Hajime Uno, PhD*; Janet Wittes, PhD*; Haoda Fu, PhD*; Scott D. Solomon, MD; Brian Claggett, PhD; Lu Tian, ScD; Tianxi Cai, ScD; Marc A. Pfeffer, MD, PhD; Scott R. Evans, PhD; and Lee-Jen Wei, PhD**

A noninferiority study is often used to investigate whether a treatment's efficacy or safety profile is acceptable compared with an alternative therapy regarding the time to a clinical event. The empirical quantification of the treatment difference for such a study is routinely based on the hazard ratio (HR) estimate. The HR, which is not a relative risk, may be difficult to interpret clinically, especially when the underlying proportional hazards assumption is violated. The precision of the HR estimate depends primarily on the number of observed events but not directly on exposure times or sample size of the study population. If the event rate is low, the study may require an impractically large number of events to ensure that the prespecified noninferiority criterion for the HR is attainable. This article discusses deficiencies in the current approach for the design and analysis of a noninferiority study. Alternative procedures are provided, which do not depend on any model assumption, to compare 2 treatments. For a noninferiority safety study, the patients' exposure times are more clinically important than the observed number of events. If the patients' exposure times are long enough to evaluate safety reliably, then these alternative procedures can effectively provide clinically interpretable evidence on safety, even with relatively few observed events. These procedures are illustrated with data from 2 studies. One explores the cardiovascular safety of a pain medicine; the second examines the cardiovascular safety of a new treatment for diabetes. These alternative strategies to evaluate safety or efficacy of an intervention lead to more meaningful interpretations of the analysis results than the conventional strategy that uses the HR estimate.

*Ann Intern Med.* doi:10.7326/M14-1741 **www.annals.org**

For author affiliations, see end of text.

* Drs. Uno, Wittes, and Fu have contributed equally to this work.

This article was published online first at www.annals.org on 9 June 2015.

Several statistical and clinical publications highlight concerns about superiority studies in which the hazard ratio (HR) is used as a summary measure for assessing the efficacy of a new therapy (1–3), but few if any address the use of the measure in noninferiority studies. The HR is a model-based measure of differences between 2 groups, and it therefore assumes a specific relationship between the 2 distributions of the outcome variable. The interpretability of such a summary measure depends heavily on the validity of the model assumptions. Noninferiority studies have often been used for comparative evaluations of the efficacy or safety of therapies (4–6). This article uses 2 examples to illustrate the limitations of using the HR when designing and interpreting such studies and discusses the pros and cons of using alternative measures, such as the risk difference and the difference between 2 restricted mean survival times (RMSTs). (See **Appendix Table 1**, available at www.annals.org, for a glossary of terms.)

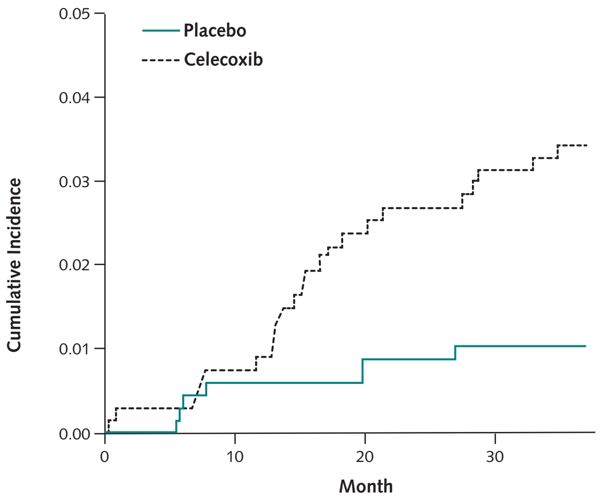
## Example 1: Celecoxib Study

The APC (Adenoma Prevention With Celecoxib) trial tested whether celecoxib, 400 mg twice daily, would reduce the recurrence of colorectal adenoma after polypectomy (7). The study randomly assigned 671 and 679 patients to celecoxib and placebo, respectively. The end point for cardiovascular (CV) safety was the time to a composite outcome of death from CV causes, myocardial infarction, stroke, and heart failure. At the advice of the data monitoring committee, the trial ended early with 23 and 7 events in the celecoxib and placebo groups, respectively. Although the observed event rates were low, the cumulative incidence curves, which indicate the event rates over time (**Figure 1**), seem markedly different.

A conventional way to quantify the between-group difference is to calculate the HR under the assumption of proportional hazards (PH). The PH assumption requires the ratio of the 2 hazard functions to be approximately constant over time (8). For this example, the estimated HR was 3.35 (95% CI, 1.44 to 7.81; $P$ = 0.005) (7). Clinically, even if the hazards were truly proportional, it is difficult to interpret a 3.4-fold increase in hazard for celecoxib compared with placebo because the hazard is not a probability measure and the HR is not a relative risk. Rather, the HR is a ratio of hazard rates. Like other ratio-based measures, the estimated HR may convey a dramatic contrast between 2 groups when the observed event rates are low. For the celecoxib trial, the estimated event rates at 36 months for the celecoxib and placebo groups were 3.0% and 1.0%, respectively. Thus, the tripling of the HR corresponded to only a 2.0% absolute increase (CI, 0.8% to 3.2%) in rates (**Table 1**).

The precision of the estimated HR depends mainly on the number of observed events and not on the number of patients or their exposure times. If we artificially added 1000 exposure times censored at the end of the study—that is, no additional events—to each group of the celecoxib trial, the estimated HR would change little (HR, 3.29 [CI, 1.41 to 7.67]). On the other hand, with those additional observations, the rate difference at 36 months would be 0.9% and the 95% CI of 0.3% to 1.6% would be a much more precise estimate. Further, when the PH assumption is violated (the HR is not actually constant over time), the clinical meaning of the HR is unclear. The 2 empirical cumulative incidence curves

***Figure 1.*** Empirical cumulative incidence curves for patients in the celecoxib study.



| Risk, *n* | | | | |
|---|---|---|---|---|
| Placebo | 679 | 675 | 669 | 667 |
| Celecoxib | 671 | 666 | 653 | 648 |

Patients randomly assigned to celecoxib, 400 mg twice daily (*dashed line*), and placebo (*solid line*).

for the celecoxib trial (**Figure 1**) separate after 10 months but not during the initial study period. This indicates a possible violation of the PH assumption. Checking the plausibility of the PH assumption when there are few events, as in the celecoxib study, is problematic because no goodness-of-fit test would have sufficient power to detect the inadequacy of the PH model.

## Example 2: Saxagliptin Study

A randomized, placebo-controlled clinical trial was conducted to assess the potential CV risk of saxagliptin, which is a dipeptidyl peptidase-4 inhibitor for patients with type 2 diabetes mellitus (9). The primary end point was the time to the first occurrence of CV death, nonfatal myocardial infarction, or nonfatal ischemic stroke. To claim that saxagliptin is noninferior to placebo, the study investigators followed guidelines of the U.S. Food and Drug Administration (10) and prespecified a noninferiority margin for the HR (saxagliptin vs. placebo) of less than 1.30 under the PH model (10–12). If the upper bound of the observed 95% CI was less than 1.30, saxagliptin would be concluded to be safe. If noninferiority was established, the investigators planned to assess whether the CV safety of saxagliptin was superior to placebo.

Because the CI for the HR depends mainly on the observed number of events, investigators needed 1040 events by the end of the study to satisfy noninferiority and superiority objectives, regardless of the number of participants or duration of follow-up. To obtain 1040 events, the investigators randomly assigned 16 492 patients to saxagliptin or placebo in a 1:1 ratio; these patients were followed up to 2.9 years (median, 2.1 years). At the end of the study, 613 and 609 events had occurred in the saxagliptin and placebo groups, respectively. The estimated HR (saxagliptin vs. placebo) was 1.00 (CI, 0.89 to 1.12). Because the upper bound of this interval was less than 1.30, the trial satisfied the prespecified criterion for noninferiority; however, the drug failed to meet the claim that the CV safety of saxagliptin was superior to placebo.

Designing and analyzing a safety trial to establish noninferiority using the HR is not ideal. First, the threshold of 1.30 for the HR (10–12) does not account for any background absolute hazard value for the placebo group. If the event rate for the placebo group is very low, a potential 30% (or higher) additional hazard may not represent a clinically meaningful increase in risk. If the event rate is high, a 30% increase may be unacceptably high. Second, the width of the CI for the HR depends mainly on the observed number of events but not on the exposure times. Had the few events been distributed evenly over a reasonably long follow-up, the new therapy would have shown sufficient evidence of safety; however, with few events, the resulting CI for the HR would be unacceptably wide. Conventionally, but in some cases inappropriately, a wide CI suggests that evidence is insufficient to make conclusions about safety. Third, if the PH assumption (8) is violated (especially when the hazard functions cross during the study period), the standard inferential procedure based on the HR may not detect a potential excess risk because the study would have inadequate power to detect a difference between 2 groups.

## Alternatives to the HR

The event rates are low in most safety studies (9, 11, 13, 14). The conventional design for a noninferiority study, such as the saxagliptin study, requires a large number of study patients, long study duration, or both to show noninferiority. Using data from the saxagliptin study, we discuss several well-known model-free

***Table 1.*** Between-Group Difference Measure Estimates for the Celecoxib and Saxagliptin Studies*

| Measure | Study | |
|---|---|---|
| | **Celecoxib** | **Saxagliptin** |
| HR† | 3.35 (1.44 to 7.81) | 1.00 (0.89 to 1.12) |
| Risk difference, %‡ | | |
| At 36 mo | 2.0 (0.8 to 3.2) | – |
| At 900 d | – | −0.2 (−1.2 to 0.9) |
| RMST difference§ | 0.43 (0.08 to 0.78) | 0 (−5 to 4) |

HR = hazard ratio; RMST = restricted mean survival time.
* Estimates presented as point estimates (95% CIs) and contrasts relative to a placebo group. For the difference measures, a positive value indicates an increased risk of active treatment.
† Active treatment divided by placebo.
‡ Active treatment minus placebo.
§ Placebo minus active treatment; RMSTs calculated to 36 mo (celecoxib) and 900 d (saxagliptin). The units are month and day, respectively.

*Table 2.* Pros and Cons of Between-Group Difference Measures for Event-Time Analyses of Noninferiority Safety Studies

| Measures | Pros | Cons |
|---|---|---|
| HR (model-based) | A valid summary for the difference between 2 cumulative incidence distributions (when the PH assumption is correct), with statistically efficient inference procedures. | Lacks a clinically meaningful reference value for the hazard from the control group to assess the difference between groups.<br>Difficult to interpret when the PH model is far from correct because it estimates a population quantity that depends in part on the censoring distributions.<br>May not have adequate power to detect a safety signal especially when the 2 hazard functions cross during study follow-up.<br>May require an impractically large study because the precision of the estimated HR depends on the number of observed events and not directly on the number of patients and their exposure times.<br>May selectively study a higher-risk population than the indicated patient population for the new treatment because many observed events are needed. |
| Relative time (model-based) | Provides a clinically meaningful summary of the differences between the groups if the model is correctly specified. For example, if the estimated ratio (treatment group vs. control group) of 2 event times is 1.3, one can claim that, on average, a patient in the control group would gain an extra 30% "survival time" if treated with the new therapy. This, coupled with the survival distribution of the control group, provides a clinically meaningful interpretation of the treatment benefit. | Difficult to interpret when the model is not correct because the empirical relative time estimates a population quantity that depends on the censoring distributions. |
| Difference of percentiles (model-free) | Provides a clinically meaningful summary of the differences between groups and does not depend on a model assumption.<br>Has a well-developed inference procedure for the difference (ratio). | May not be estimable if follow-up is short or the event rate is low because not all the percentiles can be observed in such studies.<br>May be an unstable estimate because the median (the 50th percentile) is heavily dependent on the local shape of the cumulative incidence curve. |
| The *t*-year event-rate difference (model-free) | Provides an easy-to-interpret and clinically meaningful summary of the differences between groups.<br>Has a well-established and robust inference procedure.<br>Probably the most relevant quantity for decision making when one is interested in long-term survival. | Only reflects cumulative information at time *t*, and does not reflect any differences in the profile of the cumulative incidence curves up to *t*. |
| RMST difference (model-free) | Provides a clinically meaningful summary of the differences between groups.<br>Provides a more stable estimate than the median in survival time studies.<br>Uses more information than its *t*-year event-rate counterpart.<br>May not need an impractically large study to assess noninferiority if the patients' exposure time is sufficiently large for safety evaluation. | Needs prespecification of the time point of interest.<br>May selectively study a relatively healthy population with low event rates rather than the indicated patient population in order to obtain a noninferiority claim. |

HR = hazard ratio; PH = proportional hazards; RMST = restricted mean survival time.

alternatives to the HR and show that, had this study been much smaller, it would still have led to a statistically valid conclusion based on a clinically interpretable measure of safety. Table 2 summarizes advantages and disadvantages of the alternative measures. Note that a model-free alternative to the HR does not require the assumption of a specific relationship between 2 groups with respect to the outcome distribution.

## Risk Difference

An obvious choice for a model-free measure is the difference in event rates at a specific time point. For example, in the saxagliptin study, we might choose 900 days (approximately 2.5 years) after randomization, which is the last time point shown in the cumulative incidence curves (9). The estimated risk difference (saxagliptin minus placebo) is −0.2% with an 8.9% event rate for the placebo group, which indicates a small absolute reduction from saxagliptin. This between-group difference at a specific time provides a clinically interpretable comparison but may not capture the overall profile of the difference between the 2 cumulative incidence curves. The method for event-rate differences at a specific time has been extensively discussed (15).

## Percentile Difference

Another common measure, the difference of 2 median event times, can be easily obtained using the cumulative incidence curves (16); however, when the event rate is low or the follow-up is short, the median event time may not be observable. Instead, one may use a difference of percentiles between 2 groups (17). For this example, the difference of the 5th percentiles (saxagliptin vs. placebo) is 0. Although percentiles other than the median have a simple mathematical interpretation, their meanings may not be intuitively obvious to investigators or patients.

## RMST Difference

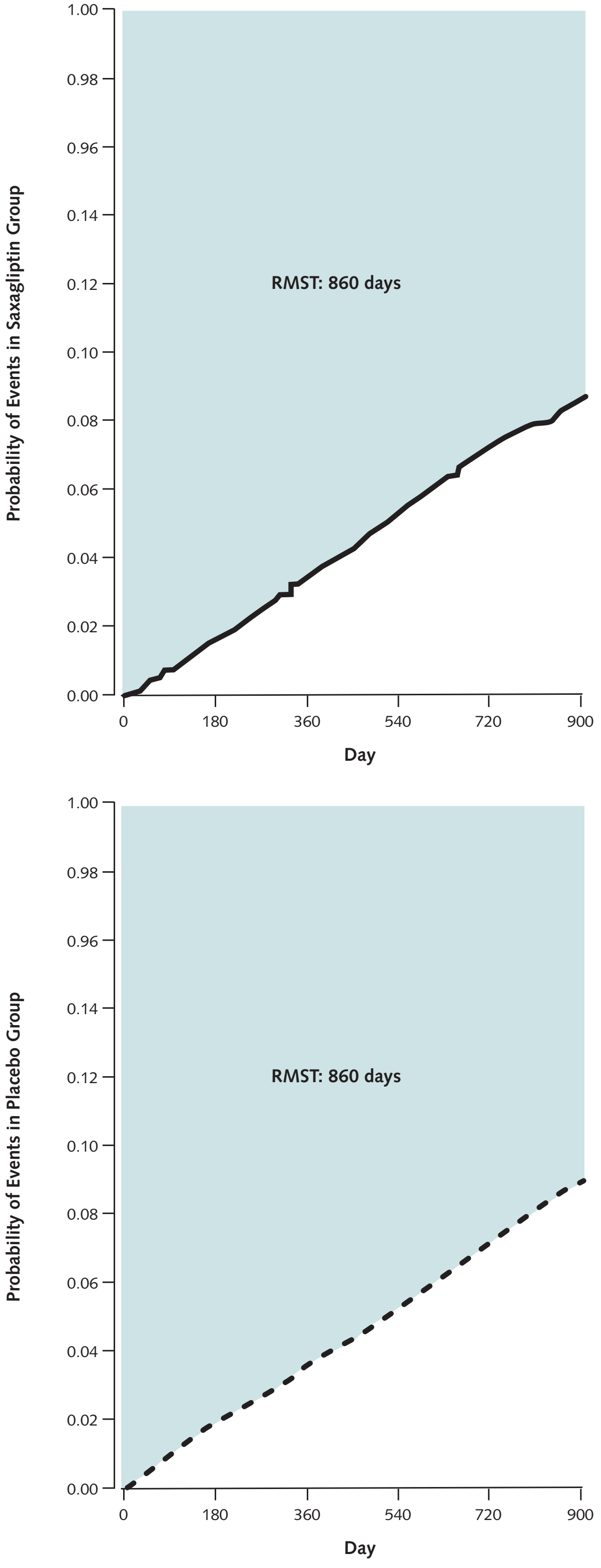
An attractive but seldom used alternative is the RMST (18–21) up to a specific time point. The RMST is the expected time spent event-free for a future patient followed for a specified time. It is estimated by the area above the empirical cumulative incidence curve. **Figure 2** presents the observed cumulative incidence curves up to 900 days as reported in the saxagliptin study (9). The areas above the curves for saxagliptin and placebo are both approximately 860 days. That is, if we treat future patients from the study population and follow them for 900 days, the average time spent event-free would be approximately 860 days for both groups, with an observed difference of 0 days between the 2 groups. These RMST estimates incorporate the number of events and the exposure times. With the observed RMST for the placebo group, the corresponding CI estimate for the difference in RMSTs can be interpreted statistically and clinically for assessing noninferiority. As with models for HRs that can adjust for baseline covariates, models for event-rate or RMST differences can also adjust for baseline imbalances (22–24).

To illustrate our proposals, we applied an algorithm developed by Guyot and colleagues (25) to the observed incidence rate curves in the publication of the saxagliptin study to reconstruct individual-patient–level data. We used these reconstructed data to make inferences about the risk and RMST differences at 900 days. Risk differences are presented as saxagliptin minus placebo and RMST differences as placebo minus saxagliptin so that positive values consistently indicate increased risk of saxagliptin treatment. The 95% CI for the risk difference is −1.2% to 0.9%. For the difference in RMSTs, the 95% CI is −5 to 4 days (**Table 1**). That is, at the confidence level of 95%, future patients treated with saxagliptin for 900 days would be expected to be free of CV events for as many as 5 days more to as many as 4 days less than their placebo counterparts on average. These bounds on the difference in RMSTs, coupled with the summary measure for the control group (RMST of 860 days), provide more clinically interpretable information about the group difference than a HR of 1.00 (CI, 0.89 to 1.12). **Appendix 1** (available at www.annals.org) provides computer programs for implementing RMST analyses with a documented example.

## How Group Difference Measures Affect the Study Size and Precision of Estimates

To explore the connection between the study size and the observed noninferiority bound for group difference measures with the reconstructed data from the saxagliptin study, we randomly selected a subset of patients using a fixed proportion of the original study size and constructed 95% CIs for the HR and the difference in RMSTs at day 900. We repeated the process 1000 times and obtained the average of the resulting 1000 95% CI estimates for each measure. **Table 3** has the results with several sample sizes: 15%, 20%, and 25% of the saxagliptin study. For example, had the saxagliptin trial enrolled only 2474 study participants (15% of those actually enrolled), the resulting average 95% CI for the HR would have been 0.76 to 1.36, with the upper

*Figure 2.* Empirical cumulative incidence curves with reconstructed event-time data for the saxagliptin study.



The shaded area (the area above the cumulative incidence curve) in each panel is the RMST up to 900 days. RMST = restricted mean survival time. **Top.** Saxagliptin group (*solid line*). **Bottom.** Placebo group (*dashed line*).

***Table 3.*** Study Size and Corresponding Noninferiority Bounds for HRs, RMST Difference, and Risk Difference in the Saxagliptin Study*

| Variable | Entire Study Population (*n* = 16 492) | Subsamples of the Total Study Population† | | |
|---|---|---|---|---|
| | | **25% (*n* = 4123)** | **20% (*n* = 3298)** | **15% (*n* = 2474)** |
| HR | 0.89 to 1.12 | 0.80 to 1.26 | 0.78 to 1.29 | 0.76 to 1.36 |
| RMST difference, *d*‡ | −5 to 4 | −9 to 9 | −11 to 10 | −12 to 12 |
| Risk difference at 900 d, % | −1.2 to 0.9 | −2.3 to 2.0 | −2.6 to 2.2 | −2.9 to 2.6 |

HR = hazard ratio; RMST = restricted mean survival time.

* The numbers are presented as 95% CI and contrasts relative to the placebo group. For the difference measures, a positive value indicates an increased risk of saxagliptin treatment. Data are reconstructed from the original report. See **Appendix 3** (available at www.annals.org) for details.

† RMSTs are calculated to 900 d.

‡ Estimates are based on 1000 repeated random samples of size 25%, 20%, and 15% of the total study population. See text ("RMST Difference" section) for details.

bound exceeding 1.3. On the other hand, for the difference in RMSTs, the average 95% CI would have been −12 to 12 days. This estimate provides a high degree of confidence that, on average, the saxagliptin group would be free of events no more than 12 days less than the placebo group through 900 days of follow-up. If a difference of 12 days out of 900 is a clinically acceptable margin of noninferiority, the saxagliptin trial could have been done with substantially fewer patients.

Note that event-free observations at the end of the study may contribute information to the difference in RMST but not to the HR. For each generated sample of 2474 patients in this simulation, if we were to add 7009 artificial observations across 900 days to each group to match the saxagliptin study's original sample size of 16 492, the resulting average 95% CI for the difference in RMSTs would be −2 to 2 days. On the other hand, the corresponding 95% CI for the HR, which depends primarily on the observed number of events, would be 0.75 to 1.35—practically identical to the previously mentioned 95% CI of 0.76 to 1.36. That is, censored observations (patients without events at 900 days) contribute essentially nothing to the precision of the HR.

## Using Alternative Measures to Design a Noninferiority Study

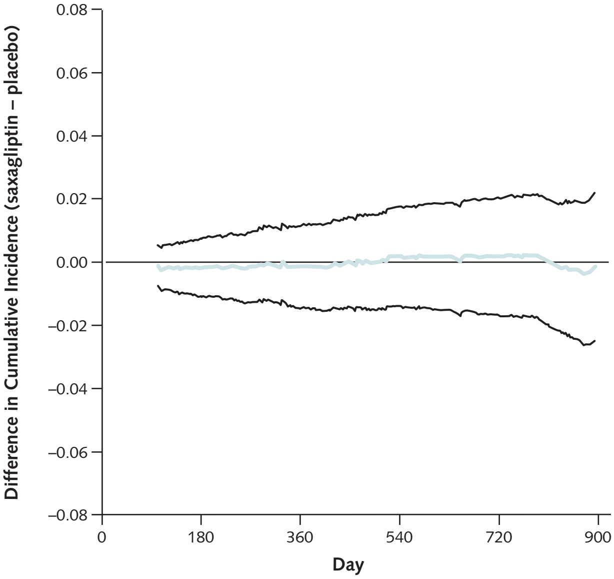
To design a noninferiority study, we usually assume that 2 groups being compared are identical to each other with respect to the distribution of the end point (the time to a specific event). We also specify such elements as the metric or parameter that will be used to compare groups, the noninferiority margin, the statistical inference procedure (for example, the 2-sided 95% CI estimate for the between-group difference) for assessing the noninferiority of the new therapy, the parametric distribution for the outcome variable, the patient's potential exposure time for safety assessment, the number of patients expected to enroll, and the accrual profile over time. For a conventional design that uses the HR as the group contrast measure, the noninferiority margin for a diabetes drug is generally 1.30 in safety studies (10–12). The rationale for this specific HR has never been clear. In designing such a trial, survival times are often assumed to be exponential (that is, constant hazard rates throughout the study) and the timing of the end of the study is determined by the total number of events such that the upper bound of the 95% CI for the HR is probably (for example, a chance of 80%) less than the noninferiority margin.

If we design a study using the difference of the RMSTs, we need to specify a time point when the RMST will be evaluated, which should be long enough to assess the treatment's clinical safety profile. Under the conventional setting with the HR, the saxagliptin study would need 456 CV events to ensure enough evidence for assessing the drug's safety regardless of the underlying event rates. To show how to design a study similar to the saxagliptin study using the difference between 2 RMSTs, we assume that the time point for evaluating the RMST is 900 days with a noninferiority margin of 18 days, which is 2% of 900 days. **Appendix 2** (available at www.annals.org) presents a simple numerical procedure to calculate the study sample size under various accrual profiles over time. For example, if 30 patients per day enter the study with at least 10% having 900 days of follow-up at the end, the study needs about 2100 participants and a total of 2.5 years to finish so that the noninferiority margin of 18 days is attainable with high probability. Note that the corresponding upper bound of the 95% CI for the HR would be 1.52.

## The Group Difference Measure and the Choice of the Study Population

When designing a safety trial, the study participants should be chosen appropriately from a target population that clinicians would treat in a real-life setting. Otherwise, the study investigators might “game” the system by selecting patients improperly to reach the study goal faster. For example, using the RMST as the primary parameter of interest, one may choose patients with low CV risk in a CV outcome trial. On the other hand, using the HR approach, the investigators might choose patients with high CV risk to collect a large number of events in a short period. Using a relatively short-term study for assessing safety can be problematic because it might be too short to identify unexpected rare events from patients treated with a new therapy. For example, in a recent large and long-term clinical trial for evaluating the safety of darbepoetin alfa, a small excess number of strokes was unexpect-

***Figure 3.*** Saxagliptin versus placebo.



Point estimate (*green line*) and 95% simultaneous confidence band (*black lines*) for the difference of the cumulative incidence.

edly detected in the group assigned to darbepoetin alfa (26). However, long-term controlled, comparative clinical studies in settings with limited resources to address safety concerns can be impractical.

## Evaluating Group Differences Over a Set of Time Points Simultaneously

Some situations require comparing 2 treatment groups across a set of time points simultaneously, rather than at a specific time point. The cumulative incidence curves (**Figure 2**) provide temporal profiles of the event rates. Although the 2 empirical curves for the saxagliptin study visually overlap, whether we can claim that their population counterparts are "equivalent" statistically and clinically over time might be of interest. To this end, one may construct simultaneous CIs for the curve of the difference between 2 cumulative incidence functions over time. **Figure 3** shows a 95% simultaneous confidence band between 100 and 900 days for the saxagliptin study (27). This band suggests that, with high probability, the true difference of 2 cumulative incidence curves would be contained entirely within the 2 dotted lines between 100 and 900 days. For example, at 300, 600, and 900 days, the true differences of the cumulative incidence curves probably fall in the intervals of −1.1% to 1.0%, −1.4% to 1.6%, and −2.3% to 1.9%, respectively. This information, coupled with the empirical cumulative incidence curve for the placebo group, provides additional useful information for decision making about the CV safety of saxagliptin beyond that obtained from a single summary measure evaluated at 900 days.

## Implications for Clinical Interpretation

**Table 2**, which presents advantages and disadvantages of various measures, shows that measures other than the HR facilitate a clinically meaningful interpretation of findings. We further illustrate this point with data from the celecoxib trial (7) in which the estimated event rates at 36 months for the celecoxib and placebo groups were 3.0% and 1.0%, respectively. The 95% CI for the difference of the event rates was 0.8% to 3.2%. The estimated RMSTs through 36 months of follow-up for the celecoxib and placebo groups were 35.33 and 35.76 months, respectively, a difference of 0.43 month (CI, 0.08 to 0.78 months; $P = 0.015$) or about 13 days (CI, 2 to 24 days) (**Table 1**). So, with 95% confidence, the patients treated with celecoxib would be event-free about, at most, 24 days shorter than their placebo counterparts. These estimates are based on the cumulative incidence rate or RMST and are easier to interpret clinically than the corresponding HR estimate of 3.4.

## Conclusion

Design and analysis of superiority and noninferiority studies differ fundamentally. Although the patients' exposure times are important for both, the number of observed events is essential for evaluating a superiority claim for a new therapy over the control but not for assessing safety by noninferiority. That is, "no news is bad news" for efficacy but "no news could be good news" for safety.

Note that it is not clear how to compare the statistical efficiency of a robust estimation procedure discussed in this article with a model-based counterpart because the underlying HR and the RMST difference parameters are not directly comparable. In some cases, the advantage of using the event rate or RMST difference to quantify the group difference is obvious. For example, if no event occurs in one treatment group of the study, the 95% CI based on the HR is infinitely wide but its counterpart for the absolute difference can be very narrow and provides sufficient evidence for assessing a noninferiority claim.

In summary, to explore toxicity, conventional study designs based on the HR have both statistical and clinical limitations. We encourage investigators at the design stage of the study to consider using the difference between 2 RMSTs or some other robust and clinically interpretable model-free metrics rather than the HR. No matter which measure is used, the RMST or the event rate for the control group is needed to provide context for clinical decision making. Using RMST or the cumulative incidence rate at a particular time point as a summary of the distribution of the event-time observations, the investigator must prespecify an expected follow-up that is sufficient for evaluating toxicity.

From Dana Farber Cancer Institute, Brigham and Women's Hospital, and Harvard University, Boston, Massachusetts; Statistics Collaborative, Washington, DC; Eli Lilly and Company, Indianapolis, Indiana; and Stanford University School of Medicine, Palo Alto, California.

**Acknowledgment:** The authors thank Dr. Robert O'Neill, 2 referees, and the editors.

**Grant Support:** Partially supported by the National Institute of Health (grant UM1AI104681 [Dr. Evans]) and contracts.

**Disclosures:** Dr. Wittes reports consulting fees paid to her employer from Astra-Zeneca, Bristol-Myers Squibb, Merck, Amgen, Reata, Relypsa, Acerta, and Janssen outside the submitted work. Dr. Pfeffer reports both grants and personal fees from Amgen; grants alone from Celladon, Hamilton Health Sciences, Novartis, and Sanofi Aventis; and personal fees alone from Aastrom, Abbot Vascular, Cleveland Clinic, FibroGen, GlaxoSmithKline, Medtronic, Merck, Roche, Servier, Teva, Daiichi Sankyo, Genzyme, Novo Nordisk, Salix, and Sanderling outside the submitted work. In addition, the Brigham and Women's Hospital has patents for the use of inhibitors of the ras genes in selected survivors of myocardial infarction with Novartis; Dr. Pfeffer is a co-inventor, and his share of the licensing agreement is irrevocably transferred to charity. Dr. Evans reports grants from the National Institute of Allergy and Infectious Diseases of the National Institutes of Health during the conduct of the study and personal fees from Millennium: The Takeda Oncology Company, Pfizer, Roche, Novartis, Achaogen, personal fees from MGH, Taris, Boston University, Huntington's Study Group, Auspex, Alcon, Merck, Chelsea, MannKind, QRx Pharma, IMMPACT (Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials), Genentech, Affymax, FzioMed, Amgen, GlaxoSmithKline, Sunovion, Boehringer-Ingelheim, American Statistical Association, U.S. Food and Drug Administration, Osaka University, City of Hope, National Cerebral and Cardiovascular Center of Japan, National Institutes of Health, Muscle Study Group, Society for Clinical Trials, Dug Information Association, University of Rhode Island, Rutgers New Jersey Medical School, PPRECISE (Preclinical Pain Research Consortium for Investigating Safety and Efficacy), and Statistical Communications in Infectious Diseases outside the submitted work. Authors not named here have disclosed no conflicts of interest. Disclosures can also be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M14-1741.

**Requests for Single Reprints:** Lee-Jen Wei, PhD, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115; e-mail, wei@hsph.harvard.edu.

Current author addresses and author contributions are available at www.annals.org.

## References

1. **Hernán MA.** The hazards of hazard ratios. Epidemiology. 2010;21: 13-5. [PMID: 20010207] doi:10.1097/EDE.0b013e3181c1ea43
2. **Muñoz A, Mongilardi N, Checkley W.** Multilevel competing risks in the evaluation of nosocomial infections: time to move on from proportional hazards and even from hazards altogether. Crit Care. 2014; 18:146. [PMID: 25042281] doi:10.1186/cc13892
3. **Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, et al.** Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol. 2014;32:2380-5. [PMID: 24982461] doi:10.1200/JCO.2014.55.2208
4. **D'Agostino RB Sr, Massaro JM, Sullivan LM.** Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. Stat Med. 2003;22:169-86. [PMID: 12520555]
5. **Head SJ, Kaul S, Bogers AJ, Kappetein AP.** Non-inferiority study design: lessons to be learned from cardiovascular trials. Eur Heart J. 2012;33:1318-24. [PMID: 22564354] doi:10.1093/eurheartj/ehs099
6. **Schumi J, Wittes JT.** Through the looking glass: understanding non-inferiority. Trials 2011;12:106-18.
7. **Solomon SD, Pfeffer MA, McMurray JJ, Fowler R, Finn P, Levin B, et al; APC and PreSAP Trial Investigators.** Effect of celecoxib on cardiovascular events and blood pressure in two trials for the prevention of colorectal adenomas. Circulation. 2006;114:1028-35. [PMID: 16943394]
8. **Cox DR.** Regression models and life-tables. J R Stat Soc Series B Stat Methodol. 1972;34:187-220.
9. **Scirica BM, Bhatt DL, Braunwald E, Steg PG, Davidson J, Hirshberg B, et al; SAVOR-TIMI 53 Steering Committee and Investigators.** Saxagliptin and cardiovascular outcomes in patients with type 2 diabetes mellitus. N Engl J Med. 2013;369:1317-26. [PMID: 23992601] doi:10.1056/NEJMoa1307684
10. **U.S. Department of Health and Human Services; Food and Drug Administration; Center for Drug Evaluation and Research.** Guidance for industry diabetes mellitus—evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. Accessed at www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm071627.pdf on 14 February 2014.
11. **Hirshberg B, Raz I.** Impact of the U.S. Food and Drug Administration cardiovascular assessment requirements on the development of novel antidiabetes drugs. Diabetes Care. 2011;34 Suppl 2:S101-6. [PMID: 21525438] doi:10.2337/dc11-s202
12. **Hirshberg B, Katz A.** Cardiovascular outcome studies with novel antidiabetes agents: scientific and operational considerations. Diabetes Care. 2013;36 Suppl 2:S253-8. [PMID: 23882054] doi:10.2337/dcS13-2041
13. **Hiatt WR, Kaul S, Smith RJ.** The cardiovascular safety of diabetes drugs–insights from the rosiglitazone experience. N Engl J Med. 2013;369:1285-7. [PMID: 23992603] doi:10.1056/NEJMp1309610
14. **White WB, Cannon CP, Heller SR, Nissen SE, Bergenstal RM, Bakris GL, et al; EXAMINE Investigators.** Alogliptin after acute coronary syndrome in patients with type 2 diabetes. N Engl J Med. 2013; 369:1327-35. [PMID: 23992602] doi:10.1056/NEJMoa1305889
15. **Fleming TR, Harrington DP.** Counting Processes and Survival Analysis. New York: J Wiley; 1991.
16. **Su JQ, Wei LJ.** Nonparametric estimation for the difference or ratio of median failure times. Biometrics. 1993;49:603-7. [PMID: 8369391]
17. **Wei LJ.** The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat Med. 1992;11: 1871-9. [PMID: 1480879]
18. **Zucker DM.** Restricted mean life with covariates: modification and extension of a useful survival analysis method. J Am Stat Assoc. 1998;93:702-9.
19. **Royston P, Parmar MK.** 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: 2409-21. [PMID: 21611958] doi:10.1002/sim.4274
20. **Zhao L, Tian L, Uno H, Solomon SD, Pfeffer MA, Schindler JS, et al.** Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. Clin Trials. 2012;9:570-7. [PMID: 22914867] doi:10.1177/1740774512455464
21. **Royston P, Parmar MK.** 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. [PMID: 24314264] doi:10.1186/1471-2288-13-152
22. **Tian L, Zhao L, Wei LJ.** Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostatistics. 2014;15:222-33. [PMID: 24292992] doi:10.1093/biostatistics/kxt050

23. **Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP.** SAS and R functions to compute pseudo-values for censored data regression. Comput Methods Programs Biomed. 2008;89:289-300. [PMID: 18199521] doi:10.1016/j.cmpb.2007.11.017

24. **Parner ET, Andersen PK.** Regression analysis of censored data using pseudo-observations. Stata J. 2010;10:408-22.

25. **Guyot P, Ades AE, Ouwens MJ, Welton NJ.** Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol. 2012;12:9. [PMID: 22297116] doi:10.1186/1471-2288-12-9

26. **Pfeffer MA, Burdmann EA, Chen CY, Cooper ME, de Zeeuw D, Eckardt KU, et al; TREAT Investigators.** A trial of darbepoetin alfa in type 2 diabetes and chronic kidney disease. N Engl J Med. 2009; 361:2019-32. [PMID: 19880844] doi:10.1056/NEJMoa0907845

27. **Parzen MI, Wei LJ, Ying Z.** Simultaneous confidence intervals for the difference of two survival functions. Scand J Stat. 1997;24:309-14.

**Current Author Addresses:** Dr. Uno: Department of Medical Oncology, Dana-Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02215.
Dr. Wittes: Statistics Collaborative, 1625 Massachusetts Avenue, Northwest, Suite 600, Washington, DC 20036.
Dr. Fu: Eli Lilly and Company, 893 South Delaware Street, Indianapolis, IN 46285.
Drs. Solomon, Claggett, and Pfeffer: Division of Cardiovascular Medicine, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115.
Dr. Tian: Department of Health Research and Policy, Stanford University School of Medicine, Palo Alto, CA 94305.
Drs. Cai, Evans, and Wei: Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, MA 02115.

**Author Contributions:** Conception and design: H. Uno, J. Wittes, H. Fu, L. Tian, S.R. Evans, L. Wei.
Analysis and interpretation of the data: H. Uno, H. Fu, B. Claggett, L. Wei.
Drafting of the article: H. Uno, J. Wittes, H. Fu, B. Claggett, L. Tian, S.R. Evans, L. Wei.
Critical revision of the article for important intellectual content: H. Uno, J. Wittes, H. Fu, S.D. Solomon, B. Claggett, L. Tian, M.A. Pfeffer, S.R. Evans, L. Wei.
Final approval of the article: H. Uno, J. Wittes, H. Fu, S.D. Solomon, B. Claggett, L. Tian, S.R. Evans, T. Cai, M.A. Pfeffer, L. Wei.
Statistical expertise: H. Uno, J. Wittes, H. Fu, B. Claggett, L. Tian, T. Cai, S.R. Evans, L. Wei.
Obtaining of funding: S.D. Solomon.
Administrative, technical, or logistic support: H. Fu.
Collection and assembly of data: H. Uno, S.D. Solomon.

## Appendix 1: Computer Programs for RMST

Computer programs to compare RMST between groups are available in 3 popular platforms (R [R Foundation for Statistical Computing], SAS [SAS Institute], and Stata [StataCorp]). We briefly describe the implementation with R (survRM2 packages). The R package is available from the Comprehensive R Archive Network Web site (http://cran.r-project.org/web/packages/survRM2/index.html). Similar program code is available for both SAS and Stata (http://bcb.dfci.harvard.edu/~huno/computer-program/).

For illustration, we use data from a randomized study (the primary biliary cirrhosis study) by the Mayo Clinic. The details of the study and the data elements are seen in the help file in the survival package. The sample data set used here can be loaded by the function *rmst2.sample.data()* in the surv2RM2 package. A listing of part of the sample data set can be found in **Appendix Table 2**.

Here, *time* is time from randomization to either death or censoring; *status* indicates the survival status (1 means dead, and 0 means alive); *arm* is the variable that indicates treatment assignment. In this sample, 0 denotes the placebo group and 1 represents the active treatment. The other 4 variables are covariates.

The following command implements the test of between-group differences based on RMST:

*rmst2(time, status, arm, tau = 10)*

Here, *tau* is the truncation time used in the RMST calculation. The output generated from this command can be found in **Appendix Tables 3** and **4**.

In this example, the difference in RMST (the first row of the "Between-group contrast" block in the output) was −0.137 year. The point estimate indicated that patients in the placebo group survive 0.137 year longer than those in the active treatment group on average when following up the patients for 10 years. Although no statistical significance was seen ( *P* = 0.74), the 95% CI (−0.939 to 0.665 years) was relatively tight around 0, which suggests that the difference in RMST would be at most ± 1 year. For more detailed illustrations, please see the package vignette that accompanies the survRM2 package.

## Appendix 2: Designing a Noninferiority Study With RMST Difference

We assume a Weibull distribution for the time to the composite CV events in the saxagliptin study (9). The observed data give shape and scale parameters for this Weibull distribution of 1.05 and 8573, respectively. The observed accrual rate for this study was about 30 patients per day. Moreover, at the end of this study, about 10% of patients had follow-up beyond 900 days. Assume that this Weibull distribution is the true model for the event times for both groups. The resulting RMSTs are about 860 days. Under this setting, for a range of potential numbers of study patients and a 1:1 treatment allocation, we generate 2000 sets of realizations with each sample size. This results in 2000 interval estimates for the difference between 2 RMSTs. We then calculate the chance that the upper bounds of these 2000 intervals fall below 18 days. If the chance is lower than 80%, we increase the current sample size and repeat the process. Then the final study sample size is chosen such that there is an 80% chance for the upper bound of the 95% CIs for the difference between the 2 RMSTs to be below 18 days. An alternative way to design the study is to fix the sample size but choose the timing of the end of the study. (For example, with fewer patients, we need more than 10% of patients whose follow-up would be beyond 900 days.) In **Appendix Table 5**, we report cases with various accrual rates. For example, when we enroll 2094 patients with an accrual rate of 30 patients per day, we will need a total of 908 days to confirm the noninferiority with an RMST difference. At the time of the analysis (for example, 908 days after the study activation), the expected total number of observed events is 182. The corresponding upper bound of the interval estimates for the HR is 1.52, which is much larger than 1.30.

## Appendix 3: Reconstructing Data From the Saxagliptin Study

Making inferences about the difference between 2 RMSTs requires an individual patient's event-time observations. The patient-level data from the saxagliptin study (9), however, are not publicly available. Therefore, we used an algorithm proposed by Guyot and colleagues (25) to reconstruct an individual-level event-time data set from the saxagliptin study using the information presented in that study. Of note, we used the software DigitizeIt, version 2.0 (I. Bormann), to scan the cumulative incidence curves with the reported numbers of patients at risk at various times to recreate the event-time observations. The reconstructed data led to cumulative incidence curves that are nearly identical to the originally published counterparts (not shown). Moreover, our reconstructed data yield a 95% CI for the HR of 0.89 to 1.12, which is identical to the interval reported in the saxagliptin study.

***Appendix Table 1.*** Glossary

*Cumulative incidence rate:* The probability that an event has occurred before a specific time point.

*Event-driven study:* A study with time to an event as the end point whose total information, study size, or study time is based solely on the number of observed events and not directly on the patients' exposure times or the number of patients involved.

*Hazard rate:* An instantaneous "force of mortality" ("mortality" is a generic term for an event in survival analysis) at a specific time point approximating the probability that an event-free patient would have the event in the next small period divided by the length of such a period (for example, a day or week); this rate may not be well-estimated empirically.

*Hazard ratio (HR):* Ratio of the hazard rates.

*Model-based between-group summary measure:* A population parameter for quantifying the difference between groups by imposing a specific relationship between 2 cumulative incidence functions; examples include the HR (assuming the HRs are constant over the entire study period) and the relative time (assuming that the ratio of percentiles of the 2 event time outcomes are constant); in the **Appendix Figure**, the slope in the top panel indicates the HR and the slope in the middle panel indicates the relative time.

- *Proportional hazards (PH) model:* The ratio of 2 hazard curves is assumed to be constant over the study duration.
- *Model-based relative time model (accelerated failure time):* The ratio of percentiles between 2 survival distributions is assumed to be constant over the study duration.

*Model-free between-group summary measure:* A population parameter for quantifying the between-group difference. This measure does not need to impose any relationship between 2 cumulative incidence curves; examples include the difference or ratio of RMSTs, the difference or ratio of *t*-year event rates, and the difference or ratio of median event times.

- *Difference or ratio of* t*-year event rates:* Difference or ratio of the event rates at a specific time point *t*; for example, the vertical distance between the 2 squares in the bottom panel of the **Appendix Figure** is the difference of 36-month event rates.
- *Difference or ratio of percentiles between 2 event times:* Difference or ratio of percentiles between 2 event-time distributions; the horizontal distance between the 2 closed circles in the bottom panel of the **Appendix Figure** is the difference of 2 median times.
- *Difference in or ratio of RMSTs:* For example, the shaded area in the bottom panel of the **Appendix Figure** is the difference in RMSTs at 48 months.

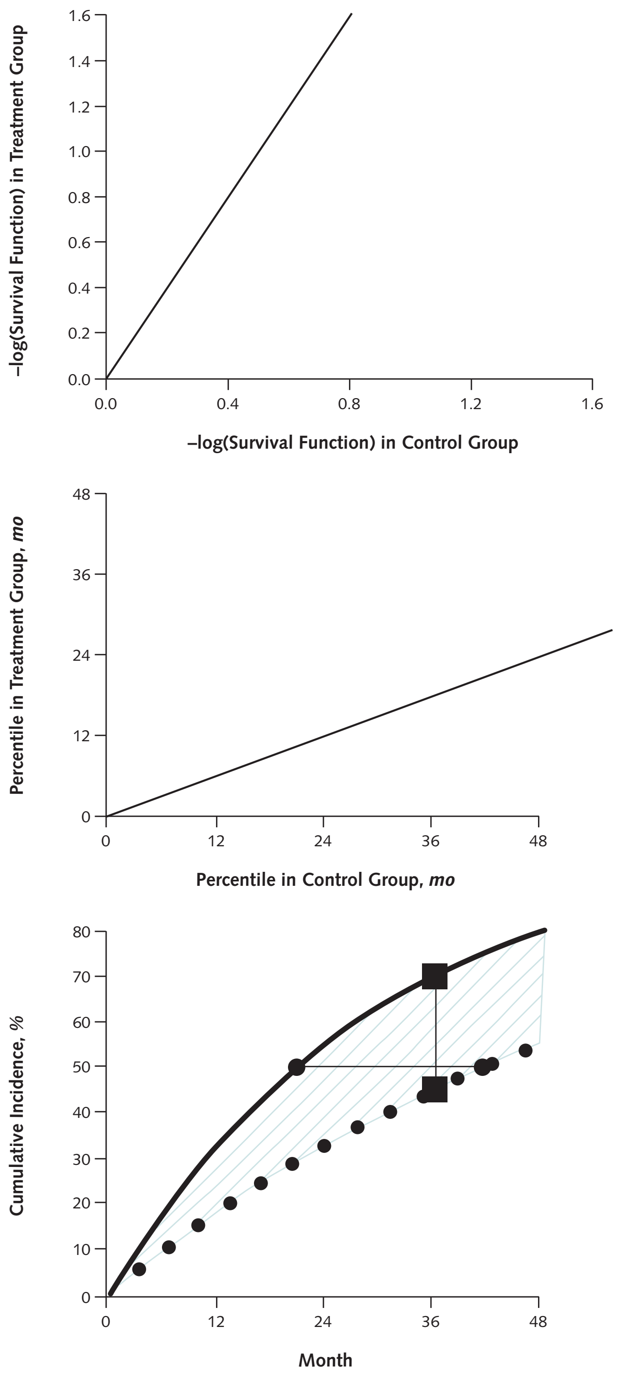
*95% simultaneous confidence band for a curve:* A collection of CIs over a time interval of interest such that the true curve (for example, the difference of 2 cumulative incidence functions) is entirely contained in the upper and lower boundaries of the band with a confidence level of 95%.

*Noninferiority margin:* A value for a between-group difference measure (for example, HR or the difference in RMSTs) under which a new treatment can be claimed to be noninferior to the control with respect to safety or efficacy.

*Percentile of the event time:* The time at which a given percentage of patients have had the clinical event.

*Restricted mean survival time (RMST) at a specific time point*: The average "survival" (event-free) time of the patient followed up to a specific time point, which is measured by the area above the cumulative incidence curve from 0 to this time point; also equivalent to the area under the survival curve.

***Appendix Figure.*** Graphical presentation of between-group difference metrics.



**Top.** Hazard ratio depicted by the slope of the line. **Middle.** Relative time depicted by the slope of the line. **Bottom.** Various model-free, between-group difference measures for a new treatment (*solid curve*) and a control (*dotted curve*). The distance between the 2 closed circles (*horizontal line*) is the difference of 2 medians, and the distance between the 2 closed squares (*vertical line*) is the difference of 2 event rates at 36 months. The shaded area is the difference in the RMST up to 48 months. RMST = restricted mean survival time.

***Appendix Table 2.*** Data Listing of a Part of the Sample Data Set

| Time | Status | Arm | Age | Edema | Bilirubin | Albumin | Protime |
|---|---|---|---|---|---|---|---|
| 1.095140 | 1 | 1 | 58.76523 | 1.0 | 14.5 | 2.60 | 12.2 |
| 12.320329 | 0 | 1 | 56.44627 | 0.0 | 1.1 | 4.14 | 10.6 |
| 2.770705 | 1 | 1 | 70.07255 | 0.5 | 1.4 | 3.48 | 12.0 |
| 5.270363 | 1 | 1 | 54.74059 | 0.5 | 1.8 | 2.54 | 10.3 |
| 4.117728 | 0 | 0 | 38.10541 | 0.0 | 3.4 | 3.53 | 10.9 |
| 6.852841 | 1 | 0 | 66.25873 | 0.0 | 0.8 | 3.98 | 11.0 |

***Appendix Table 3.*** RMST and RMTL, by Arm*

| | Estimate | SE | Lower 95% CI Bound | Upper 95% CI Bound |
|---|---|---|---|---|
| **RMST, by arm** | | | | |
| RMST (arm = 1) | 7.146 | 0.283 | 6.592 | 7.701 |
| RMST (arm = 0) | 7.283 | 0.295 | 6.704 | 7.863 |
| **RMTL, by arm** | | | | |
| RMTL (arm = 1) | 2.854 | 0.283 | 2.299 | 3.408 |
| RMLT (arm = 0) | 2.717 | 0.295 | 2.137 | 3.296 |

RMST = restricted mean survival time; RMTL = restricted mean time lost.
* The truncation time: tau = 10 was specified.

***Appendix Table 4.*** Between-Group Contrast*

| | Estimate | Lower 95% CI Bound | Upper 95% CI Bound | *P* Value |
|---|---|---|---|---|
| RMST (arm = 1) – (arm = 0) | −0.137 | −0.939 | 0.665 | 0.738 |
| RMST (arm = 1) / (arm = 0) | 0.981 | 0.878 | 1.096 | 0.738 |
| RMTL (arm = 1) / (arm = 0) | 1.050 | 0.787 | 1.402 | 0.738 |

RMST = restricted mean survival time; RMTL = restricted mean time lost.
* The truncation time: tau = 10 was specified.

*Appendix Table 5.* Upper 95% Confidence Bounds for HR and Risk Difference From a Noninferiority Study Designed Using RMST*

| Total Study Size, *n* | Accrual Rate Per Day, *d* | Study Duration, *d* | Total Events Observed, *n* | Estimated Upper Bound of the 95% CI for HR | Estimated Upper Bound of 95% CI for Risk Difference on 900 d, % |
|---|---|---|---|---|---|
| 2216 | 5 | 949 | 160 | 1.56 | 4.4 |
| 2172 | 10 | 924 | 176 | 1.53 | 4.0 |
| 2094 | 30 | 908 | 182 | 1.52 | 3.6 |

HR = hazard ratio; RMST = restricted mean survival time.

* Estimates of upper 95% confidence bounds for the HR and risk difference were calculated assuming the following fixed inputs: total sample size, accrual rate, and total study time. All configurations were determined to allow the estimated upper bound of the 95% CI for difference in RMST at 900 d to meet a noninferiority margin of 18 d (2% of the 900 d).