Restricted mean survival time for the analysis of cardiovascular outcome trials assessing non-inferiority: Case studies from antihyperglycemic drug development

David H. Manner, PhD,[a] Chakib Battioui, PhD,[b] Stefan Hantel, PhD,[c] B. Nhi Beasley, PharmD,[d] Lee-Jen Wei, PhD,[e] Mary Jane Geiger, MD, PhD,[f] J. Rick Turner, PhD, DSc,[g] and Markus Abt, PhD[h] *Indianapolis, IN; IN 46285, USA; Biberach, Germany; Silver Spring, MD; Boston, MA; North Wales, PA; Buies Creek, NC; and CH-4070, Basel, Switzerland*

Abstract Cardiovascular outcome trials (CVOTs) have been employed in multiple therapeutic areas to explore whether a noncardiovascular drug increases the risk for cardiovascular events. These studies are now a central part of drug development programs for antihyperglycemic drugs. These programs are expected to demonstrate that new antihyperglycemic drugs for patients with Type 2 diabetes do not have unacceptable cardiovascular risk. The hazard ratio, which is usually provided as evidence that patients receiving the investigational treatment are not at statistically significantly greater cardiovascular risk than patients on the control treatment, can be difficult to interpret for various reasons. Therefore, an alternative approach known as the Restricted Mean Survival Time (RMST) or τ-year mean survival time is presented, and its ability to overcome interpretation challenges with the hazard ratio discussed. The RMST approach is applied to five completed CVOTs and is compared with the corresponding hazard ratios. Additionally, detailed considerations are given on how to design a non-inferiority CVOT using the RMST approach. The RMST methodology is shown to be a practical alternative to the hazard ratio methodology for designing a non-inferiority CVOT. (Am Heart J 2019;215:178-86.)

Cardiovascular (CV) disease is the leading cause of morbidity and mortality in patients with Type 2 diabetes (T2D). These patients have a 2- to 4-fold increased risk of CV disease and a 3-fold increased risk of mortality.[1] Although a wide range of therapies is already available, developing new effective treatments for diabetes remains important to meet the needs of patients. Given concerns a decade ago that some therapies for T2D may increase CV risk (see Geiger et al[2]), international regulatory requirements for prospectively assessing the CV safety of new antidiabetic therapies to treat T2D have been put in place.[3,4] The ensuing investigations assist regulators in making decisions about market authorization, and knowing the effect of new treatments on CV risk also enables prescribers to make more informed decisions about the management of T2D.

Although the two guidances[3,4] differ in specific details, as noted shortly, the approaches discussed focus on the use of the hazard ratio (HR) to provide compelling evidence that the rate of occurrence of CV events in patients receiving the investigational treatment is not unacceptably greater than the rate for patients receiving the control treatment (typically a placebo). However, while the HR is often used to estimate treatment effect in a time-to-event analysis, its interpretation is not straightforward. Because of the limitations of this analytical strategy, interest in other methodologies has increased.

This White Paper, prepared by a Task Force consisting of members of the Cardiac Safety Research Consortium (CSRC), considers restricted mean survival time (RMST) analysis, an analysis approach for Cardiovascular Outcome Trials (CVOTs) that utilizes patients' exposure times in addition to the number of events in the analysis,

From the [a]Eli Lilly and Company, Lilly Corporate Center Drop Code 2240, Indianapolis, IN, [b]Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, USA, [c]Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Strasse 65, Biberach, Germany, [d]Food and Drug Administration, Center for Drug Evaluation and Research, Division of Cardiovascular and Renal Products, 10903 New Hampshire Ave, WO 22 - 4169, Silver Spring, MD, [e]Department of Biostatistics, Harvard University, 677 Huntington Ave, Boston, MA, [f]VP & Therapeutic Area Lead – Cardiovascular, Drug Development Services, ICON plc, 2100 Pennbrook Parkway, North Wales, PA, [g]Adjunct Professor of Pharmacy Practice, Campbell University College of Pharmacy & Health Sciences, P.O. Box 1090, 180 Main Street, Buies Creek, NC, and [h]F. Hoffmann-La Roche AG, Grenzacherstrasse 124, CH-4070, Basel, Switzerland.

Submitted May 30, 2019; accepted May 30, 2019.

Reprint requests: David H. Manner, PhD, Senior Research Advisor, Eli Lilly and Company, Lilly Corporate Center Drop Code 2240, Indianapolis, IN 46285, USA.
E-mail: mannerdh@lilly.com

0002-8703

https://doi.org/10.1016/j.ahj.2019.05.016

providing a more clinically meaningful interpretation of the results. The limitations of the HR analysis will be discussed. The RMST procedure is introduced and discussed, and its performance is compared with the HR method for studies designed to test non-inferiority.

The CSRC is a public-private partnership created to advance scientific knowledge in cardiac safety for new and existing medical products by building a collaborative environment based upon the principles of the FDA Critical Path Initiative and other public health priorities.[5] The opinions expressed in this paper do not represent regulatory policy.

Superiority and non-inferiority study designs in Clinical trials

The terms superiority and non-inferiority clinical trials are well known in the realm of drug efficacy and now find application in the realm of drug safety. The classical superiority study design, as developed for use in randomized, concurrently controlled clinical trials by visionaries such as Ronald Fisher and Paul Meier (see Turner[6]) employs a treatment arm and a control arm and determines if statistically significantly greater efficacy is associated with the treatment arm. However, efficacy alone is not the only factor that influences a drug's suitability for therapeutic use and the degree to which it will be successful in the market place if approved: other factors include its tolerability, safety profile, and cost.[7]

Consider a scenario in which the safety profile of a new drug is clearly better than that of an existing (reference) drug for the same indication, typically the existing gold-standard treatment for that indication, but its efficacy is slightly less. A somewhat lower degree of efficacy is acceptable in light of its better safety profile since the overall benefit–risk balance may be deemed (considerably) more favorable. The clinical question of relevance is: what degree of reduced (inferior) efficacy is acceptable? This is considered in the FDA Non-inferiority guidance.[8] The goal of a non-inferiority trial in this context is to look for compelling evidence that the new drug's efficacy is not worse by more than a specified margin, the non-inferiority margin. In some cases that margin can be the entire effect of the control drug (M1 M1), so that ruling out that margin shows that the new drug has some effect (greater than zero). When the clinical effect is important, a smaller margin (M2) is chosen ruling out an unacceptable degree of inferiority.

The degree of inferior efficacy we are prepared to accept is a clinical judgment, and one that must be made before a non-inferiority trial commences, and hence included in the study's Protocol or Statistical Analysis Plan. The associated statistical analysis will reveal whether or not compelling evidence of non-inferiority is demonstrated by the study. Readers are referred to Turner and Durham[7] for examples couched in terms of a new antihypertensive drug's efficacy. It should be appreciated that a new drug could be non-inferior according to the NI margin but actually less effective.

Attention now turns to the employment of non-inferiority designs in the safety realm. We recognize that CVOT trials as described in the 2008 FDA Guidance[3] are not non-inferiority trials but because they are designed similarly, that is, to demonstrate that a new antidiabetic drug for T2D does not have unacceptable increase in cardiovascular risk (analogous to the loss of effect that would be clinically acceptable in an NI trial), we refer to the CVOT trials as NI trials in this paper. The logic here is as follows: if a new antidiabetic drug shows clear evidence that it reduces blood sugar and hemoglobin A1C, we are prepared to accept the possibility of some increased risk of CV events. For initial approval, the FDA Guidance requires at a minimum that sponsors provide evidence that the upper bound of the two-sided 95 percent confidence (CI) interval placed around the HR point estimate (investigational treatment versus control treatment) for observed CV events in pooled Phase 2 and Phase 3 trials of the drug's clinical development program is below 1.8. This requirement translates to the prospective exclusion of excess cardiovascular risk of 80% or greater generally also with a reassuring point estimate. If the upper bound of the 95 percent CI for the HR is between 1.3 and 1.8, and the overall benefit–risk analysis supports approval, a post-marketing CVOT trial is likely to be required to further evaluate the CV risk profile of the new therapy employing a more stringent Hazard Ratio upper bound of 1.3. This requirement translates to the prospective exclusion of excess cardiovascular risk of 30% or greater. While the general goal of the EMA's Guideline is the same, explicit levels of what is cardiovascular harm are not provided therein.

Superiority considerations for diabetes cardiovascular outcome trials

Before continuing, it should be noted at this point that several trials have now demonstrated a favorable CV effect, i.e., a cardiovascular risk HR upper bound of the 95% CI less than 1.0, meaning that there is statistically significant evidence that the drug reduces CV risk compared to the control treatment. For example, EMPA-REG,[9] LEADER,[10] and SUSTAIN-6[11] have reported various types and degrees of CV benefit with the antidiabetic agents empagliflozin, liraglutide, and semaglutide, respectively (see Turner[12] for discussion). In December 2016 the FDA announced a new indication for empagliflozin to reduce the risk of cardiovascular death in adult patients with type 2 diabetes and established cardiovascular disease, the first time it had granted such an indication,[13] and there is discussion in the literature

regarding certain antihyperglycemic agents being used to lower cardiovascular and renal risk in patients without T2D.[14,15] That said, while CV benefit is certainly an attractive characteristic of a new antidiabetic drug, the primary focus of the required CVOT trials for new diabetes drugs in the United States and Europe remains exclusion of an unacceptable increase in CV risk, and accordingly our discussions now move back to that topic.

The nature of cardiovascular safety outcome trials

The primary endpoint of the CVOT is typically a composite outcome that includes CV death, nonfatal myocardial infarction (MI), and nonfatal stroke, commonly referred to as the major adverse cardiovascular events (MACE) composite endpoint. Sometimes, the composite endpoint may also include other events such as hospitalization for unstable angina: such expanded composites are generally referred to as MACE-plus.

While the HR is often used to estimate treatment effect in a time-to-event analysis, its interpretation is not straightforward for several reasons. First, the HR assumes proportional hazards, i.e., the ratio of the hazard rates for treatment versus control is constant over time. If this assumption is not met, the HR is difficult to interpret clinically. For this case, the estimated HR is not a simple average of hazards over time and it raises the question of its usefulness. Secondly, the HR is the ratio of two hazard rates. A reference hazard rate for the control arm is needed to understand the HR. The same hazard ratio may refer to a non-clinically meaningful difference if the hazard function in the control arm is small but may be clinically meaningful otherwise. Lastly, the precision of the HR estimate solely depends on the observed number of events, and does not consider the patients' exposure times during the trial. Therefore, if the risks for patients in both arms in the trial are low, the total number of patients needed for the CVOT is (very) large. That is why high risk patients are sought.

As an example, a two arm-trial designed to show that a new treatment has an acceptable CV risk (using the upper bound of 1.3, a two-sided Type 1 error rate of 0.05, and a power of 90%) requires 611 total events to be observed.[2] Even in diabetic patients considered at high risk for CV events, the yearly event rate for MACE is relatively low. Assuming an annual event rate (AER) of 3%, an enrollment period of 1.5 years, and a trial duration of 4 years, 6484 patients need to be enrolled to observe the required 611 events. More than 90% of the patients will be event-free during the trial, and hence do not contribute to the NI assessment via the HR.

Restricted mean survival time analysis.

This analytical approach utilizes the restricted mean survival time (RMST) or tau (τ)-year mean survival time as a summary measure. Fundamental aspects of this approach are captured here; detailed overviews of the RMST methodology are provided by Uno and colleagues.[16,17]

In this setting, the term survival denotes being free of MACE over a τ-year time window. The term τ is the point in time pre-specified in the design of the trial when the comparison between two treatment arms will occur. The choice of τ should have clinical relevance (that is, it represents a clinically meaningful follow-up duration) for the hypothesis being tested. For example, if the RMST for the standard of care evaluated over a τ = 3-year period is 2.5 years, this means that future patients treated by the standard of care with 3-year follow-up would enjoy 2.5 years, on average, without a MACE event. For the final analysis, all data from all patients (regardless of whether or not they experienced a MACE event) will be used to calculate RMST. The latter is a key point of distinction from the analysis of HR in which patients without events do not contribute to the precision of the point estimate.

To quantify the group difference, the difference or ratio of two RMSTs may be used. The RMST estimates use all of the patient's exposure time. Because RMST uses more available data, the analysis generally requires fewer patients than the HR counterpart, especially for low event rates (as discussed in more detail later).

Lastly, unlike the HR, the RMST procedure is model-free. That is, its validity does not need any model assumptions such as proportional hazards.

Measures of treatment effect based on RMST.

Various RMST-derived measures have been proposed to report the difference between investigational treatment (T) and control (C) arms in a trial. Using the example presented in Figure 1 for illustration, we describe several of these measures and provide a recommendation from this Working Group of a preferred measure for reporting results.

For the example presented in Figure 1, a constant AER of 4% has been assumed for the control arm; the treatment arm assumes an AER of 5.2%, which corresponds to a 30% increase in hazards, i.e., HR = 1.3. An overall trial duration of 60 months from enrollment of the first patient is assumed. For the control arm, RMST, being the mean survival time over the τ = 60 months period, amounts to 54.3 months. For the treatment arm, the assumed 30% increase in hazards results in a lower RMST of 52.7 months, meaning that, over a period of 60 months, patients would lose on average 1.6 months of event-free survival under the investigational treatment.

An alternative measure is the ratio of RMST values, defined as the ratio of the RMST for the investigational treatment arm to the RMST for the control arm. It represents the relative change in RMST between the two arms and amounts to 52.7/54.3 = 0.97, i.e., a 3% loss due

Figure 1

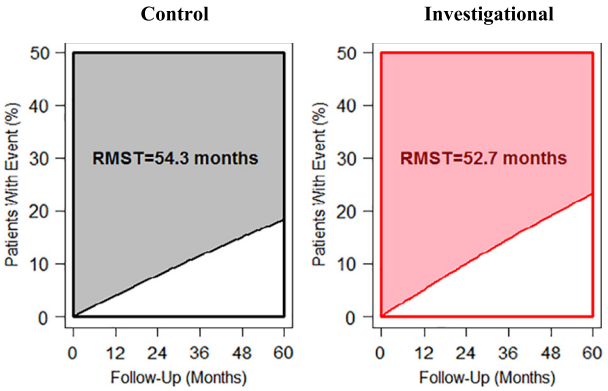


Illustration of RMST. RMST values are provided for both Control and Investigational treatment arms, assuming an exponential distribution with annual event rates of 4.0% and 5.2%, respectively, corresponding to a HR of 1.3.

to the investigational treatment. One concern with this measure is that interpretation of ratios is not straightforward and hence, similarly to HRs, the clinical interpretation of the results is difficult.

Another measure is the Integrated Risk Difference (IRD), expressing the difference in RMST between the two arms as a percentage of τ: that is;

$$\mathrm{IRD(in\%)} = 100 \left[\frac{\mathrm{RMST_T - RMST_C}}{\tau}\right] = 100 \left[\frac{52.7-54.3}{60}\right] = -2.7\%.$$

In this calculation, the IRD is interpreted as a 2.7% reduction in the survival time for the treatment arm relative to the control arm over the 60-month period. Alternatively, one could express the IRD in days per month (DpM): that is;

$$\mathrm{IRD(in\ DpM)} = 30.5 \left[\frac{\mathrm{RMST_T - RMST_C}}{\tau}\right] = 30.5 \left[\frac{52.7-54.3}{60}\right] = -0.81\ \mathrm{DpM}.$$

This result means that, on average, over a follow-up period of 60 months, 0.81 days of event-free survival time are lost each month under the investigational treatment. Although both IRD measures allow for comparison between studies that might use different tau's, using IRD expressed in days per month is more straightforward for evaluation of clinical relevance and hence the measure recommended by this group. These RMST measures incorporate both the exposure time and number of events.

Reanalysis of cardiovascular outcomes trials

Data from five randomized clinical trials—ALECARDIO,[18] SAVOR,[19] EXAMINE,[20] ORIGIN,[21] and TECOS[22]—that examined the CV safety of a new T2D treatment versus placebo were reanalyzed to compare the RMST methodology with the Cox Proportional Hazards (CPH) model. The primary endpoint of these trials was time to the first occurrence of a MACE event. The five CVOTs found there was not an increase in CV risk in the primary outcome between the new T2D treatments versus the placebo groups. The upper bound of the two-sided 95% CI for the HR for each of the trials was below the NI boundary of 1.3 required by the FDA guidance.[3] These 5 studies are representative of diabetes CVOTs to discharge CV risk.

A summary of the basic characteristics of these five exemplar CVOTs is presented in Table I.

The analysis of the ALECARDIO trial was performed on the original patient-level data, while data were reconstructed for the remaining four trials from the published Kaplan–Meier survival curves or the cumulative

Table I. Summary of the characteristics of the 5 exemplar CVOTs

Trial	Drug Name	Drug Class	Primary Outcome	Sample Size	Median Follow-up (years)
SAVOR-TIMI 53	Saxagliptin	DPP-4 Inhibitor	MACE	16,492	2.1
EXAMINE	Alogliptin	DPP-4 Inhibitor	MACE	5380	1.5
ORIGIN	Glargine	Insulin	MACE	12,537	7.0
TECOS	Sitagliptin	DPP-4 Inhibitor	MACE-plus	14,671	3.0
ALECARDIO	Aleglitazar	Dual PPAR Agonist	MACE	7226	2.0

incidence rate curves using the Guyot[23] algorithm. HRs with their corresponding 95% CIs using the reconstructed datasets showed very close proximity to the published results (data not shown).

For each of the five trials, the Cox Proportional Hazards (CPH) model was applied on the entire patient-level data, and the treatment effect (investigational treatment versus control) was estimated using the HR with the corresponding 95% CI. For example, in the ALECARDIO trial, the estimated HR was 0.96 (CI: 0.83, 1.11). The ALECARDIO trial would have fulfilled the FDA's requirement that the upper bound of the confidence interval is less than 1.3, satisfying the NI criterion.

In this trial, 7226 patients were randomized to receive either Aleglitazar or placebo in a 1:1 ratio. Patients were followed for up to 3 years. A total number of 704 events occurred by the end of the trial, 344 (9.5%) and 360 (10.0%) events in the Aleglitazar and placebo arms, respectively. Figure 2 shows the cumulative incidence curves for Aleglitazar (treatment arm) and placebo (control arm) from the ALECARDIO trial. For illustrative purposes, RMST was applied on ALECARDIO actual patient-level data, with $\tau = 30$ months. Several RMST statistics were recorded and compared with the HR, as shown in Table II.

The observed RMST difference (aleglitazar minus placebo) was 1.2 days and the lower limit of the two-sided 95 percent confidence interval was about 8 days. This means that on average, over a period of 30 months, future patients treated with Aleglitazar would at worst be free of CV events 8 days less than their corresponding placebo patients. Similarly to the interpretation of the HR, these results assume, on average, adherence as observed in the clinical trial.

Uno and colleagues' method[17] was followed to investigate the opportunity of designing the ALECARDIO trial with a smaller sample size if RMST would be pre-specified as the primary analysis method instead of the HR. Random sampled occurred for 2890 (40% of 7226) and 1444 (20% of 7226) patients 1000 times without replacement from the original dataset. The HR and RMST statistics in each of the 1000 samples were estimated and then averaged over all the samples. By fitting the CPH model to 20% of the original data, the HR method did not, on average, meet the 1.3 NI margin. On the other hand, if the NI margin based on RMST difference were to be 20 days out of 30 months according to the clinical rational, the ALECARDIO trial could have demonstrated safety of CV risk with only 20% of the original sample size. These results are summarized in Table II.

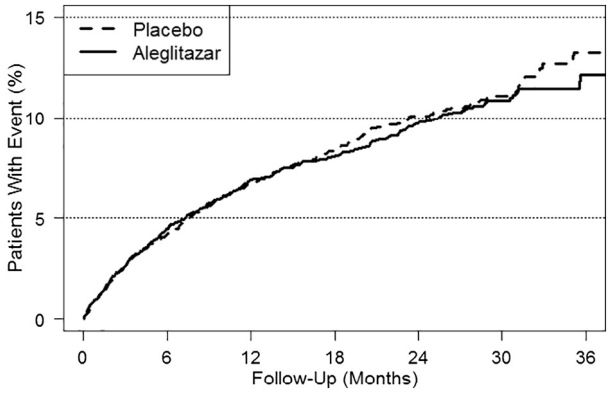
We conducted the similar RMST analyses on the reconstructed datasets from SAVOR, EXAMINE, ORIGIN, and TECOS. As shown in Table III, similar results were derived from these analyses, suggesting that the use of RMST with a much smaller sample size of patients can demonstrate the CV safety of these new type 2 diabetes treatments. Therefore, by applying RMST methodology to these five CVOTs, a much smaller sample size could have been used to demonstrate CV safety.

Table II. RMST and HR Results Estimated from ALECARDIO (τ = 30 months) using subsamples of different size sampled from the entire population of 7226 patients

Method	Entire data (95% CI)	40% data (95% CI)	20% data (95% CI)
Hazard ratio	0.96 (0.83, 1.11)	0.96 (0.76, 1.22)	0.96 (0.69, 1.34)
RMST difference (days)	1.2 (−8.3, 10.7)	1.3 (−13.8, 16.3)	1.2 (−20.0, 22.4)
IRD (%)	0.13 (−0.91, 1.17)	0.14 (−1.50, 1.79)	0.13 (−2.19, 2.46)
IRD (DpM)	0.04 (−0.28, 0.36)	0.04 (−0.46, 0.54)	0.04 (−0.67, 0.75)

DpM, days per month.

Figure 2



Cumulative incidence curve for Aleglitazar (Treatment arm) and Placebo (Control arm) from the ALECARDIO trial.

Table III. RMST and HR results estimated based on reconstructed datasets from four CVOTs (SAVOR, EXAMINE, ORIGIN, TECOS)

Method	Entire data (95% CI)	40% data (95% CI)	20% data (95% CI)
SAVOR (Saxagliptin), N = 16,492; τ = 900 days (~30 months)			
Hazard ratio	1 (0.89,1.12)	1 (0.84, 1.20)	1 (0.78, 1.31)
RMST difference	0 (–5, 5)	0 (–7, 7)	0 (–10, 10)
IRD (%)	0 (–0.51, 0.51)	0 (–0.77, 0.77)	0 (–1.11, 1.11)
IRD (DpM)	0 (0.15, 0.15)	0 (–0.23, 0.23)	0 (–0.39, 0.39)
EXAMINE (Alogliptin), N = 5380; τ = 900 days (~30 months)			
Hazard ratio	0.95 (0.81, 1.12)	0.96 (0.75, 1.24)	0.96 (0.67, 1.36)
RMST difference	4 (–9, 17)	4 (–17, 23)	5 (–24, 33)
IRD (%)	0.4 (–1.03, 1.85)	0.4 (–1.88, 2.55)	0.5 (–2.66, 3.66)
IRD (DpM)	0.12 (–0.31, 0.56)	0.12 (–0.57, 0.77)	0.15 (–0.81, 1.12)
ORIGIN (Glargine), N = 12,537; τ = 2555 days (~85 months)			
Hazard ratio	1.03 (0.94, 1.12)	1.03 (0.90, 1.18)	1.03 (0.85, 1.26)
RMST difference	–8 (–29, 12)	–8 (–40, 25)	-8 (–54, 38)
IRD (%)	–0.31 (–1.13, 0.47)	–0.31 (–1.56, 0.97)	–0.31 (–2.11, 1.48)
IRD (DpM)	–0.09 (–0.34, 0.14)	–0.09 (–0.46, 0.3)	–0.09 (–0.64,0.45)
TECOS (Sitagliptin), N = 14,671; τ = 1440 days (~48 months)			
Hazard ratio	0.98 (0.89, 1.08)	0.99 (0.85, 1.15)	0.99 (0.8, 1.23)
RMST difference	0 (–10, 11)	0 (–17, 17)	0 (–24, 23)
IRD (%)	0 (–0.69, 0.76)	0 (–1.18, 1.18)	0 (–1.66, 1.6)
IRD (DpM)	0 (–0.21, 0.23)	0 (–0.35, 0.35)	0 (–0.5, 0.49)

DpM, days per month.

Designing a non-inferiority CVOT using RMST

In this section, we discuss more detailed considerations for trial protocols when designing a CVOT for NI using RMST. Traditionally, based on the CPH model, the primary analysis is designed to establish that an increase in risk (the "hazard") by more than a pre-defined margin (eg, 1.3 or 1.8[3]) can be ruled out. Using the RMST methodology, the primary analysis attempts to establish that the loss in mean survival time over time τ does not exceed a pre-defined margin. The margin should preferentially be expressed relative to τ in days per year or days per month. Common to both settings is the primary endpoint, the time to the first occurrence of MACE.

If we assume the AER is 4% with equal allocation between the two arms (control and investigational treatment arm), an accrual period of 1.5 years, no drop-outs, a total trial duration of 4 years, and 90% power with a two-sided type 1 error rate of 5%, the trial would require 4922 patients to be randomized to reach 611 events to achieve NI based on a margin of 1.3 for the HR: see Table IV.

Suppose that for designing the same trial, RMST is used to analyze the primary endpoint. For a total trial duration of 4 years one may choose τ = 4 years. The endpoint will be expressed as integrated risk difference (IRD) in "Days per Year" or "Days per Month" as described in Section 2. The null and alternative hypotheses associated with the primary NI analysis may then be written as follows:

H_0: $IRD_{T\text{-}C} \le -7.9$ days per year versus H_1: $IRD_{T\text{-}C} \gg -7.9$ days per year.

The non-inferiority margin of 7.9 days per year (approximately 0.66 days per month) was in this case chosen to correspond to the NI margin of 1.3 assuming a 4% AER and exponential distribution. Over a 4-year trial, the NI margin thus ensures that the loss in mean event-free survival will not exceed about 1 month or less with 97.5% confidence. To achieve 90% power for the NI analysis, 4350 patients would need to be randomized (Table IV).

Although the NI margin for the RMST analysis previously discussed has been selected as the one corresponding to HR = 1.3 under the CPH model, this is for illustration only and is not intended as a suggestion of how NI margins for an RMST based analysis should be derived. Under CPH, to facilitate the discussion of NI margins in teams, HRs are frequently converted to eg, event-free rates when designing CVOTs, and therefore the discussion of clinical relevance is disconnected from the statistical analysis approach. In contrast to that, the RMST methodology is expected to better support discussions with clinical and other scientific experts of what a clinically irrelevant worst-case loss in mean survival time may be, which directly corresponds to the statistical analysis.

In fact, in absence of guidance, clinical discussions around NI are often held in terms of median or mean time to event or event rates at a certain landmark – and only then "back transformed" to the HR scale. Thus, there is a discrepancy between the discussions in terms of more clinically meaningful quantities and the later statistical formulation of the NI test. The advantage of the RMST concept is that it enables the clinical discussion as well as the statistical analysis to be on the same scale. Figure 3 illustrates the correspondence between the NI margins

Table IV. Required sample sizes to reach 90% power for non-inferiority under the CPH model (margin 1.3 for the hazard ratio) and when using RMST (margin 7.5 days per year)

Annual event rate	Analysis model	Number of patients	Number of events	Power
4%	CPH	4922	611	90%
4%	RMST	4350	540	90%

for the RMST based IRD measure expressed in days per year that correspond to an NI margin of 1.3 for the HR.

In preparation for discussion of the RMST methodology with clinical trialists who may not yet be familiar with the concepts, a more detailed understanding of the properties may be required. To offer further insights, we examined the power of an RMST based NI analysis for a range of AER, enrollment times and enrollment patterns. In Table V, we varied the AER by assuming that trial duration and enrollment times are fixed at 4 years and 1.5 years, respectively. Using RMST, with a total sample size of 4350 patients the power decreases with increasing event rate, and hence an increasing number of events. Statistically this is a result of greater variability in the estimate of the RMST when the AER increases from 3% to 5%. If the event rate is low, the variance of estimated event rate and hence the variance of RMST is small, but when the event rate approaches 0.5, then the variance increases. Consequently, the trialist should be careful not to underestimate the expected AER at the design stage.

It is worth emphasizing though that all results in Table V have been based on the same NI margin of 7.9 days per year. Assuming this margin has been considered clinically meaningful for an AER of 4%, it may no longer be if the AER is smaller or bigger. From Figure 3, for AER of 3% and 5%, the NI margin corresponding to a HR of 1.3 is 6.1 and 9.6 days per year. If these NI margins had been used for designing the trial, with the same number of 4350 patients, the power to conclude NI becomes 82% and 95%, respectively.

This may give rise to the discussion of what may happen to trial power when the assumed event rate originally stipulated in a protocol turns out to be different

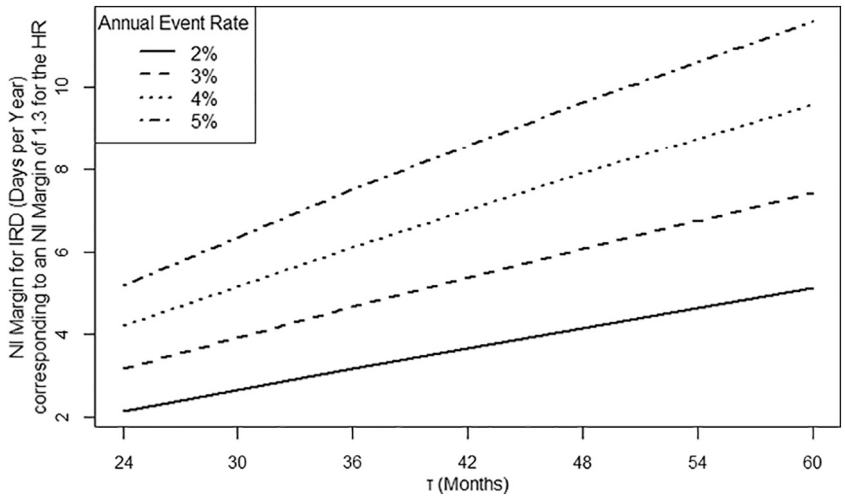
Table V. Power of an RMST based non-inferiority analysis for different annual event rates assuming enrollment of 4350 patients over a period of 1.5 years and total trial duration of 4 years; the non-inferiority margin was fixed at 7.9 days per year

Annual event rate	Number of patients	Number of events	Power
3%	4350	409	96%
4%	4350	540	90%
5%	4350	666	85%

in the clinical trial. For a trial using the traditional CPH model, smaller or larger observed event rates in the trial compared to the protocol assumptions mean that the trial may take more or less time than expected to reach the pre-defined required number of events. When enrollment is still open, the risk that a trial will require considerably longer follow-up than originally planned may be mitigated by enrolling more patients. In the example used in Table IV, if the event rate is found to be only 3% rather than the assumed 4%, increasing enrollment from 4922 to 5722 patients and extending the total trial duration to 55 months will retain the power at 90%.

Under the RMST methodology, should the event rate appear to be larger than expected, Table V suggests that this may lead to less power for the same NI margin. Also here, the trialist has the possibility to mitigate this by enrolling more patients. For example, if the actual AER in a trial were 5% rather than 4% as assumed when the trial with 4350 patients was planned, extending enrollment by 3 months for a total of 5075 patients again achieves a power of 90%. Note that, different from the situation based on the HR above, τ and hence the trial duration

Figure 3



Non-inferiority margins for the RMST-based integrated risk difference (expressed in days per year) corresponding to a non-inferiority margin of 1.3 for the hazard ratio.

does not increase. Of course, τ has to be agreed upon when the trial is designed and cannot be changed later.

When designing CVOTs, the assumptions for the AER at the planning stage are often based on data from previous studies. With ever improving patient care, the observed event rates in clinical trials are mostly smaller than what was assumed when these same trials were planned. Hence, a detrimental effect on power is unlikely. Although it may be tempting in certain situations, it is well known that a non-inferiority margin may not be changed after a design has been agreed and a trial started. Likewise, switching from a planned RMST based NI analysis to one using the CPH model (or vice versa), is generally not possible. Any such modifications imply a change to the originally stated null hypothesis, which means a change of the trial objective, and are therefore not acceptable.

Next, we examined the impact of different rates of enrollment. As discussed above, assuming a 4% AER and an NI margin of 7.9 days per year, enrollment of 4350 patients over a period of 1.5 years, and a trial duration of 4 years provides 90% power for an RMST based NI test with an expected number of 540 events. In case the same number of patients would be enrolled faster (in 1 year) or slower (over 2 years), this would lead to 580 and 501 events, respectively. Despite this difference, the power of the trial would be largely unchanged at approximately 90%.

When designing a CVOT to rule out a certain level of worsening by RMST, the total trial duration will need to slightly exceed τ to enable evaluation of the RMST over the interval from 0 to τ. For all simulations performed here we assumed the trial duration to exceed τ by 0.1 years, i.e., less than 1 month. Under the same assumptions used for Table IV, retaining $\tau = 4$ years but extending the trial duration to 4.5 or even 5 years, retains the power at 90%. For practical purposes, this implies that when closing a trial, a clinical cutoff may be chosen as a calendar date slightly (i.e., 2 weeks) after τ. All patients should then be contacted as soon as possible after the clinical cutoff and assessed for MACE events. This will enable the RMST analysis and also ensure that the data for analysis are complete and missing values are minimized. Of note, this approach is not unique to the RMST methodology but is recommended for all CVOTs.

Publicly accessible statistical software to perform all the computations using RMST is readily is available in the R package survRM2[24] and has been used throughout this article.

Concluding comments

In this study, the RMST methodology has been shown to be a practical alternative to the hazard ratio methodology when designing a CVOT to show an acceptable risk of CV harm. The RMST methodology has been applied to five completed diabetes CVOTs to demonstrate the difference in sample sizes and interpretation for assessing risk relative to the hazard ratio methodology. The sample sizes needed using RMST to show an acceptable CV risk were greatly reduced when using a clinically meaningful threshold. The RMST results from these analyses, such as RMST difference or IRD expressed in days per month, allow easy interpretation of the results.

As effective therapies for CV disease continue to be developed and incorporated into clinical practice, the event rates will continue to decrease. Consequently, using the traditional Hazard Ratio concept, a larger number of patients and/or longer trial durations will be needed to accrue the required number of MACE events. RMST can provide a clinically meaningful assessment of cardiovascular harm in a timely manner, likely with fewer patients.

By describing the results of a CVOT in terms such as the average days difference between the investigational treatment and control arm, the interpretation of the RMST results are straightforward. RMST will offer an opportunity to have discussions in clinical teams about the acceptable worst-case loss on a clinically meaningful scale, something to which patients and clinicians should be able to relate more readily than hazard ratios.

The focus of this paper has been ruling out increased risk compared to a control in diabetes CVOTs using the RMST methodology. Superiority can also be tested using RMST methodology. Examples in the literature of RMST methodology being applied to test superiority are provided.[25,26] Furthermore, as discussed throughout this paper, the RMST methodology can be used in other therapeutic areas to assess safety by ruling out a certain level of worsening.

Although the use of hazard ratios is well established and commonly used in the literature, methods such as RMST should be considered when an opportunity to answer an important clinical question about a medication using a statistically sound methodology that provides results in an intuitive and timely manner is available. The establishment of RMST as a common primary analysis for a CVOT is unlikely to be simple or quick. Therefore, sponsors and regulators should be encouraged to actively propose and discuss the RMST approach. It is hoped that the considerations in this paper will provide some additional encouragement.

Acknowledgments

The authors thank Dr. Robert Temple, Dr. Tzu-Yun McDowell, Dr. Norman Stockbridge, and Professor Lu Tian for their insightful comments that contributed to this manuscript.

Disclosures

Dr Manner is an employee and owns stock in Eli Lilly and Company.

Dr Battioui is an employee and owns stock in Eli Lilly and Company.

Dr Hantel is an employee of Boehringer Ingelheim Pharma GmbH & Co. KG.

Dr Beasley has no relevant disclosures.

Dr Wei has no relevant disclosures.

Dr Geiger is an employee of ICON and member of the Cardiac Safety Research Consortium's Executive Committee.

Dr Turner is a member of the Cardiac Safety Research Consortium's Executive Committee.

Dr Abt is an employee and owns stock in F. Hoffmann-La Roche Ltd.

References

1. Meigs JB. Epidemiology of cardiovascular complications in type 2 diabetes. Acta Diabetol 2003;40:S358-61.
2. Geiger MJ, Mehta C, Turner JR, et al. Clinical development and statistical approaches to assessing cardiovascular risk of new type 2 diabetes therapies. Ther Innov Regul Sci 2015;49:50-64.
3. FDA Guidance for Industry. Diabetes mellitus– evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. Available at, http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm071627.pdf December 2008. Accessed 11 April 2019.
4. European Medicines Agency. Guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus. Available at, http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129256.pdf May 2012. Accessed 11 April 2019.
5. Turner JR, Kowey PR. Rodriguez, et al: on behalf of the Cardiac Safety Research Consortium. The Cardiac Safety Research Consortium enters its second decade: an invitation to participate. Am Heart J 2016;177:96-101.
6. Turner JR. The 50th anniversary of the Kefauver-Harris Amendments: efficacy assessment and the randomized clinical trial. J Clin Hypertens (Greenwich) 2012;14:810-5.
7. Turner JR, Durham TA. Must new drugs be superior to those already available? The role of non-inferiority clinical trials. J Clin Hypertens (Greenwich) 2015;17:319-21.
8. FDA Guidance for Industry. Non-inferiority clinical trials to establish effectiveness. November 2016. Available at: https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf (Accessed 11 April 2019)
9. Zinman B, Wanner C, Lachin JM, et al. EMPA-REG OUTCOME Investigators. Empagliflozin, cardiovascular outcomes, and mortality in type 2 diabetes. N Engl J Med 2015;373:2117-28.
10. Marso SP, Daniels GH, Brown-Frandsen K, et al. LEADER Steering Committee; LEADER Trial Investigators. Liraglutide and cardiovascular outcomes in type 2 diabetes. N Engl J Med 2016;375:311-22.
11. Marso SP, Bain SC, Consoli A, et al. SUSTAIN-6 Investigators. Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. N Engl J Med 2016;375:1834-44.
12. Turner JR. Integrated cardiovascular safety: multifaceted considerations in drug development and therapeutic use. Expert Opin Drug Saf 2017;16:481-92.
13. News Release FDA, 2nd December. FDA approves Jardiance to reduce cardiovascular death in adults with type 2 diabetes. Available at, http://www.fda.gov/newsevents/newsroom/pressannouncements/ucm531517.htm 2016. Accessed 12th April 2019.
14. Raschi E, Poluzzi E, Marchesini G, et al. Dapagliflozin and cardiovascular outcomes: anything else to DECLARE? Expert Opin Pharmacother 2019;20:1087-90.
15. Herrington WG, Preiss D, Haynes R, et al. The potential for improving cardio-renal outcomes by sodium-glucose co-transporter-2 inhibition in people with chronic kidney disease: a rationale for the EMPA-KIDNEY study. Clin Kidney J 2018;11:749-61.
16. 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;32:2380-5.
17. Uno H, Wittes J, Fu H, et al. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in non-inferiority studies. Ann Intern Med 2015;163:127-34.
18. Lincoff AM, Tardif JC, Schwartz GG, et al. AleCardio Investigators. Effect of aleglitazar on cardiovascular outcomes after acute coronary syndrome in patients with type 2 diabetes mellitus: the AleCardio randomized clinical trial. JAMA 2014;311:1515-25.
19. Scirica BM, Bhatt DL, Braunwald E, 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.
20. White WB, Cannon CP, Heller SR, et al. EXAMINE Investigators. Alogliptin after acute coronary syndrome in patients with type 2 diabetes. N Engl J Med 2013;369:1327-35.
21. Gerstein HC, Bosch J, Dagenais GR, et al. The ORIGIN Trial Investigators. Basal insulin and cardiovascular and other outcomes in dysglycemia. N Engl J Med 2012;367:319-28.
22. Green JB, Bethel MA, Armstrong PW; for the TECOS Study Group. Effect of sitagliptin on cardiovascular outcomes in type 2 diabetes. N Engl J Med 2015;373:232–242.
23. 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.
24. Uno H, Tian L, Cronin A, Battioui C, Horiguchi M. Comparing Restricted Mean Survival Time. Package 'survRM2'. 2017. Available at: , https://cran.r-project.org/web/packages/survRM2/survRM2.pdf (Accessed 11 April 2019)
25. 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.
26. Pak K, Uno H, Kim DH, et al. Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. JAMA Oncol 2017;3:1692-6.