Choice of treatment-effect measures when noninferiority margins originally defined in absolute difference translated into relative difference influenced the results of clinical trials

Xianmixikemaier Abulizi*, Philippe Flandre
Sorbonne Universités, UPMC Univ Paris 06, INSERM, Institut Pierre Louis d’épidémiologie et de Santé Publique (IPLESP UMRS 1136), F75013, Paris, France

Abstract

Objectives: The aim of this study was to investigate the use of three treatment-effect measures in the analysis of randomized trials using a time-to-event endpoint and assess their influence on the results.

Study Design and Setting: A recent equivalence trial showed discordant results with the use of different measures. Different hypotheses may explain such discordant results including a mistaken hypothesized distribution of time to failure and an overestimation of failure rates in the protocol. In a simulation study, we investigate different situations in comparing analyses based hazard ratio (HR), risk difference estimated by Kaplan–Meier curves, and difference in restricted mean survival time. We also compared these three analyses on genuine data from a recent equivalence trial.

Results: In the equivalence trial, two analyses would have concluded equivalence, whereas the original analysis based on HR estimate did not declare equivalence. Results of our simulation study indicate little to moderate differences between the three analyses when the true distribution of time to failure is different to the hypothesized distribution. The main discordant results are found when failure rates have been overestimated or underestimated regardless of the distribution.

Conclusion: With the distributions investigated, differences between analyses based on different measures are much more driven by mistaken hypothesized failure rates than by the shape of the distribution of time to failure.

Keywords: HIV; Effect measures; Time-to-event endpoint; Survival analysis; Simulation study

1. Introduction

Many randomized clinical trials used a time-to-event primary endpoint to compare efficacy or safety of two or more treatments. Clinicians and investigators would like to summarize and represent their data as simply as possible but also want that the effect measure depict the significance of the intervention being studied in the clinically useful way. Different effect measures can be employed to report results of such studies including the relative risk (RR) and the risk difference (RD). The former is a measure of relative effect, whereas the latter is a measure of absolute effect [1–5]. Quite different in nature, there is not a clear consensus between investigators concerning the choice of effect measure. The selection of the most appropriate effect measure has been long debated, but no general guidelines have yet been developed for a single study as well as for meta-analyses [2,6–8]. The choice of the effect measure is of paramount importance when different measures lead to different conclusions for a given trial.

In clinical trials with a time-dependent event, the treatment effect is often estimated by the hazard ratio (HR) from a Cox proportional hazards model. The RD could be assessed by the difference of cumulative failure probabilities at a given time estimated by Kaplan–Meier curves (RD_KM). The RD_KM provides a clinically interpretable summary of differences between groups at a specific time and does not require a parametric assumption for the outcome distribution in the two groups, but it presents only a comparison at a given time point without reflecting the whole difference [9]. An HR is a simple and generally well-understood measure; however, the magnitude of the HR depends on the length of follow-up and which may
What is new?

**Key findings**
- Incorrect hypotheses of failure rates and distribution of time to failure have an impact on the performance of analyses based on hazard ratio, risk difference estimated by Kaplan–Meier curves (RDKM), and restricted mean survival time difference in trials using a time-to-event primary endpoint.

**What this adds to what was known?**
- The influence of incorrect hypotheses of failure rates and distribution of time to failure might not be negligible in noninferiority trials.
- In the distributions investigated, differences between analyses based on different measures are much more driven by mistaken hypothesized failure rates than by the shape of the distribution of time to failure.

**What is the implication and what should change now?**
- Caution is needed in the choice of effect measures in trials when the hypothesis of failure rate or distribution of time to failure is incorrect because these might have an impact on their results.
- Our recommendation is to use the RD KM analysis.

have a built-in selection bias [10]. In addition, HR may also be difficult to interpret when the proportional hazards assumption is violated [9]. The restricted mean survival time (RMST) has recently been proposed to estimate average event-free time from randomization to a suitable follow-up time point $t$. The measure is an estimate of the integral under the survival curve [9,11–13]. In general, the Kaplan–Meier method is used to estimate such a curve. RMST difference (Δ-RMST) provides a stable estimate of the difference and it is valid under any distribution of the time to event in the treatment groups of which the proportional hazard assumption is far from correct. The measure, however, also needs prespecification of the time point of interest so investigators could envisage a temptation to choose a specific time point so as to obtain more significant results [12].

Most randomized controlled trials including HIV-infected patients are noninferiority trials. In a noninferiority trial, null and alternative hypotheses are reversed because the null hypothesis is inferiority of the experimental treatment over the standard, and the alternative is noninferiority of the experimental over the standard treatment [14,15]. The noninferiority is demonstrated if the efficacy of the experimental treatment is not unacceptably worse than that of the standard based on a prestated margin of noninferiority [15,16]. Rejecting the null hypothesis (rejecting the inferiority of the experimental treatment) leads to conclude noninferiority. The noninferiority margin is defined according to the choice of the measure for measuring the treatment effect. In many trials including HIV-1-infected patients, the primary endpoint is most often the proportion of patients in success (or failure) at a given week, for example, week 48 [17]. In such trials, the RD, computed as a simple difference in proportions, is the summary effect measure, and noninferiority margins around 10–12% have been widely used. In trials using an HR analysis, the noninferiority margin is often based on RD due to the large number of previous trials using the RD measure and then subsequently translated into the HR metric. To carry out this translation, an assumption of the survival distribution is required. For simplicity, the exponential distribution considering a constant hazard assumption is employed.

A recent trial including HIV-infected patients showed discordant conclusions between analyses based on these two effect measures. Indeed, results based on RD KM at week 96 would have led to demonstrate equivalence, whereas results based on HR did not conclude to such equivalence [18,19]. Different hypotheses may explain such a discordance including (i) an inherent difference between analyses based on RD KM and HR, (ii) an overestimation of failure rates in the protocol, and (iii) a distribution of failure times that does not follow a simple exponential distribution. In this work, we aim to investigate the reasons that lead to such discordant conclusions in the use of different effect measures. We then compare analyses based on HR, RD KM, and Δ-RMST as summary effect estimates in a simulation study when sample sizes are computed from an HR-based analysis. The equivalence trial introduced above is also analyzed using the effect measure based on Δ-RMST.

2. Materials and methods

2.1. Motivating example

The AIDS Clinical Trials Group (ACTG) 5202 was an equivalence study comparing atazanavir + ritonavir (ATV/r) vs. efavirenz (EFV) for each of the two combinations abacavir/lamivudine (ABC/3TC) and tenofovir/empiricritabine (TDF/FTC) [18]. The primary efficacy endpoint was the time from randomization to virologic failure. The study team considered that a difference in the probability of virologic failure by week 96 of approximately 10% was acceptable for equivalence (see the Supplementary Appendix available with reference [20]). Assuming a probability of virologic failure of 31.9% by week 96 in the reference arm, a rate difference of 10% implies a probability of failure rate by week 96 in the other regimen of 41.6% (an exact difference of 9.7%). In other words, equivalence
margins for \( R_{\text{KM}} \) estimates by week 96 are \((-9.7\% \text{ to } 9.7\%)\). Assuming that the time to virologic failure follows an exponential distribution at a rate of 31.9\% by week 96 corresponds to an HR of 0.004 for the reference arm \([-\log (1-0.319)/96]\) and a rate of 41.6\% by week 96 to an HR of 0.0056 for the other regimen arm. Then margins for the HR-based analysis are 0.71 (0.004/0.0056) and 1.40 (0.0056/0.004), respectively [18]. A sample size of 450 subjects in each arm was computed to provide approximately a 90\% statistical power of rejecting the null hypothesis of inferiority (probability that the 95\% two-sided confidence interval [CI] upper bound for the HR will be less than 1.40), that is, concluding noninferiority.

Among the patients randomized in the ATV/r or EFV groups, with EFV as reference, the estimated HR was 1.13 (95\% CI, 0.82–1.56) in the group receiving ABC/3TC and 1.01 (95\% CI, 0.70–1.46) in the group receiving TDF/FTC [18]. Neither 95\% CIs met the prespecified equivalence boundaries of 0.71 and 1.40. Differences in Kaplan–Meier estimates by week 96 were also provided [18]. In particular, \( R_{\text{KM}} \) by week 96 was 1.9\% (95\% CI, –2.9 to 6.8) for patients randomized in the ABC/3TC group and was 0.8\% (95\% CI, –3.3 to 4.9) for patients randomized in the TDF/FTC group. We noted that both \( R_{\text{KM}} \) CIs are within the equivalence margin of \( \pm 9.7\% \) hypothesized in the protocol to compute the HR bounds [19]. Thus analyses based on \( R_{\text{KM}} \) would have led to equivalence for both comparisons, whereas HR-based analyses did not conclude to equivalence.

At least three hypotheses may explain such discordance. First, the difference is due to the nature of the two measures, one is a relative measure, whereas the other is an absolute measure. Second, the overestimation of the failure rate in the protocol would explain such a difference between the two measures. Indeed, computation of the sample size was based on a failure rate of 32\%, whereas the observed rates were 11–17\% [18]. Third, the time to virologic failure did not follow an exponential distribution that would lead to different results according to the measure used. Indeed, more virologic failures were observed before week 48 than after week 48 in violation of the constant hazard assumption. For example, in the ABC/3TC/EFV group, 72 virologic failures were observed over the full follow-up, 52 failure occurred before week 48, 11 failures occurred between weeks 48 and 96, and 9 failures occurred beyond week 96. Similar findings are found in the other treatment groups suggesting that the instantaneous risk may be not constant along the follow-up time.

### 2.2. Alternative analysis and ACTG 5202 data reconstruction

To reanalyze the ACTG 5202 with an RMST difference, we need to compute corresponding bounds for such an analysis. Using the same parameters than those defined in the ACTG 5202 protocol, we simulated data sets to estimate the upper limit that provides a probability that the 95\% two-sided CI upper bound for the \( \Delta \)-RMST is approximately 90\% (see below). Based on 5,000 replications, the upper bound for a difference in RMST was 5.48 weeks.

We have reconstructed ACTG 5202 individual-patient-level data by applying an algorithm developed by Guyot et al [21]. Briefly, we used the Digitizelt software to measure the time and survival probability co-ordinates on the Kaplan–Meier curves [22]. We extracted the numbers of patients at risk and the total numbers of events, when available. The data were then input into an algorithm on the basis of iterative numerical methods to solve the inverted Kaplan–Meier equations [21]. More details can be found in Guyot [21]. We also test the proportional hazards assumption in comparing ATV/r vs. EFV in both combinations ABC/3TC and TDF/FTC using a statistic based on scaled Schoenfeld residuals [23].

Those reconstructed data were then reanalyzed with HR, \( R_{\text{KM}} \), and \( \Delta \)-RMST as effect measures. The data were also analyzed using two parametric models: exponential and Weibull models. Those parameters were used to generate 1,000 data sets mimicking the ACTG 5202 data. Each simulated data set was analyzed according to the three different measures to compute the proportion of data set leading to conclude noninferiority.

#### 2.3. Simulation study

We conducted two series of simulation studies to compare analyses based on HR, \( R_{\text{KM}} \), and \( \Delta \)-RMST and to assess their influence on the results of trial. In our simulation study, we consider that an exponential distribution for the time to failure is assumed in the protocol and that the computation of the power and sample size is based on an HR analysis.

##### 2.3.1. Simulation study 1

The first series investigate the performance of the three measures when an exponential or a Weibull distribution is used to generate the data. We investigate low (10\%), moderate (30\%), and high (50\%) failure rates for the reference arm. As in the ACTG 5202, we consider \( \pm 10\% \) noninferiority bounds for the \( R_{\text{KM}} \) analysis by week 96. As above, bounds for the HR-based analysis are computed from failure rates using corresponding hazards from an exponential distribution. For example, a 10\% failure rate corresponds to an HR of 0.0011 in the reference arm, and a 20\% failure rate in the other arm corresponds to an HR of 0.00232, providing an upper bound of 2.12 (0.00232/0.0011) for an HR-based analysis. Bounds for analyses based on \( \Delta \)-RMST are computed by simulation after computation of the sample size (see below). Bounds are provided in the Appendix Table 1.

We determine the sample size corresponding to low, moderate, and high failure rates from an HR-based
analysis. Each sample size is computed to give approximately 90% probability of concluding noninferiority (probability of 90% that the 95% two-sided CI upper limit for the HR estimate will be less than the corresponding upper bound value). The above calculations make the following additional assumptions: (1) accrual takes 96 weeks after a uniform distribution, (2) follow-up continues for 96 weeks after the last entry, (3) time to lost to follow-up from entry follows also an exponential distribution, (4) 10% lost to follow-up at 96 weeks, and (5) equal allocation between the two arms. Sample sizes are provided in the Appendix Table 2.

Once sample sizes were obtained, bounds for RMST-based analyses were computed by simulation. Those bounds correspond to the mean difference in RMST between the two arms at week 96 based on 5,000 replications. For example, assuming a 10% failure rate in the reference arm and then a 20% failure rate in the other arm, a mean of 4.42 weeks was found for the difference in RMST by week 96 (see Appendix Table 1).

Once upper bounds and sample sizes were computed, we then generated a time-to-event outcome for each subject according to two different distributions for the time to failure: exponential and Weibull. For each subject, we generated a random number from a standard uniform distribution: \( u \sim U(0,1) \). Half of the subject is in the reference arm, the other half in the other arm. We generated a survival or event time for each subjects as follows: \(-\log(u)/\lambda^{1/\rho}\). For the exponential distribution, \( P = 1 \) and two values of \( P \) were investigated to explore monotonic decreasing \( (P = 0.75) \) or increasing \( (P = 1.5) \) hazard function from a Weibull distribution (see the Appendix). The hazard \( \lambda \) takes different values to provide the expected probability of failure rate by week 96 in the two arms. Thus data are generated according to the proportional hazards assumption. For a given level of failure rate in the reference arm (low, moderate, or high), data were generated with an increasing probability of failure in the other arm, providing by week 96 a difference in probability from 0% to 7%. For example, with a low failure rate, the probability of failure in the reference arm is 10%, whereas the probability of failure in the other arm varies from 10% to 17%.

2.3.2. Simulation study 2

The second series investigates the impact of an underestimation or overestimation of the probability of failure in a protocol. We consider a design assuming a 20% probability of failure by week 96 in the reference arm. Upper bounds and sample sizes are computed in a similar way than previously (see Appendix Tables 3 and 4). Survival data were also generated as previously with a sample size corresponding to a 20% failure rate by week 96. Investigating an overestimation (underestimation) of failure rates in the protocol, we simulated data providing a 10% (30%) probability of failure rate by week 96 in the reference arm. Data were simulated in a similar way than previously to provide difference in the probability of failure by week 96 between the two arms varying from 0% to 7%. Thus, investigating overestimation, the probability of failure by week 96 in the reference arm was 10% and varies from 10% to 17% in the other arm. Investigating underestimation, the probability of failure by week 96 in the reference arm was 30% and varies from 30% to 37% in the other arm.

In both simulation studies, for each situation we simulated 1,000 data sets. Within each simulated data set, we estimated the HR using a Cox proportional hazards model censoring the data by week 96. \( \text{RD}_{\text{KM}} \) estimate by week 96 and \( \Delta \text{-RMST} \) by week 96 from Kaplan–Meier survival curves between the two treatment groups. Contrary to the ACTG 5202 study, in our simulation studies only the follow-up up to week 96 was used to estimate the HR. Proportions of simulated trials that did not conclude noninferiority (did not reject the null hypothesis) were computed using corresponding upper bounds. These analyses were done using a discrete time corresponding to study evaluations in clinical settings. Discrete times were used to mimic the ACTG 5202 study. The observed time span in the simulated study was divided into a series of fine intervals so that there is a separate interval for each time of failure. All events occurring in the certain interval are fixed at the largest time point of that interval. For example, all events occurring before week 16 had their time to failure at week 16, and for all events occurring between weeks 16 and 24, had their time to failure at week 24, and so on. The discrete times to failure were at week 16 and every 12 weeks thereafter. Computations were performed using the SAS statistical software, version 9.4.

3. Results

Estimates of HR and \( \text{RD}_{\text{KM}} \) based on original or reconstructed data of the ACTG 5202 are similar (Table 1). Both 95% CI of \( \Delta \text{-RMST} \) between EFV and ATV/r for ABC/3TC and TDF/FTC met equivalence bounds. Therefore, all analyses based on \( \text{RD}_{\text{KM}} \) and \( \Delta \text{-RMST} \) would have led to equivalence for both comparisons, whereas analyses based on HR estimates did not meet the prespecified criteria for equivalence. Testing the proportional hazards assumption using a score test based on scaled Schoenfeld residuals leads to \( P = 0.45 \) in the ABC/3TC combination and \( P = 0.20 \) in the TDF/FTC combination. Then we cannot reject the proportional hazards assumption for both comparisons.

Parameters of the analysis of the reconstructed ACTG 5202 data with both exponential and Weibull models are given in the Appendix Table 5. Of note, the shape parameter of the Weibull distribution corresponds to a decreasing hazard for all treatment regimens (\( P \approx 0.65 \)). Results based on 1,000 simulated data using those parameters estimates are displayed in Table 2. In the ABC/3TC group, we found that the probability to conclude equivalence using an
HR-based analysis was around 24—27%, whereas it was greater than 90% with an analysis based on RDKM and Δ-RMST, whatever the survival distribution. Similar findings were found with a piecewise exponential distribution of time to failure (data not shown). Similar results were observed in the TDF/FTC groups although with smaller differences between analyses.

Results of the first series of the simulation study are displayed in Fig. 1. The y-axis provides the probability of not demonstrating noninferiority; that is, not rejecting inferiority. The power of rejecting the null hypothesis of inferiority corresponds to no difference in failure rates between the two arms and corresponds to one minus that probability. In Fig. 1A the distribution of failure time is exponential that corresponds to the distribution both assumed to compute sample sizes and boundaries to simulate the data. The main differences between the three analyses were obtained for low (panel A) failure rates. For low failure rates, analyses based on RDKM (HR) provided lower (higher) probabilities of not demonstrating noninferiority. For moderate and high failure rates in the reference arm, all analyses provided almost similar results and differences between analyses decreased with increasing failure rates.

HR-based analysis was around 24—27%, whereas it was greater than 90% with an analysis based on RDKM and Δ-RMST, whatever the survival distribution. Similar findings were found with a piecewise exponential distribution of time to failure (data not shown). Similar results were observed in the TDF/FTC groups although with smaller differences between analyses.

Table 1. Summary of the analyses in the ACTG 5202 study and analyses based on reconstructed data. With efavirenz as the reference in all comparisons

<table>
<thead>
<tr>
<th>Analysis according to the three treatment-effect measures</th>
<th>Equivalence bound</th>
<th>ABC/3TC</th>
<th>Conclusion</th>
<th>TDF/FTC</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR (95% CI)</td>
<td>(0.71—1.40)</td>
<td>EFV</td>
<td>Not equivalence</td>
<td>EFV</td>
<td>Not equivalence</td>
</tr>
<tr>
<td>K-M Difference at W96 (95% CI)</td>
<td>(−9.69 to 9.69)</td>
<td>(−2.9 to 6.8)</td>
<td>0.19 (−3.3 to 4.9)</td>
<td>0.82 (−3.24 to 4.88)</td>
<td></td>
</tr>
<tr>
<td>RMST difference at W96 (95% CI)</td>
<td>(−5.48 to 5.48)</td>
<td>−0.53 (−3.61 to 2.55)</td>
<td>−1.03 (−3.58 to 1.51)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: EFV, efavirenz; ATV/r, atazanavir + ritonavir; ABC/3TC, abacavir/lamivudine; TDF/FTC, tenofovir/emtricitabine; HR, hazard ratio; K-M, Kaplan-Meier; RMST, restricted mean survival time; CI, confidence interval.

Table 2. Probability that the 95% two-sided CI upper margin for the estimated HR will be <1.40 for full follow-up, RDKM will be <9.69% and RMST difference will be <5.48 weeks at week 96, with sample size of 450 subjects per arm; results of simulation based on the ACTG 5202 parameters

<table>
<thead>
<tr>
<th>Distribution model</th>
<th>HR &lt; 1.40</th>
<th>RDKM &lt; 9.69%</th>
<th>RMST difference &lt; 5.48 weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC/3TC group</td>
<td>24.80%</td>
<td>98.80%</td>
<td>99.90%</td>
</tr>
<tr>
<td>Exponential</td>
<td>27.00%</td>
<td>97.00%</td>
<td>99.30%</td>
</tr>
<tr>
<td>Weibull</td>
<td>40.50%</td>
<td>99.10%</td>
<td>99.60%</td>
</tr>
<tr>
<td>TDF/FTC group</td>
<td>38.60%</td>
<td>99.90%</td>
<td>99.90%</td>
</tr>
<tr>
<td>Exponential</td>
<td>38.60%</td>
<td>99.90%</td>
<td>99.90%</td>
</tr>
</tbody>
</table>

Abbreviations: ABC/3TC, abacavir/lamivudine; TDF/FTC, tenofovir/emtricitabine; HR, hazard ratio; RDKM, risk difference estimated by Kaplan-Meier curves; RMST, restricted mean survival time; CI, confidence interval.

ABC/3TC group = EFV + ABC/3TC vs. ATV/r + ABC/3TC. TDF/FTC group = EFV + TDF-FTC vs. ATV/r + TDF-FTC.
provide a 90% power with an HR-based analysis, whereas the simulation study provides a power of approximately 65% (Fig. 2A). Note that the ACTG 5202 study with an observed decreasing hazard and overestimation of the failure rate in the protocol corresponds to Fig. 2B left panel. Inversely, analyses based on RDkm and Δ-RMST provided higher probabilities of not declaring noninferiority in the case of an underestimation of failure rates. For each difference in the probability of failure rates, we also computed the mean and the standard deviation of the probability of not demonstrating noninferiority (Appendix Table 6). The results indicated that analyses based on RDkm provided much lower standard deviations compared with analyses based on the two other measures.

Fig. 1. Probability of not demonstrating noninferiority according to three effect measures used. In the figure, the probability of failure in the reference arm is 10% (panel A), 30% (panel B), and 50% (panel C), respectively. HR, hazard ratio; RDkm, risk difference in Kaplan–Meier estimates (by week); Δ-RMST, difference in restricted mean survival time estimates (by week).
4. Discussion

The objective of this work was to investigate differences between analyses based on HR, RD estimated by Kaplan–Meier curves and RMST difference in a genuine equivalence trial and in a simulation study. The HR-based analysis was considered as the reference analysis to compute sample sizes. This work was motivated by a recent equivalence trial, leading to different conclusions with the use of HR estimates or RD$_{KM}$ [18,19]. Our results based on reconstructed data showed that an RMST-based analysis would also have led to a different conclusion that the trial’s primary analysis is based on HR estimate. Such differences between effect measures need to be highlighted and more widely understood because they potentially explain differences in the results observed between trials using different measures. Of course, conclusions of a trial must be based on the analysis defined in the protocol. Similar findings, however, based on different measures reinforce conclusions of a trial especially when noninferiority margins are described according to different measures.

With the distributions and shape parameters investigated, our results showed that violation of the assumption of constant hazard rate had less impact in the difference observed between the analyses than mistaken levels of failure rates. Nevertheless, the difference due to the distribution shape increase with low failure rates and a stronger increasing hazard. In particular, an overestimation of the failure rates would likely lead to not demonstrating noninferiority with an HR-based analysis even with no or little difference between the two arms. Conversely, the power of demonstrating noninferiority is much larger with an HR-based analysis than with the two other analyses in a situation of underestimation. From the different situations of our simulation studies, it appears that the RD$_{KM}$ analysis provides more stable results to departures from the hypotheses of levels of failure rates or incorrect distribution of the time to virologic failure (see Appendix Table 6). Although other simulation studies can be carried out involving other situations, our results lead to recommend the RD$_{KM}$ analysis. This recommendation is valid only in our setting where a noninferiority margin for an RD measure is translated into an HR measure. Our recommendation does not cover other situations where original margins given for HR or RMST analyses would be transformed into RD measures. Further studies are required to investigate those latter situations. One difficulty, however, is to transform margins from a relative measure into margins for absolute measures.

In general, ACTG used HR-based analyses though RD$_{KM}$ was used in a recent study [19,24]. ACTG trials are very well designed and have main influence in the international guidelines for the use of antiretroviral agents. Most other trials, however, conducted by research groups or pharmaceutical companies used an RD at week 48 as primary efficacy analysis estimated by a simple difference of proportions [17]. Using results of those trials in designing a new study with an HR-based analysis, one needs to translate failure rates at a given time point into hazards assuming a statistical distribution of time to failure. For simplicity, the exponential distribution is widely used. In the ACTG 5202 study, results in the TDF/FTC group was particularly puzzled because there was no difference between EFV and ATV/r as indicated by an HR point estimate of 1.01 (95% CI, 0.70–1.46). Both lower and upper limits of the 95% CI were beyond corresponding bounds. This clearly indicates a much larger variance than expected for the log HR parameter in the proportional hazards model. The larger variance is due to a lower number of events than expected resulting in an overestimation of the hypothesized failure rates. Our simulation study with an overestimation of the assumed failure rates confirmed that finding. The ACTG 5202 team already pointed out that the low rate of virologic failure observed (11–17%) in their data compared with the hypothesized one (32%) could explain that equivalence was not declare [18]. Overestimation of failure rates or underestimation of success rates, has been previously described in HIV noninferiority trials, leading to overpowered trials [25]. We think that this point also influenced the design of a recent study because the ACTG 5257 trial used a RD$_{KM}$-based analysis to investigate equivalence between three antiretroviral regimens [24]. Results of that study showed virologic failure rates of 9% to 15%, whereas the assumed failure rate was 25%. Again there is an overestimation of the assumed failure rate although smaller than the one in the ACTG 5202 study. We may suspect that an HR-based analysis in the ACTG 5257 study would likely not conclude to equivalence, whereas equivalence was demonstrated with RD$_{KM}$-based analyses [24].

Testing the proportional hazards assumption in the ACTG 5202 data does not show violation of that assumption. Consequently, this cannot be used to rule out the HR although it is known that those tests have a low power to detect a violation of the proportional hazards assumption with a high level of censoring. Parametric distribution

---

### Table 3. Relative sample size for the three analyses with an assumed exponential distribution for the time to failure according to failure rates

<table>
<thead>
<tr>
<th>Distribution of time to failure</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$N_{KM} &lt; N_{RMST} &lt; N_{HR}$</td>
<td>$N_{KM} &lt; N_{HR} \approx N_{RMST}$</td>
<td>$N_{HR} \approx N_{KM} &lt; N_{RMST}$</td>
</tr>
</tbody>
</table>

$N_{KM}$, $N_{HR}$, $N_{RMST}$: simple size required for a given statistical power with a RD$_{KM}$, HR, RMST-based analyses, respectively.
could be envisaged instead of the Kaplan–Meier method to estimate RD. However, we did not explore such methods because Kaplan–Meier curves are widely used in trials with time-to-event primary outcome avoiding parametric assumptions. Of note, in the presence of competing risks, the Kaplan–Meier is not recommended and cumulative incidence functions can be used in this setting [26]. Our results are limited to randomized trials and further investigation should be conducted in the observation cohort context.

Noninferiority trials present specific difficulties in the design, analysis, and interpretation [15]. In superiority
trials, the null hypothesis \( H_0 \) is that there is no difference between treatments and the alternative is that they differ \[15\]. Investigators want to demonstrate a difference (superiority) between a new experimental treatment and the placebo or standard treatment (rejecting \( H_0 \)). As explained previously, the power of noninferiority trial is computed with no difference in failure rates between treatment arms although alternative designs have been discussed \[16\]. Our simulation study indicates that, if failure rates are correctly hypothesized, RD\(_{KM} \) analysis requires lower sample sizes compared with other analyses with all levels of failure rates (Table 3). In noninferiority trials, a type I error is the erroneous acceptance of an inferior experimental treatment, that is, we choose the experimental treatment when the standard is actually better, whereas a type II error is the erroneous rejection of a truly noninferior treatment, we choose the standard therapy when the experimental is essentially just as good \[14\]. The y-axis in the figures corresponds to the type II error, in particular, when there is no difference between failure rates in the two arms. Overestimation of failure rates lead to an inflation of the type II error with an HR-based analysis, whereas underestimation lead to overpowered trials.

Previous studies have also provided empirical evidence regarding the comparison between effects measures of time-to-event outcomes in clinical trials. Seruga et al. \[27\] compared the difference in survival probabilities at specific time points with the difference in RMST, and provided that difference in RMST is less dependent on the shape of survival curves and might be advantageous in the absolute treatment-effect investigations. Deeks \[28\] already showed that in 551 systematic reviews of clinical trials with binary outcomes, the RR and odds ratio models are on average more consistent than RD model. Trinquart et al. \[29\] compared the treatment effects measured by the HR and RMST difference (and ratio) in oncology randomized trials and point out that treatment-effect measures on the basis of RMST analysis yielded more conservative estimates than HR. To our knowledge, our study is the first to compare empirically the probabilities of not demonstrating non-inferiority of HR, RD estimated by Kaplan–Meier curves (RD\(_{KM} \)), and RMST difference in the context of different failure rate and different time to failure distribution hypotheses.

A limitation of our work is that other distributions of failure could have been investigated, maybe leading to slightly different findings. However, both distributions and shape parameters used in our simulation studies explore common situations of constant, monotonic increasing, and monotonic decreasing hazards. The use of piecewise exponential distribution with monotonic increasing or decreasing hazard lead to similar results than those obtained with the Weibull distribution (data not shown).

5. Conclusions

According to the results of our simulation study, we showed that discordance between analyses based on the three effect measures are much more driven by mistaken hypothesized failure rates than by the shape of the distribution of time to failure. When an HR-based analysis has been designed in the protocol, an overestimation of failure rates would lead to increase the probability of not concluding noninferiority even when there is no difference between the two arms. Inversely, underestimation of assumed failure rates would tend to declare noninferiority even in the presence of a subsequent difference between treatment failure rates. When margins are originally defined for RD measures, analyses based on RD\(_{KM} \) could be preferable because results with that measure are less influenced by mistaken levels of failure rates leading to more stable findings.

Acknowledgments

Funding: X.A. was supported by a Ph.D fellowship from the French Ministry of Higher Education and Research.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2017.12.010.

References


\[7\] Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the