509

The Use of the Restricted Mean Survival Time as a Treatment Measure in HIV/AIDS Clinical Trials: Reanalysis of the ACTG A5257 Trial

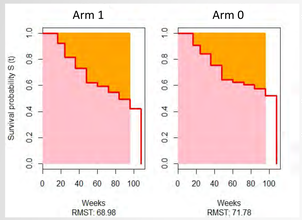
Xian Abulizi, MPH[1], Heather J. Ribaudo, PhD[2], Philippe Flandre, PhD[1]

[1]INSERM, Sorbonne Université, Institut Pierre Louis d'Epidémiologie et de Santé Publique (IPLESP), Paris, FRANCE; [2]Harvard TH Chan School of Public Health, Boston, MA, USA.

Heather Ribaudo, PhD
ACTG Statistics and Data Analysis Center
Center for Biostatistics in AIDS Research (CBAR)
Harvard TH Chan School of Public Health
Phone: 617/432.2897
Fax: 617/432.3163
ribaudo@sdac.harvard.edu

BACKGROUND

- The Restricted Mean Survival Time (RMST) has recently been proposed to estimate the event-free time over a given time period.
  - It is estimated as the area under the Kaplan-Meier (KM) curve as shown in pink below:



- *The average "event-free" time during 96 weeks of follow-up is 68.98 weeks on Arm 1 and is 74.11 weeks on Arm 0.*
- *The difference in RMST is 68.98-71.78=-2.8 weeks*
- *For future patients the "event-free " time on Treatment 1, is 2.8 weeks shorter than for those on Treatment 0 on average.*

- RMST measure has not been used as primary measure of efficacy in HIV/AIDS clinical trials.
- Under or over estimation of the hypothesized failure rates in the definition of non-inferiority bounds for a hazard ratio(HR) based analysis can significantly impact on the probability of a trial demonstrating non-inferiority for a HR and complicate the interpretation of the study findings. [1]
- ACTG A5257 was a US-based Phase III randomized trial comparing 3 modern NNRTI-sparing regimens for initial treatment of HIV-1. [2]
  - 1809 ART naive participants were randomized to FTC+TDF with ATV/RTV , RAL, DRV/RTV.
  - Primary efficacy endpoint: Time to virologic failure. Analysis based on the pairwise comparisons of difference in Kaplan-Meier estimates (RD_{KM}) by week 96 with equivalence accepted if the 97.5% CI was wholly contained within -/+10%.
  - Primary tolerability endpoint: Time to discontinuation of treatment for toxicity. Analysis based on RD_{KM} estimated by method Gray in the presence of the competing risk. A composite endpoint combining virologic and tolerability endpoint was also analyzed.
  - Based on an exponential distribution model, assuming rates of virologic failure and lost to follow-up of 25% and 12% respectively, a sample size of 600/arm was targeted.
  - Based on the RD_{KM} analysis, equivalence was demonstrated in all comparisons for the virologic endpoint, and for one comparison for the tolerability endpoint (DRV/RTV versus RAL).

1. *Abulizi X, Flandre P. Choice of treatment-effect measures when noninferiority margins originally defined in absolute difference translated into relative difference influenced the results of clinical trials. J Clin Epidemiol 2018;96:63–72. doi:10.1016/j.jclinepi.2017.12.010.*
2. *Lennox JL, Landovitz RJ, Ribaudo HJ, Ofotokun I, Na LH, Godfrey C, et al. Efficacy and Tolerability of 3 Nonnucleoside Reverse Transcriptase Inhibitor–Sparing Antiretroviral Regimens for Treatment-Naive Volunteers Infected With HIV-1: A Randomized, Controlled Equivalence Trial. Annals of Internal Medicine 2014;161:461. doi:10.7326/M14-1084.*

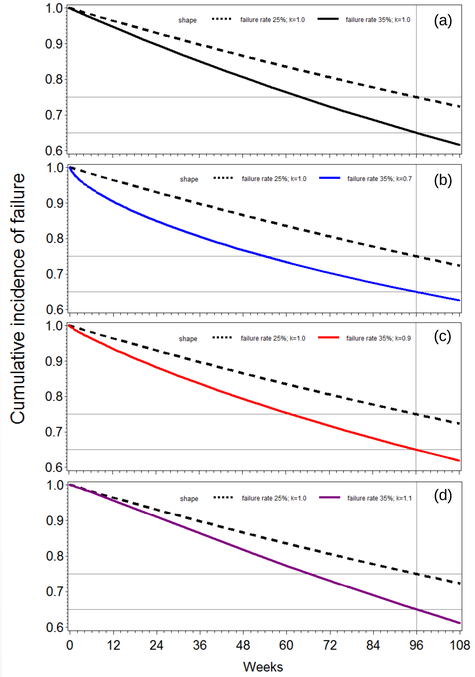
OBJECTIVES

- **To compare analysis based on the difference in RMST (Δ-RMST) measure to RD_{KM} and hazard ratio (HR) in the ACTG A5257 equivalence trial.**
- **To investigate the performance and characteristics of Δ-RMST-based analysis in the context of proportional and non-proportional hazards.**

METHODS

- The primary efficacy and tolerability as well as the combined outcome measures from ACTG A5257 trial were reanalyzed using hazard ratio(HR) and difference in RMST (Δ-RMST) and compared to the original study findings based on RD_{KM}.
- A5257 equivalence bounds were transformed for each measure assuming exponential time-to-event distribution and A5257 design characteristics.
- The performance of Δ-RMST-based analyses with proportional and non-proportional hazards in terms of power of the study and false positive rate was investigated in simulation study.
- Assuming 25% of failure rate by week 96 in the reference arm and a sample size of 600/arm, clinical trial dataset with time-to-event outcomes were generated with Weibull shape parameters to simulate proportional and non-proportional hazards for hypothetical pairwise comparisons (Figure 1).
  - Parameters for the simulation study were determined using Weibull parametric models fit for each randomized group of A5257.

Figure 1. Simulation study time-to-event distributions. *In each case the time-to-event outcome for the reference group (dashed line) follows an exponential distribution (shape parameter k=1.0) with a failure rate at 96 weeks of 25%. In the other group the failure rate at 96 weeks is 35% and the time-to-event outcome follows an exponential (k=1.0) (a) or Weibull (k=0.7, 0.9 and 1.1) distribution (b-d, respectively).*



RESULTS

Table 1. Summary of the analyses in the ACTG 5257 study and alternative analyses.

		Comparaison ATV/r vs RAL		Comparaison DRV/r vs RAL		Comparaison ATV/r vs DRV/r	
Method	Equivalence Bounds	Point estimate	97.5% CI	Point estimate	97.5% CI	Point estimate	97.5% CI
Virologic failure endpoint							
RD_{KM}	-/+10%	3.4%	[-0.6 to 7.4]**	5.6%	[1.4 to 9.8]**	-2,2%	[-6.7 to 2.2]**
HR	0.56; 1.50	1.4	[0.82 to 1.58]	1.40	[1.0 to 1.93]	0.81	[0.6 to 1.10]**
Δ-RMST	-5.18; +5.43	2.3wk	[-0.08 to 4.7]**	3.7wk	[1.1 to 6.2]	-1.4wk	[-4.1 to 1.4]**
Tolerability failure endpoint							
RD_{KM}	-/+7%	12.7%	[9.4 to 16.1]	3.6%	[1.4 to 5.8]**	9.2%	[5.5 to 12.9]
HR	0.29; 1.77	12.7	[5.6 to 29.1]	4.07	[1.7 to 9.9]	3.14	[2.0 to 4.9]
Δ-RMST	-3.44; +3.53	7.9wk	[5.5 to 10.3]	1.6wk	[0.1 to 3.0]**	6.4wk	[3.8 to 9.0]
Combined endpoint (Virologic or tolerability failure endpoint)							
RD_{KM}	-/+10%	14.9%	[10.2 to 19.6]	7.5%	[3.2 to 11.8]	7.5%	[2.3 to 12.7]
HR	0.63; 1.43	2.4	[1.7 to 3.0]	1.55	[1.1 to 2.1]	1.46	[1.1 to 1.9]
Δ-RMST	-5.30; +5.57	9.5wk	[6.3 to 12.6]	4.1wk	[1.4 to 6.8]	5.4wk	[1.9 to 8.5]

- Analyses of Δ-RMST globally led to similar conclusions as the published A5257 findings based on RD_{KM}.
- Analyses based on HR provided some discordant equivalence conclusions compared both with the initial analyses bases on RD_{KM} and the Δ-RMST

***Equivalence shown.*

Table 2. RMST estimate for each endpoint and randomized groups of ACTG 5257 study, with Δ-RMST estimate for pairwise comparisons.

Endpoint	Arm	RMST by week 96 (weeks)	Pairwise comparison	Δ-RMST Estimate (weeks)
Virologic failure	ATV/r	89.2	ATV/r vs. RAL	2.3 [-0.08 to 4.7]
	RAL	91.5	DRV vs. RAL	3.7 [1.1 to 6.2]
	DRV/r	87.5	ATV/r vs. DRV/r	-1.4 [-4.1 to 1.4]
Tolerability failure	ATV/r	87.1	ATV/r vs. RAL	7.9 [5.5 to 10.3]
	RAL	95.0	DRV vs. RAL	1.6 [0.1 to 3.0]
	DRV/r	93.4	ATV/r vs. DRV/r	6.4 [3.8 to 9.0]
Combined endpoint	ATV/r	81.5	ATV/r vs. RAL	9.5 [6.3 to 12.6]
	RAL	90.9	DRV vs. RAL	4.1 [1.4 to 6.8]
	DRV/r	86.9	ATV/r vs. DRV/r	5.4 [1.9 to 8.8]

Table 3. Probability of concluding equivalence using the Δ-RMST* when time-to-event outcomes follow exponential (shape parameter =1) or Weibull (shape parameter ≠ 1) distributions and treatment groups have an underlying 96-week failure rate of 25% (i.e., power).

		35% failure rate shape parameter (*k*)			
	Shape parameter (*k*)	0.7	0.9	1.0	1.1
25% failure rate	0.7	81.5	66.5	51.2	38.3
	0.9		89.4	86.8	77.4
	1.0			**93.6**	89.9
	1.1				95.2

- With an exponential distribution, the expected power for the pairwise comparison with the Δ-RMST analysis is 93.6%.
- When the PH assumption is valid (same shape parameter in both groups), power decreases (increases) with monotonic decreasing (increasing) hazard (i.e., Weibull shape parameter).
- When the PH assumption is not valid (i.e., a shape parameter that is greater in one group), the power of the study is markedly decreased.

Table 4. Probability of concluding equivalence using Δ-RMST* when time-to-event outcomes follow exponential (shape parameter =1) or a Weibull (shape parameter ≠ 1) distributions and treatment groups have underlying 96-week failure rates of 35% and 25% failure rates (i.e., false positive rate).

		35% failure rate shape parameter (*k*)			
	Shape parameter (*k*)	0.7	0.9	1.0	1.1
25% failure rate	0.7	1.6	19.6	38.9	61.3
	0.9	0,1	3.9	12.2	27.7
	1.0	0.1	1.6	**5.4**	15.2
	1.1	0.0	0.7	2.5	9.2

- With an exponential distribution, the false positive rate for the pairwise comparison with the Δ-RMST analysis is 5.4%.
- When the PH assumption is valid, the false positive rate decreases (increases) with monotonic decreasing (increasing) hazard.
- When the PH assumption is not valid, with a shape parameter that is greater (lower) in the group with the highest failure rate, the false positive rate is markedly increased (decreased).

** RD_{KM} and HR-based analyses were not evaluated as part of the simulation study since the RD_{KM} estimate will not change as the difference in the rate of failure is the same whatever the time-to-event distribution (see Figure 1), and the HR estimate is not appropriate in the case of non-proportional hazards.*

CONCLUSIONS

- Analyses based on Δ-RMST globally led to similar conclusions as the published findings of ACTG A5257 study based on RD_{KM}. In contrast, analyses based on HR provided some discordant equivalence conclusions compared both with initial analysis based on RD_{KM} and the Δ-RMST. Such discordance is mainly explained by violation of the PH assumption and by an under- or over-estimation of the hypothesized failure rates in the study design.
- Though there are some advantages to use the RMST measure, further discussion is needed with clinicians involved in the HIV/AIDS field to determine which estimand best informs the clinical question and to suggest equivalence bounds. We also show that finding provided by RMST-based analyses are sensitive to departures from the PH assumption.

Acknowledgements

X. Abulizi is supported by a Ph.D fellowship from the French Ministry of Education and Research.
H.J. Ribaudo is supported by National Institute of Allergy and Infectious Diseases (NIAID) grant UM1AI68634 for the ACTG Statistics and Data Management Center.
We gratefully acknowledge the efforts of the ACTG 5257 protocol team, the participating Clinical Trial Units of the ACTG, and the study participants.
A5257 was supported by the NIAID under Awards UM1 AI068634, UM1 AI068636 and UM1 AI106701. Bristol-Myers Squibb, Merck Inc., Janssen Therapeutics, and Gilead Sciences provided study medications (atazanavir, raltegravir, darunavir, and tenofovir-emtricitabine, respectively).