Adv Ther (2015) 32:854–862
DOI 10.1007/s12325-015-0238-9

ORIGINAL RESEARCH

Converging or Crossing Curves: Untie the Gordian Knot or Cut it? Appropriate Statistics for Non-Proportional Hazards in Decitabine DACO-016 Study (AML)

Jörg Tomeczkowski · Ansgar Lange · Andreas Güntert · Pushpike Thilakarathne · Joris Diels · Liang Xiu · Peter De Porre · Christoph Tapprich

To view enhanced content go to www.advancesintherapy.com
Received: July 20, 2015 / Published online: September 14, 2015

ABSTRACT

Introduction: Among patients with acute myeloid leukemia (AML), the DACO-016 randomized study showed reduction in mortality for decitabine [Dacogen® (DAC), Eisai Inc., Woodcliff Lake, NJ, USA] compared with treatment choice (TC): at primary analysis the hazard ratio (HR) was 0.85 (95% confidence interval 0.69–1.04; stratified log-rank $P = 0.108$). With two interim analyses, two-sided alpha was adjusted to 0.0462. With 1-year additional follow-up the HR reached 0.82 (nominal $P = 0.0373$). These data resulted in approval of DAC in the European Union, though not in the United States. Though pre-specified, the log-rank test could be considered not optimal to assess the observed survival difference because of the non-proportional hazard nature of the survival curves.

Methods: We applied the Wilcoxon test as a sensitivity analysis. Patients were randomized to DAC ($N = 242$) or TC ($N = 243$). One-hundred and eight (44.4%) patients in the TC arm and 91 (37.6%) patients in the DAC arm selectively crossed over to subsequent disease modifying therapies at progression, which might impact the survival beyond the median with resultant converging curves (and disproportional hazards).

Results: The stratified Wilcoxon test showed a significant improvement in median (CI 95%) overall survival with DAC [7.7 (6.2; 9.2) months] versus TC [5.0 (4.3; 6.3) months; $P = 0.0458$].

Conclusion: Wilcoxon test indicated significant increase in survival for DAC versus TC compared to log-rank test.

Funding: Janssen-Cilag GmbH.

Electronic supplementary material The online version of this article (doi:10.1007/s12325-015-0238-9) contains supplementary material, which is available to authorized users.

J. Tomeczkowski (✉) · A. Güntert · C. Tapprich
Janssen-Cilag GmbH, Neuss, Germany
e-mail: jtomeczk@its.jnj.com

A. Lange
Center for Health Economics Research Hannover,
Leibniz University Hannover, Hannover, Germany

P. Thilakarathne · J. Diels · P. De Porre
Janssen Pharmaceutica NV, Beerse, Belgium

L. Xiu
Janssen Research & Development, LLC, Raritan, NJ,
USA

Keywords: Acute myeloid leukemia; Dacogen®; Decitabine; Non-proportional hazards; Log-rank test; Oncology; Wilcoxon test

INTRODUCTION

In clinical trials of end-stage or potentially fatal diseases, survival is often the primary outcome measure of efficacy. To get a more granular understanding of underlying mechanisms, survival as one component of overall treatment efficacy is often analyzed by comparing the survival distributions of two or more treatment groups. There are several potential non-parametric and parametric tests available to compare two survival distributions. Among non-parametric tests based on the ranks of censored survival times, the two classic procedures are the log-rank test and the generalized Wilcoxon procedure. Each of these tests has been shown to be powerful in its ability to detect certain differences between survival distributions [1].

Among patients with acute myeloid leukemia (AML), the DACO-016 randomized study (ClinicalTrials.gov identifier, NCT00260832) showed a reduction in mortality for decitabine [Dacogen® (DAC), Eisai Inc., Woodcliff Lake, NJ, USA] compared with treatment choice (TC) which could be Cytarabine or Supportive Care [2]. With the primary analysis only showing a trend, these data resulted in approval of DAC in the European Union (EU), though not in the United States (US) and, furthermore, the French Haute Autorité de Santé negated a mortality benefit. Though pre-specified, the log-rank test could be considered not optimal to assess the observed survival difference because of the non-proportional hazard nature of the survival curves. For testing the observed treatment effect, Wilcoxon test is considered more powerful compared to the log-rank test, as the former assigns more weight to earlier events [3]. The objective of this article was to show the differences between the log-rank and Wilcoxon tests for the comparison of survival distributions in randomized controlled trials (RCTs). We use the recent illustrative example of decitabine in the phase III DACO-016 RCT in patients with AML to underline the relevance of this topic and to discuss the consequences for designing and conducting RCTs.

METHODS

The Wilcoxon test was applied as a sensitivity analysis by the same programmer (co-author L. X.) using the same software (SAS, SAS Institute Inc, NC, USA) with the same adjustments. In the DACO-016 study, patients (aged 65 years or older, ineligible for chemotherapy) were randomized to DAC ($N = 242$) or TC ($N = 243$). One-hundred and eight (44.4%) patients in the TC arm and 91 (37.6%) patients in the DAC arm selectively crossed over to subsequent disease modifying therapies at progression, which might impact the survival beyond the median with resultant converging curves (and disproportional hazards). Full methodology has been described previously by Kantarjian et al. [2].

This article is based on previously conducted studies and does not involve any new studies of human or animal subjects performed by any of the authors. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1964, as revised in 2013. Informed consent was obtained from all patients for being included in the study.

RESULTS

The stratified Wilcoxon test showed a significant improvement in overall survival (OS) with DAC [7.7 (6.2; 9.2) months] versus TC [5.0 (4.3; 6.3) months]; $P = 0.0458$ while at primary analysis the stratified log-rank test showed a non-significant improvement in OS [hazard ratio (HR): 0.85; 95% confidence interval (CI): 0.69–1.04; $P = 0.108$]. With 2 interim analyses, two-sided alpha was adjusted to 0.0462. With one-year additional follow-up the HR reached 0.82 (log-rank test $P = 0.0373$; Wilcoxon test $P = 0.0302$; Table 1).

Wilcoxon test indicated significant increase in survival already at primary analysis (CCO 2009) for DAC versus TC in patients with AML compared to log-rank test. The overall median survival for patients randomized to DAC ($N = 242$) or TC ($N = 243$) was 7.7 months or 5.0 months, respectively. The overall response rate [complete remission (CR) + complete remission with incomplete platelet recovery (CRp)] was 17.8% of patients in the DAC arm and 7.8% in the TC arm, which is statistically significant ($P = 0.001$; Table 2).

As expected was the median survival for patients who were in CR and CRp better than for the overall population: 18.6 months for DAC and 21.2 months for TC ($P = 0.13$; data from CCO 2010).

DISCUSSION

Almost all older patients with AML have a poor prognosis and the majority of patients are treated with therapies that are more than 40 years old and became standard for AML in the 1970s. In recent years a number of new therapies, like tipifarnib (Zarnestra®, Tibotec Therapeutics, a Division of Ortho Biotech, L.P), clofarabine (Clolar®; Genzyme, Cambridge, MA, USA), and laromustine (Onrigin®; Vion Pharmaceuticals, New Haven, CT, USA) were tested for AML in the older population, but failed to convince Food and Drug Administration's (FDA) Oncologic Drugs Advisory Committee (ODAC) or garner approval by the FDA or Committee for Medicinal Products for Human Use (CHMP). In 2012, the FDA decided to not grant approval for DAC, a DNA methyltransferase inhibitor,

Table 1 Log-rank test versus Wilcoxon test for overall survival at CCO 2009 and 2010

Decitabine		**TC**		**Decitabine vs. TC**	
N	**Death, *n* (%)** **Median OS, months** **[min; max]**	***N***	**Death *n* (%)** **Median OS, months** **[min; max]**	**HR [95% CI]**	***P* value**
CCO 2009 (protocol pre-specified)					
242	197 (81.4%)	243	199 (81.9%)	0.85 [0.69–1.04]	Log-rank test: 0.1079
	7.7 [6.2; 9.2]		5.0 [4.3; 6.3]		Wilcoxon test: 0.0458
CCO 2010 (analysis with one-year additional follow-up)					
242	219 (90.5%)	243	227 (93.4%)	0.82 [0.68–0.99]	Log-rank test: 0.0373
	7.7 [6.2; 9.2]		5.0 [4.3; 6.3]		Wilcoxon test: 0.0302

CCO Clinical cutoff, *CI* Confidence interval, *HR* Hazard ratio, *OS* Overall survival, *TC* Treatment choice

Table 2 Median overall survival at clinical cutoff 2010 for patients with and without CR + CRp

	Decitabine			**TC**		
	CR + CRp		**Total**	**CR + CRp**		**Total**
	No	**Yes**		**No**	**Yes**	
N (%)	199 (82.2)	43 (17.8)	242	224 (92.2)	19 (7.8)	243
Median survival, months	5.6	18.6[a]	7.7	4.4	21.2[a]	5.0

CR Complete remission, *CRp* Complete remission with incomplete platelet recovery, *TC* Treatment choice

[a] Difference is not statistically significant

which is already approved by the FDA and in over 40 countries globally for use in myelodysplastic syndromes.

Sekeres et al. [4] discussed the shortcomings in the development of new drugs in AML and cited decitabine as another drug emblematic of the "Boulevard of Broken Dreams." In their opinion, the negative decision is rooted in rigorous statistical principles, as the primary end point, survival difference, was not statistically significant ($P = 0.108$) at the primary analysis. An unplanned one-year additional follow-up indicated the same median OS while the HR reached 0.82 (nominal $P = 0.037$). These results, together with significant outcomes in secondary endpoints and a positive benefit–risk assessment, were the basis of EMA approval of DAC in the EU and other countries outside of the US.

As stated by Kantarjian et al. [5], there is little debate about the importance of rigorous statistical principles. However, there are more examples from clinical trials [6, 7] where the pre-specified test, particularly for comparing survival curves, was not the optimal choice once the results were on the table. Since survival curves cannot be predicted and, therefore, the most appropriate test not always pre-specified this need to be discussed and considered in a better way in the future.

Comparing Time-to-Event Distributions

Time-to-event data concern elapsed time until the occurrence of some pre-defined specific event. In simple cases, the event is death (survival), but also other events like progression of a disease, recurrence of disease (progression-free survival), or a complication (e.g., time to first incidence of neuropathy) might be considered. Study participants in principle are at risk for the occurrence of the respective event continuously over time from the beginning of the observation [8]. However, in RCTs the duration of observation may also vary from one subject to another because the subject might withdraw from the study (e.g., due to an adverse drug reaction), the study ends before the subject experiences that event or due to loss of follow-up. This type of censoring is also called "right censored" data since times of failure to the right are missing.

The most established way to summarize and compare survival data is calculating survivor functions (or survival curves). Therefore, the Kaplan–Meier estimate is the simplest way of computing the survival over time despite censoring. The Kaplan–Meier survival curve is defined as the probability of surviving in a given length of time while considering time in many small intervals (e.g., day) [9].

Tests of Equality of Survivor Functions

General Procedure

The two survival curves can be compared statistically by testing the null hypothesis, that is, there is no difference regarding survival between two interventions. There are several non-parametric tests to compare two survival distributions which are, with the exception of the Cox test, members of a family of statistical tests that are extensions to right censored data of non-parametric rank tests for comparing distributions [10]. Basically, all of these tests follow the same procedure: at each distinct failure time in the survival data, the contribution to the test statistic is obtained as a weighted standardized sum of the difference between the observed and expected number of deaths in each of the two groups. The expected number of deaths is obtained under the hypothesis of no differences between the survivals of the two groups.

Mantel proposed the use of the procedure for combining a series of 2 × 2 tables [10]. In this procedure, each time, t_j, a death occurs in either group, a 2 × 2 table is formed. The entry a_j represents the observed number of deaths at time t_j in the intervention group, and c_j represents the observed number of deaths at time t_j in the control group (Table 3). Of the n_j participants at risk just prior to time t_j, $a_j + b_j$ were in the intervention group and $c_j + d_j$ were in the control group. The expected number of deaths in the intervention group, denoted as $E(a_j)$, can be calculated as shown in Fig. 1. The weighting factor w_j, which is used for the calculation, determines the test statistic. The test statistics $W^2/V(W)$ has approximately a Chi-square distribution with one degree of freedom. If $w_i = 1$, we obtain the Mantel–Haenszel or log-rank test. If $w_i = n_j/(N + 1)$, where $N =$ the combined sample size, we obtain the Gehan version of the Wilcoxon test.

Table 3 Procedure to calculate survival time according to Mantel

	Death at time t_j	**Survivors at time t_j**	**At risk prior to time t_j**
Intervention	a_j	b_j	$a_j + b_j$
Control	c_j	d_j	$c_j + d_j$
	$a_j + c_j$	$b_j + d_j$	n_j

Thus, in simpler words, in survival analysis it is possible to obtain different results using different weighting factors depending on where the survival curves separate [11]. If the distribution of the survival curve of the study population is known, a test with an optimal weight function in the weighted log-rank family might be selected before study initiation [11, 12]. However, in practice the shape is unknown and the selection of the weights is problematic as an inappropriate choice may result in a loss of power [13]. However, in real-life practice a choice is often made between versions of the log-rank and the Gehan–Wilcoxon tests [12].

Log-Rank Test

The log-rank test, proposed by Mantel, is the standard test used in many trials [10]. It has been shown that the log-rank test is the best choice for testing differences, if the so-called proportional hazards assumption holds. This means that the risk for an event (e.g., death) in the intervention group is a constant multiple of the hazard in the control group. The assumption definitely does not hold in case survival curves cross.

Wilcoxon Test

On the other hand, Lee et al. [14] have shown that the Wilcoxon procedure has more power

$$E(a_j) = \frac{(a_j + c_j)(a_j + b_j)}{n_j}$$

with

$$W = \sum_{j=1}^{K} w_j[a_j - E(a_j)]$$

and

$$V(W) = \sum_{j=1}^{K} w_j^2 \frac{(a_j + c_j)(b_j + d_j)(a_j + b_j)(c_j + d_j)}{n_j^2(n_j - 1)}$$

Fig. 1 Procedure to calculate the expected number of deaths in the intervention group. The expected number of deaths in the intervention group is denoted $E(a_j)$. n_j participants at risk just prior to time t_j, $a_j + b_j$ were in the intervention group and $c_j + d_j$ were in the control group. The weighting factor w_j, which is used for the calculation, determines the test statistic. The test statistics $W^2/V(W)$ has approximately a Chi-square distribution with one degree of freedom. If $w_i = 1$, we obtain the Mantel–Haenszel or log-rank test. If $w_i = n_j/(N + 1)$, where $N =$ the combined sample size, we obtain the Gehan version of the Wilcoxon test

than the log-rank test when the HR is non-constant (proportional hazard assumption must be refused). As indicated above, the Wilcoxon procedure differs from the log-rank test only in that the deviations of observed from expected for both groups are weighted by the number of subjects at risk of failure at each distinct failure time.

Thus, the Wilcoxon test is more sensitive to differences between groups that occur at earlier time points in the conduct of a study (more weights to early events) whereas the log-rank test gives equal weights to all failures regardless to when they occur. As a result, the Wilcoxon test is susceptible to differences in the censoring patterns of the groups.

The HR estimate is routinely used to empirically quantify the between-group difference under the assumption that the ratio of the two hazard functions is approximately constant over time. When the underlying proportional hazards assumption is violated (i.e., the HR is not constant over time) the clinical meaning of such a ratio estimate is difficult, if not impossible, to interpret. Selective cross-over to subsequent therapies, which is routinely the case in oncology trials, contributes to non-proportional hazards. In this situation a Wilcoxon test can help to interpret the results because it gives more weight to earlier events when no or less subsequent therapy was given.

The Cox proportional hazards regression model depends on parametric assumptions. When there is a substantial difference in treatment effect estimates between the covariate-adjusted and unadjusted analyses, concerns about the proportional hazards assumption can arise [15].

Example of Decitabine

The aforementioned differences between the log-rank and Wilcoxon tests can be quite crucial with regards to the interpretation of results from clinical trials. In light of past decision by the FDA, the RCT for decitabine (DACO-016) in the treatment of AML is used as an illustrative example.

The study showed a non-significant increase in median OS with decitabine (7.7 months; 95% CI 6.2 to 9.2) compared with TC (5.0 months; 95% CI 4.3 to 6.3): at primary analysis the HR for OS was 0.85 (95% CI 0.69–1.04; stratified log-rank $P = 0.108$). With two interim analyses, two-sided alpha was adjusted to 0.0462. An unplanned one-year additional follow-up indicated the same median OS: the HR reached 0.82 (nominal $P = 0.037$).

These data, together with significant outcomes in secondary endpoints and a positive benefit–risk resulted in approval of decitabine in the EU, but not in the US. Though pre-specified, the log-rank test could be considered sub-optimal to assess the observed survival difference between treatment arms because of the non-proportional hazard nature of the survival curves. Patients (aged ≥65 years, ineligible for chemotherapy) were randomized to decitabine ($N = 242$) or TC ($N = 243$) but 91 (37.6%) patients in the decitabine arm and 108 (44.4%) patients in the TC arm selectively crossed over to subsequent disease modifying therapies at progression. This might have impacted the survival beyond the median with resulting in converging curves (and disproportional hazards). Therefore, the Wilcoxon test was applied as a sensitivity analysis. Naïve censoring of patients with subsequent therapies led also like Wilcoxon to statistically significant results when using log-rank (data not shown).

The Wilcoxon test stratified by baseline age, cytogenetic-risk and ECOG performance status (an accepted classification for disease progression) showed a significant improvement in OS with decitabine for the primary analysis [$P = 0.0458$ (and also for the secondary analysis with one-year follow-up ($P = 0.0302$)].

In this example, the Wilcoxon test indicated a significant increase in survival for decitabine versus TC in patients with AML compared to results generated by applying the log-rank test in the primary analysis. It looks like Wilcoxon test is better than log-rank to detect the early difference between the two Kaplan–Meier curves in this example. These results were favorably considered by the German Federal Joint Committee [Gemeinsamer Bundesausschuss (G-BA)] in the benefit assessment.

CONCLUSIONS

In general, the log-rank test and the Wilcoxon test are actually two members of a larger class of linear rank tests. These tests will have different power under different circumstances. If hazards are proportional between groups then the log-rank test has greatest power. This fact is approximately true for many trials and accounts for its popularity. However, this property sometimes does not hold; when it does not, it is necessary to consider other tests instead, like the Wilcoxon procedure.

Although this might be very unlikely, trials could perhaps have been stopped earlier had the appropriate test statistics been the pre-specified statistic of choice [13]. In the absence of clear indications for using the log-rank or Wilcoxon test for the comparison of survival curves legislators might be challenged to specify unambiguous rules if necessary. There is a need in the clinical community to clarify and educate which tests are appropriate when survival curves are non-proportional. It should be possible to ex ante include different test options in a statistical analysis plan making their respective use dependent on the proportionality of hazard rates.

In the future, an adaptively weighted log-rank test might be appropriate because it maintains optimality at the proportional alternatives, while improving the power over a wide range of non-proportional alternatives [14]. Uno et al. [15] suggest that when there is not sufficient information about the profile of the between-group difference at the design stage of the study, practitioners should consider a pre-specified, clinically meaningful, model-free measure for quantifying the difference and to use robust estimation procedures to draw primary inferences.

ACKNOWLEDGMENTS

Sponsorship, article processing charges, and the open access charge for this study were funded by Janssen-Cilag GmbH, Germany. All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this manuscript, take responsibility for the integrity of the work as a whole, from inception to publication and have given final approval for the version to be published. All authors had full access to all of the data in this study and take complete responsibility for the integrity of the data and accuracy of the data analysis. The authors would like to acknowledge the investigators of the DACO-016 trial Hagop M. Kantarjian, Xavier G. Thomas, Anna Dmoszynska, Agnieszka Wierzbowska, Grzegorz Mazur, Jiri Mayer, Jyh-Pyng Gau, Wen-Chien Chou, Rena Buckstein, Jaroslav Cermak, Ching-Yuan Kuo, Albert Oriol, Farhad Ravandi, Stefan Faderl, Jacques Delaunay, Daniel Lysák, Mark Minden, and Christopher Arthur.

Conflict of interest. Jörg Tomeczkowski is an employee of Janssen-Cilag GmbH. Andreas Güntert is an employee of Janssen-Cilag GmbH. Pushpike Thilakarathne is an employee of Janssen Pharmaceutica. Joris Diels is an employee of Janssen Pharmaceutica. Liang Xiu is an employee of Janssen Research & Development LLC. Peter De Porre is an employee of Janssen Pharmaceutica. Christoph Tapprich is an employee of Janssen-Cilag GmbH. Janssen-Cilag GmbH, Janssen Pharmaceutica, and Janssen Research & Development LLC are companies belonging to Johnson & Johnson. Ansgar Lange received an honorarium from Janssen-Cilag GmbH Germany.

Compliance with ethics guidelines. This article is based on previously conducted studies and does not involve any new studies of human or animal subjects performed by any of the authors. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1964, as revised in 2013. Informed consent was obtained from all patients for being included in the study.

REFERENCES

1. Fleming T, O'Fallon J, O'Brien P, Harrington D. Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. Biometrics. 1980;36(4):607–25.

2. Kantarjian HM, Thomas XG, Dmoszynska A, et al. Multicenter, randomized, open-label, phase III trial of decitabine versus patient choice, with physician advice, of either supportive care or low-dose cytarabine for the treatment of older patients with newly diagnosed acute myeloid leukemia. J Clin Oncol. 2012;30(21):2670–7.

3. Cleves M, Gould WW, Gutierrez RG, Marchenko Y. An introduction to survival analysis using stata. 2nd ed. ISBN-10 59718-041-6 Stata Press; 2008.

4. Sekeres M, Steensma D. Boulevard of broken dreams: drug approval for older adults with acute myeloid leukemia. J Clin Oncol. 2012;30(33):4061–3.

5. Kantarjian H, Wilson W, Ravandi F, Estey E. Decitabine in older adults with acute myeloid leukemia: why was the dream broken? J Clin Oncol. 2013;31(14):1795–6.

6. Mok T, Wu Y, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. N Engl J Med. 2009;361(10):947–57.

7. van Besien K, Loberiza F, Bajorunaite R, et al. Comparison of autologous and allogeneic hematopoietic stem cell transplantation for follicular lymphoma. Blood. 2003;102(10):3521–9.

8. Agarwal G. Statistics for surgeons—understanding survival analysis. Indian J Surg Oncol. 2012;3(3):208–14.

9. Altman D. Analysis of survival times. Practical statistics for medical research. London: Chapman & Hall/CRC Texts in Statistical Science; 1990. p. 365–93.

10. Tarone R, Ware J. On distribution-free tests for equality of survival distributions. Biometrika. 1977;64(1):156–60.

11. Friedman L, Furberg C, DeMets D. Fundamentals of clinical trials. 4th ed. New York: Springer; 2010.

12. Bathke A, Kim M, Zhou M. Combined multiple testing by censored empirical likelihood. J Stat Plan Inference. 2009;139(3):814–27.

13. Kosoroka M, Lin C. The versatility of function-indexed weighted log-rank statistics. J Am Stat Assoc. 1999;94(445):320–32.

14. Jiang H, Symanowski J, Paul S, Qu Y, Zagar A, Hong S. The type I error and power of non-parametric logrank and Wilcoxon tests with adjustment for covariates—a simulation study. Stat Med. 2008;27(28):5850–60.

15. 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(22):2380–5.