Clinical Cancer Research

Deviation from the Proportional Hazards Assumption in Randomized Phase 3 Clinical Trials in Oncology: Prevalence, Associated Factors, and Implications

Rifaquat Rahman[1,2], Geoffrey Fell[2,3], Steffen Ventz[2,3,4], Andrea Arfé[5], Alyssa M. Vanderbeek[6], Lorenzo Trippa[2,3,4], and Brian M. Alexander[1,2]

Abstract

Purpose: Deviations from proportional hazards (DPHs), which may be more prevalent in the era of precision medicine and immunotherapy, can lead to underpowered trials or misleading conclusions. We used a meta-analytic approach to estimate DPHs across cancer trials, investigate associated factors, and evaluate data-analysis approaches for future trials.

Experimental Design: We searched PubMed for phase III trials in breast, lung, prostate, and colorectal cancer published in a preselected list of journals between 2014 and 2016 and extracted individual patient-level data (IPLD) from Kaplan–Meier curves. We re-analyzed IPLD to identify DPHs. Potential efficiency gains, when DPHs were present, of alternative statistical methods relative to standard log-rank based analysis were expressed as sample-size requirements for a fixed power level.

Results: From 152 trials, we obtained IPLD on 129,401 patients. Among 304 Kaplan–Meier figures, 75 (24.7%) exhibited evidence of DPHs, including eight of 14 (57%) KM pairs from immunotherapy trials. Trial type [immunotherapy, odds ratio (OR), 4.29; 95% confidence interval (CI), 1.11–16.6], metastatic patient population (OR, 3.18; 95% CI, 1.26–8.05), and non-OS endpoints (OR, 3.23; 95% CI, 1.79–5.88) were associated with DPHs. In immunotherapy trials, alternative statistical approaches allowed for more efficient clinical trials with fewer patients (up to 74% reduction) relative to log-rank testing.

Conclusions: DPHs were found in a notable proportion of time-to-event outcomes in published clinical trials in oncology and was more common for immunotherapy trials and non-OS endpoints. Alternative statistical methods, without proportional hazards assumptions, should be considered in the design and analysis of clinical trials when the likelihood of DPHs is high.

Introduction

Results from randomized controlled trials (RCTs) play an essential role in therapeutic development and clinical decision making. Standard clinical trial designs, summary statistics, and analytic procedures based on proportional hazards (PHs) assumptions work well when the treatment effects are constant over time. Deviations from proportional hazards (DPHs) occur when treatments exhibit variation of treatment effects (hazard ratio) over time. Time-varying treatment effects are increasingly recognized in modern trials (1). Understanding the characteristics and degree of time-varying treatment effects in oncology clinical trials has implications for design, analysis, and interpretability of trials, and may ultimately lead to improvements in the testing of therapies. In this context, the use of statistical models that poorly represent the true underlying time-varying treatment effect can constitute a substantial obstacle to therapeutic development.

Time-varying treatment effects are one example of treatment effect heterogeneity that may be induced by several factors. Immuno-oncology (IO) trials have shown evidence of possible delayed treatment effects (1, 2). Subgroups of patients with different responses to a given therapy can also manifest as a treatment effect that is not constant over time (3). Time-varying treatment effects, which define DPHs, indicate that the hazard ratio (HR) varies over time after randomization. With DPHs, estimates of treatment effects from PHs models may not be interpretable (4).

The widely used log-rank test and the Cox proportional hazards model are not optimal in terms of power when DPHs are present and render sample size calculations, subsequent results, and their interpretation questionable (5, 6). A better understanding of

[1]Department of Radiation Oncology, Center for Neuro-Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts. [2]Program in Regulatory Science Research, Dana-Farber Cancer Institute, Boston, Massachusetts. [3]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, Massachusetts. [4]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts. [5]Department of Decision Sciences, Bocconi University, Milano, Italy. [6]Department of Biostatistics, Columbia University Mailman School of Public Health, New York, New York.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

B.M. Alexander and L. Trippa contributed equally to this article as co-senior authors.

Prior Presentation: The study findings were presented at the American Society of Clinical Oncology 2018 Annual Meeting in Chicago, Illinois.

Corresponding Author: Brian M. Alexander, Dana-Farber/Brigham and Women's Cancer Center, 450 Brookline Avenue, Boston, MA 02115. Phone: 617-732-6313; Fax: 617-975-0932; E-mail: bmalexander@lroc.harvard.edu

Clin Cancer Res 2019;25:6339–45

doi: 10.1158/1078-0432.CCR-18-3999

Translational Relevance

Deviation from proportional hazards (DPHs), which occurs when a treatment exhibits variation of treatment effect over time, may be more prevalent in modern trials and can lead to underpowered clinical trials or misleading conclusions. We used reconstructed individual patient-level data from published clinical trials in recent phase III oncology clinical trials in breast, lung, colorectal and prostate cancer to investigate the prevalence of DPHs, its associated factors across different clinical settings and its implications. A notable proportion of published clinical trials in oncology exhibited evidence of DPHs, particularly in trials testing immunotherapy agents and analyses of nonsurvival endpoints. In re-analyzing immunotherapy trials, the use of alternative statistical approaches allowed for more efficient clinical trials requiring fewer patients relative to conventional trial design. Alternative statistical methods, without the proportional hazards assumptions, should be considered in the design and analysis of clinical trials in settings where DPHs occur more frequently.

the prevalence of DPHs, along with factors that increase the likelihood of DPHs, could substantially improve the therapeutic development process by anticipating DPHs and suggesting appropriate alternative trial designs.

We evaluate whether the PHs model—assumed in the vast majority of oncology trials (7)—is representative of modern RCT data. We use reconstructed individual patient-level data (IPLD; ref. 8) from published clinical trials to investigate DPHs in recent phase III oncology clinical trials. We illustrate the magnitude of DPHs and types of time-varying treatment effects observed across different clinical settings (IO, hormonal therapies, chemotherapies, etc.), and identify associated factors. This information can be used to evaluate if alternative statistical methodologies, instead of a standard PHs analysis, should be considered in a given context for the design of future studies. We then evaluate alternative procedures for testing treatment effects and demonstrate how data from completed clinical trials can be used to evaluate operating characteristics of these alternative methods for IO trials.

Materials and Methods

Study inclusion criteria, search strategy, and data extraction

We performed a PubMed search on December 4, 2017 (Supplementary Methods), with the key words of "breast cancer," "lung cancer," "prostate cancer," or "colorectal cancer" and limited results to phase III clinical trials published between January 1, 2014 and December 31, 2016, published in the English language in one of the following journals: *New England Journal of Medicine*, *Journal of the American Medical Association*, *Lancet*, *Lancet Oncology*, *Journal of Clinical Oncology*, *Journal of the National Cancer Institute*, *JAMA Oncology*, and *Annals of Oncology*. Included publications reported on a tumor directed intervention with at least one time-to-event outcome and a Kaplan–Meier (KM) curve. If multiple KM curves were reported in a single publication, overall survival (OS) and one non-OS time-to-event endpoint were selected for every two-arm comparison. For non-OS outcomes, priority was given to progression-free survival (PFS), disease free survival and relapse-free survival. For ease of analysis, multi-arm trials were treated as multiple two-arm comparisons.

For each publication, trial characteristics including cancer type, publication date, trial registration, type of intervention, type of experimental therapy, trial population (metastatic or nonmetastatic), trial design (superiority or noninferiority), randomization ratio, sample size, primary endpoint(s), reported HR and statistical significance level (*P* value) used for the trial primary hypothesis were extracted. References for every included publication are available in Supplementary Table S1.

IPLD reconstruction

For each KM curve (censored), survival times and the corresponding survival probabilities were extracted using *DigitizeIt* from publications. The number of patients at risk and the number of events were also extracted. The algorithm of Guyot and colleagues (8) was then used to estimate IPLD from the survival times and probabilities. Reconstructed datasets with discrepancy from publications in estimated HR above 0.15 were re-evaluated by comparison of published and reconstructed KM curves based on a previously described procedure (1).

Evaluating and characterizing deviation from PHs

With IPLD, we used a log-rank test to evaluate treatment efficacy and a Cox regression model to estimate HRs. As done in a prior meta-analysis (8), we used a Grambsch–Therneau test with a *P* value cutoff of 0.1 to test the PHs assumption. In a sensitivity analysis, we also considered a *P* value cutoff of 0.05 (Supplementary Table S2). The Grambsch–Therneau test uses residuals in a univariate Cox model with treatment included as predictor, and it evaluates potential trends of these residuals (9). We also used a Cox model, including a time–treatment interaction (10) to estimate time-varying treatment effects. In this model, the logarithm of the hazard ratio between experimental and control arms at time t is $\log(HR_t) = \beta_0 + \beta_1 * t$. Small values of HR_t (below one) indicate a large treatment effect at time t; if the time–treatment interaction coefficient β_1 is negative, then the benefit of the treatment increases over time. In a sensitivity analysis (Supplementary Material), we also used the model $\log(HR_t) = \beta_0 + \beta_1 * \log(t)$ with an interaction between $\log(t)$ and treatment.

Descriptive statistics were used to explore differences between studies with and without evidence of DPHs. Logistic regression was used to identify associations of trial characteristics with DPHs. To simplify analysis, pairs of KM curves from the same trial were treated as independent observations in our analysis. To account for the fact that the presence or absence of DPHs for multiple outcomes (for instance OS and PFS) or treatments (multi- experimental arm trials) within a single study are not independent, we also fitted a multivariable logistic regression model with random effects (Supplementary Table S3). We also report the results of an alternative model to describe the relationship between trial characteristics and evidence of DPHs (Supplementary Table S4).

Alternative analyses

We used IPLD to re-analyze trials using several alternative approaches that are easy to implement:

(i) Fleming–Harrington (FH) weighted the log-rank test (11). This test requires two parameters ρ-γ that are used to weigh the importance of events during the follow-up time. We used FH log-rank tests with higher weights on early ($\rho = 1$,

$\gamma = 0$), middle ($\rho = 1$, $\gamma = 1$) or late events ($\rho = 0$, $\gamma = 1$; ref. 11).

(ii) Restricted mean survival time (RMST) analysis, using the ratio of RMSTs (rRMST) to compare experimental and control arms (5, 6). The restriction time t^* in RMST was fixed at the 80th percentile of the (uncensored) event times, based upon evidence supporting a restriction time t^* close to the tail of survival curves (12).

(iii) Milestone analysis for the difference of the survival curves at a fixed time-point t^{**} (13). We used the $t^{**} = 80^{th}$ percentile of the (uncensored) event times. In practice, t^{**} should be chosen from considerations of clinical relevance (13).

(iv) Cox model with early (excluding first 20% of events) and late (excluding the last 20% of events) truncation. This approach was included for illustrative purposes, and as a direct evaluation of the influence of early overlapping survival curves followed by separation (as in IO trials) in diluting the evidence of treatment effects expressed by standard PHs analyses.

All analyses were performed with R using the survival, survRM2 and FHtest packages.

Results

We identified 836 results from our initial PubMed search and excluded 528 publications because they were not published in our prespecified list of journals. After excluding another 153 entries for not meeting our inclusion criteria, a total of 152 trials published in the period 2014 to 2016 were selected (Supplementary Fig. S1). From these trials, IPLD from 129,401 patients were incorporated into the analysis (median 577 patients per trial, range 71–8,381). Trial characteristics are summarized in Table 1.

IPLD reconstruction was derived from trial publication. After separating multi-arm studies into pairs with an experimental arm and a control, 263 published figures yielded a total of 304 reconstructed pairs of KM curves (141 OS and 163 non-OS comparisons). KM curves of the primary endpoint, or at least one of multiple co-primary endpoints, were reconstructed for 86% (131 of 152) of the trials, while the remaining trials did not have a time-to-event primary endpoint. HR estimated with our reconstructed IPLD, as expected, correlated strongly with published HRs (Supplementary Fig. S2) with 98.7% (300 of 304) digitized curves yielding an estimated HR within 0.15 of the published value. All KM pairs with discrepancy between published and IPLD-based HR estimates above 0.15 were manually re-digitized to check for potential errors. No issues were detected (Supplementary Table S5).

In our analysis, 75 (24.7%) pairs of KM curves exhibited evidence of DPHs with significant change in treatment effect (hazard ratio) over time. Each case is listed in Supplementary Table S6. Immunotherapy (57%, 8/14), hormonal/endocrine (32%, 7/22), targeted therapy (26%, 34/133), radiotherapy (29%, 8/28), and chemotherapy (16%, 15/92) trials had varying prevalence of evidence of DPHs. Characteristics associated with DPHs on analyses included endpoint [non-OS vs. OS endpoint, odds ratio (OR), 2.85; 95% CI, 1.61–4.00; $P < 0.001$] and trial population (metastatic vs. nonmetastatic, OR, 2.52; 95% CI, 1.22–5.22; $P = 0.008$). In our study, 15% OS curves and 33% of non-OS curves exhibited evidence of DPHs. Moreover, IO trials (reference: chemotherapy; OR, 6.58; 95% CI, 2.01–22.23), and trials in prostate cancer (reference: breast cancer; OR, 2.44; 95% CI, 1.10–5.42) were positively associated with DPHs (Table 2). We computed the power to detect covariate associations given that 75 pairs of KM curves out of 304 showed evidence of DPHs. At the 10% (5%) significance level, a univariable logit analyses would have an 80% (70%) power to detect a covariate association with an odds ratio of 2 for a factor with 25% prevalence.

On multivariable analysis, endpoint (adjusted OR, 3.23; 95% CI, 1.79–5.88; $P < 0.001$), trial population (adjusted OR, 3.18; 95% CI, 1.26–8.05; $P = 0.021$), IO trials (reference:

Table 1. Characteristics of the trials included in the analyses

Feature	No. (%)
Journal (alphabetical order)	
Lancet oncology	55 (37.4)
Journal of clinical oncology	35 (23.8)
Annals of oncology	31 (21.0)
New England journal of medicine	15 (10.2)
Lancet	11 (7.5)
Cancer type	
Breast	54 (35.5)
Lung	48 (31.6)
Colorectal	29 (19.1)
Prostate	20 (13.2)
Lung/gastrointestinal neuroendocrine	1 (0.7)
Trial population	
Metastatic or recurrent allowed	103 (67.8)
Nonmetastatic	49 (32.2)
Randomization	
Randomization of therapy agent/modality	135 (88.8)
Randomization of therapy dose, timing/sequencing, or duration of therapy	14 (9.2)
Mixed	3 (2.0)
Type of experimental therapy	
Targeted therapy	66 (43.4)
Chemotherapy	39 (25.7)
Radiation	16 (10.5)
Hormonal/endocrine	14 (9.2)
Surgery	2 (1.3)
Other	4 (2.6)
Combination of multiple therapy types	4 (2.6)
Arms	
Two	136 (89.5)
Three	7 (4.6)
Four	9 (5.9)
Randomization ratio (experimental arm: control arm)	
1:1[a]	130 (85.5)
2:1	22 (14.5)
Trial design	
Superiority	134 (88.2)
Noninferiority	16 (10.5)
Both[b]	2 (1.3)
Primary outcome	
Progression-free survival	47 (30.9)
Overall survival	38 (25.0)
Disease-free survival	22 (14.4)
Non-time-to-event outcome	16 (10.5)
Multiple primary outcomes	15 (9.9)
Other time-to-event outcome	14 (9.2)
Primary outcome result	
Positive	69 (45.4)
Negative	78 (51.3)
Mixed	5 (3.3)

[a]For multi-arm trials, this refers to ratio of patients randomized to each experimental arm versus the control arm of the study (e.g., a three-arm trial with 1:1:1 randomization is categorized as 1:1).

[b]Possible with two co-primary endpoints or a study designed to be a non-inferiority study that can become a superiority study.

Table 2. Univariable and multivariable logistic analysis for variables associated with deviation from proportional hazards

Variable	KM curve pairs with evidence of DPHs (%)	Sample odds ratio (95% CI)	Multivariable logistic regression	
			Adjusted odds ratio (95% CI)	*P* value (Likelihood ratio test)
Cancer type				
Breast	24/110 (22)	1	1	0.607
Lung	23/87 (26)	1.32 (0.77–2.54)	1.26 (0.57–2.80)	
Colorectal	12/68 (18)	0.77 (0.36–1.66)	0.90 (0.38–2.15)	
Prostate	15/37 (41)	2.44 (1.10–5.42)	1.86 (0.66–5.23)	
Trial population				
Nonmetastatic	20/102 (20)	1	1	0.010
Metastatic or recurrent patients allowed	55/202 (27)	1.53 (0.86–2.74)	3.18 (1.26–8.05)	
Type of experimental therapy				
Chemotherapy	15/92 (16)	1	1	0.221
Molecular/targeted therapy	34/133 (26)	1.70 (0.87–3.29)	1.24 (0.60–2.56)	
Immunotherapy	8/14 (57)	6.58 (2.0–21.58)	4.29 (1.11, 16.6)	
Hormonal/endocrine	7/22 (32)	2.30 (0.81–6.56)	1.40 (0.40–4.87)	
Radiotherapy	8/28 (29)	1.97 (0.74–5.26)	2.55 (0.73–8.91)	
Trial design				
Noninferiority	8/35 (23)	1	1	0.608
Superiority	65/263 (25)	1.12 (0.49–2.58)	1.28 (0.49–3.35)	
Outcome of KM				
Non-OS endpoint	54/463 (33)	1	1	<0.001
Overall survival (OS)	21/141 (15)	0.35 (0.20–0.62)	0.31 (0.17–0.56)	
Primary trial outcome result				
Negative trial	33/161 (20)	1	1	0.168
Positive trial	42/143 (29)	1.61 (0.95–2.73)	1.51 (0.84–2.72)	
Sample size of the study[a]			1.04 (1.01–1.06)	0.004

NOTE: Categorical variables comprising <5% of trial population were excluded for this analysis [cancer type: mixed; type of experimental therapy: surgery, other; randomization type: mixed; primary outcome: mixed result].

[a]Relative increase in odds of the probability of DPHs when the sample size of the study is increased by 100 patients.

chemotherapy; adjusted OR, 4.29; 95% CI, 1.11–16.6) and sample size (adjusted OR, 1.04; 95% CI, 1.01–1.06) were associated with DPHs. Results were confirmed by analysis based on a random-effects model that introduced dependence among KM pairs (classified as DPHs or non-DPHs) within the same trial (Supplementary Table S3). A complementary logit analysis, using a Grambsch–Therneau test *P* value cutoff of 0.05 to declare DPHs and define the dependent variable in the logit-model, provided similar results (Supplementary Table S2).

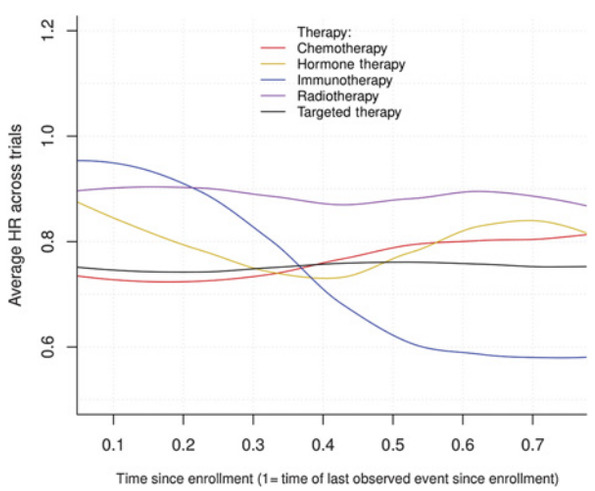

Figure 1.

Average hazard ratio as a function of time for positive trials. The hazard ratio over time for each pair of KM curves of positive trials (i.e., trials that met their primary endpoint) was averaged together for a composite visual representation, stratified by experimental treatment type. Noninferiority trials were excluded.

Among 75 cases of DPHs, 32 (43%) had a negative treatment-time interaction coefficient (treatment effect increases during time), including seven of eight (88%) KM pairs from IO trials with evidence of DPHs. The remaining 43 (57%) cases of DPHs had a positive arm–time interaction coefficient. The mean HR over time from all positive trials was plotted in Fig. 1 and Supplementary Fig. S4 shows individual HR(t) for immunotherapy trials.

For each trial, IPLD was re-analyzed with log-rank, weighted log-rank, RMST, event truncation, and milestone analysis. The null and alternative hypotheses for each test are summarized in Table 3. Using a two-sided *P* value threshold of 0.05 for all studies, after exclusion of noninferiority trials, discordance between log-rank and alternative tests ranged between 4% (early-truncated Cox) and 19% (FH late-weighted log-rank test) with alternative procedures. Among studies with DPHs, discordance between log-rank and alternative tests ranged between 3% (late-truncated Cox) and 45% (FH late-weighted log-rank test). The use of RMST analysis produced discordance in 17% of comparisons among all studies and 22% of comparisons among those with evidence of DPHs.

To assess potential trial efficiency gains with alternative analysis plans in the setting of DPHs, we re-analyzed data for a subset of four immunotherapy trials where DPHs were observed: CheckMate017 (14), CheckMate057 (15), KEYNOTE024 (16), and CA184–043 (17). These studies present evidence of increasing treatment effect over time (decreasing HRs, see Supplementary Fig. S4). Assuming enrollment rates and follow-up times as reported in each of these studies, we identified the overall sample size required for 80% power using a standard log-rank test (assuming

Table 3. Discordances in result with nonproportional hazard analyses

	All pairs of KM curves, $N = 267$				KM pairs with evidence of DPHs, $N = 67$			
Testing procedure	**Test result**	**Significant time-treatment interaction (TTA)**	**Positive TTA**	**Negative TTA**	**Test result**	**Significant time-treatment interaction (TTA)**	**Positive TTA**	**Negative TTA**
LRT: Test$^+$	107	40 (37%)	27	13	43	34 (78%)	22	12
Test$^-$	160	26 (43%)	10	16	24	20 (83%)	9	11
Alternative testing procedure	**Test$^+$**	**Total[a] number of discordances (LRT result ≠ Alt result)**	**LRT$^+$ Alt$^-$[b]**	**LRT$^-$ Alt$^+$[c]**	**Test$^+$**	**Total[a] number of discordances (LRT result ≠ Alt result)**	**LRT$^+$ Alt$^-$[b]**	**LRT$^-$ Alt$^+$[c]**
RMST (ratio)	105	46 (17%)	24	22	46	15 (22%)	6	9
Early-truncated Cox	112	12 (5%)	3	8	46	3 (4%)	0	3
Late-truncated Cox	112	13 (5%)	4	9	43	2 (3%)	1	1
Early-weighted LRT	115	26 (10%)	9	17	45	8 (12%)	3	5
Mid-weighted LRT	91	34 (13%)	25	9	33	16 (24%)	13	3
Late-weighted LRT	78	51 (19%)	40	11	27	30 (45%)	23	7
Milestone (80% of follow-up)	77	42 (16%)	36	6	50	22 (33%)	19	3

NOTES: Noninferiority trials are excluded from this table.
The LRT, early, late, and mid-weighted Fleming-Harrington (FH) tests, test the null hypothesis H_0: the survival functions in the control and experimental arm are identical against H_A: the experimental arms survival is superior to the control at some time during follow-up.
In the truncated Cox-model, we test H_0: the hazard ratio (HR) between the experimental and control arm equals one against H_A: HR<1 in the truncated subset.
The RMST tests the hypothesis H_0: the ratio of restricted mean survival time between the experimental and control arm is one against H_A: the ratio is smaller than one.
Abbreviations: Alt, alternative test; LRT, log-rank test.
[a]Number of trials where an alternative test gave a different result compared with the log-rank test when using $P < 0.05$ as a value of statistical significance.
[b]Number of trials where there was a significant treatment effect ($P < 0.05$) by LRT, but alternative analysis did not yield a significant treatment effect ($P > 0.05$).
[c]Number of trials where there was not a significant treatment effect ($P > 0.05$) by LRT, but alternative analysis yielded a significant treatment effect ($P > 0.05$).

the observed non-PHs treatment effect) and (one-sided) 2.5% type I error rate. We then considered alternative data analyses plans using either the FH weighted log-rank tests, RMST, the Cox model with early truncation of time-to-event (excluding first 20% of events) or a test of the difference in survival at $t^{**} = 80\%$ of times since randomization in both arms. For each of these alternative analyses, assuming the same enrollment rates and follow-up times as for the LRT, we determined the minimum sample size that to ensure 80% power under the observed primary outcome distribution (observed KM curves; ref. Table 4). Notably, the Fleming–Harrington-type late-weighted test reduced the overall sample size of the study by up to 74% compared with a log-rank test.

Discussion

Summary statistics such as HR are meaningful when treatment effects are constant. Here, we sought to investigate two important but separate aspects of clinical trials: testing procedures to detect potential treatment effects of experimental treatments and inference in variations of these treatment effects (for instance hazard ratios) over time. Although the majority of phase III RCTs in oncology use a testing procedure that is optimal under the PHs assumption (log-rank test) or use the Cox PHs model to analyze time-to-event endpoints, only a minority of publications (7%–9%) explicitly report testing of the PHs assumption (7). We used IPLD from published clinical trials to identify the presence of DPHs and associated factors in phase III oncology clinical trials. To describe DPHs in oncology clinical trials, we examined breast, lung, colorectal, and prostate trials published in prespecified oncology journals. This was necessary to evaluate if a single analytic approach can work well across different settings, or if the choice of a non-PHs analysis (e.g., RMST, weighted log-rank test, etc.) should depend on the context. Published trials can be used to identify areas of clinical research where DPHs are more likely to occur, such as in IO. For specific types of trials, the analysis of previously published trials can support the identification of robust non-PHs methods and guide context-specific evaluations of pivotal operating characteristics of non-PHs methods.

We found that a notable proportion of time-to-event outcomes reported in oncology clinical trials show evidence of DPHs (~25%), concordant with prior estimates. On the basis of the

Table 4. Overall sample size to ensure 80% power to declare a positive result using different statistical analyses methods

	Sample size (Sample size relative to log-rank test) for 80% power			
Analysis/study	**CheckMate017 (NCT01642004)**	**CheckMate057 (NCT01673867)**	**KEYNOTE-024 (NCT02142738)**	**CA184-043 (NCT00861614)**
Primary endpoint	**OS**	**OS**	**PFS**	**OS**
LRT	151	540	117	2,000+
RMST	158 (1.04)	724 (1.34)	126 (1.07)	1,687 (0.84)
Early-truncated Cox	213 (1.41)	275 (0.54)	109 (0.93)	772 (0.39)
Late-weighted LRT	163 (1.08)	236 (0.43)	53 (0.45)	522 (0.26)
Mid-weighted LRT	157 (1.04)	288 (0.53)	81 (0.69)	582 (0.29)
Early-weighted LRT	198 (1.31)	1,627 (3.01)	189 (1.61)	2,000+ (NA)
Milestone (80% of follow-up time)	2,000+ (NA)	200+ (NA)	648 (5.53)	2,000+ (NA)

NOTE: We assume enrollments rates and follow-up times as reported in the original manuscripts. Control of the type I error at 2.5%. Results are based on 10,000 simulated trials over a grid of sample sizes, with patient outcomes generated from the extracted KM curves.
Abbreviations: LRT, log-rank test; RMST, restricted mean survival time.

frequency of DPHs, reporting of summaries from the Grambsch–Therneau test or other tests to quantify the evidence of DPHs and visualizations of HR variations over time (e.g., Schoenfeld residual plots) should be routinely recommended when presenting trial results. If HR variations over time indicate nonmonotonic time-dependent treatment effects (HRs over time), then the evaluation and estimation of treatment effects requires complex statistical procedures. For instance, the nonparametric techniques proposed by Gray (18, 19) enable flexible estimation and testing of time-varying treatment effects.

There are several trial factors associated with DPHs. IO trials were more likely to exhibit DPHs with five of seven trials with at least one endpoint (PFS and/or OS) with evidence of DPH, consistent with other reports (1, 2, 20). IO trials show a characteristic increase in treatment effect over time (Fig. 1) in contrast with other therapies. In these trials, a delayed treatment effect, where there may not be sufficient time for an effective immune response in patients with aggressive and rapidly progressive disease, has been postulated as biological basis for DPHs (1).

Interestingly, nonsurvival clinical endpoints were associated with a higher prevalence of DPHs relative to OS endpoints (Table 2). It is possible that a higher prevalence of DPHs is seen with nonsurvival clinical endpoints because of its more proximal relationship to the trial's intervention. With 33% of examined nonsurvival KM figures exhibiting DPHs, our results suggest additional caution is warranted in designing clinical trials with a non-OS time to event endpoints as the primary outcome.

Our results support the potential in using prior trials to guide the choice of analysis. IO trials illustrate this with the presence of characteristic time-varying changes in HRs (Fig. 1; Supplementary Fig. S4). The aim of our analyses (Table 4) was to demonstrate that prior clinical trials from a specific setting where DPHs are likely to occur can be used to evaluate testing procedure with comparisons to the standard log-rank test. These analyses indicated potential efficiency gains of alternative testing procedures for specific scenarios (late-treatment effects). In our analysis, the use of late-weighted log rank tests can reduce the overall sample size requirement (up to 74%) in IO trials compared with the log-rank test (17). Given that approximately 800 IO trials were ongoing as of 2017 (21), our findings show the potential for designs that incorporate the expectation of increasing effects compared with PHs-based analyses.

When considering alternative parametric or semi-parametric non-PHs methodologies to test treatment effects, it is important to be aware of assumptions underlying these methodologies [e.g., accelerated failure time (ref. 22), proportional odds models (ref. 23), etc.], and evaluate the assumed relations between the survival functions of the experimental and control arms.

RMST has been introduced as a robust alternative to PH analyses. Several authors (5, 24–26) have investigated operating characteristics of RMST under PHs and DPHs (24, 25). Horiguchi and colleagues (27) developed an extension of RMST that adaptively selects the truncation time t^*. Other procedures to deal with DPHs include weighted log-rank tests using adaptive weights (28, 29). With respect to other approaches, Gray (18) has previously discussed tests to evaluate treatment effects under DPH using splines-based methods, and Schemper and colleagues (30) has described methods for estimating the average hazard ratio for time-varying treatment effects. More recently, Royston and Parmer described a two-stage algorithm to test violations of PHs and evaluate treatment efficacy (31). These methodologies allow for the evaluation of time-varying treatment effects, but they have yet to be implemented routinely in clinical practice.

An important limitation of our study is that there is no widely accepted metric to measure the magnitude of DPHs. We used the Grambsch–Therneau test, the most commonly used lack-of-fit test for proportionality of hazards. This may not capture DPHs in studies with a small number of events, while it may detect negligible departures from PHs in studies with a large sample size (32, 33). A limitation of the Grambsch–Therneau test for evaluating DPH is that the test may have low power to detect nonmonotone time–treatment interactions (34). Nonmonotone interactions can be detected by visualizing Schoenfeld residuals. Alternative procedure for testing generic time–treatment interactions have been reviewed by Therneau and Grambsch (10).

For our non-PHs analyses, we used prespecified parametrizations, for *en masse* application (e.g., restriction timepoint for RMST analysis, percentage follow-up for milestone analysis, etc.) and for ease of analysis, but alteration of such parameters could affect our analyses. The generalizability of our findings is limited due to our search criteria, which included only publications from a prespecified list of journals. Moreover, a limitation of our analyses of concordance/discordance between PHs-based log-rank testing and alternative non-PHs testing procedures is that we do not known whether the unknown survival functions for each of the 304 KM pairs satisfy the PHs assumption or not, neither do we know which experimental treatments has a true positive treatment effect. Some of the KM curves without evidence of DPHs might have been classified incorrectly as non-PHs violations and vice versa.

Conclusions

In conclusion, a substantial proportion of survival curves from phase III oncology clinical trials exhibited evidence of DPH, and this was more likely with non-OS endpoints and IO trials. Alternative approaches to design and analysis of clinical trials, which have been rigorously studied in the biostatistical literature, should be considered at time of study design. Data from previous trials on the experimental treatment can facilitate the choice of survival analysis methodology in phase III protocols. The use of alternative statistical procedures, including late-weighted log-rank tests for IO trials, has the potential to substantially increase trial efficiency with randomization of fewer patients and reduction of trial resources.

Disclosure of Potential Conflicts of Interest

B.M. Alexander is an employee of Foundation Medicine. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: R. Rahman, G. Fell, S. Ventz, L. Trippa, B.M. Alexander
Development of methodology: R. Rahman, G. Fell, S. Ventz, A. Arfé, L. Trippa, B.M. Alexander
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): R. Rahman, G. Fell, L. Trippa, B.M. Alexander
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): R. Rahman, G. Fell, S. Ventz, A. Arfé, L. Trippa, B.M. Alexander
Writing, review, and/or revision of the manuscript: R. Rahman, G. Fell, S. Ventz, A. Arfé, A.M. Vanderbeek, L. Trippa, B.M. Alexander

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): R. Rahman, S. Ventz, A.M. Vanderbeek, B.M. Alexander
Study supervision: S. Ventz, L. Trippa, B.M. Alexander

Acknowledgments

This work was supported by the Burroughs Wellcome Innovations in Regulatory Science Award.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 7, 2018; revised May 25, 2019; accepted July 15, 2019; published first July 25, 2019.

References

1. Alexander BM, Schoenfeld JD, Trippa L. Hazards of hazard ratios - Deviations from model assumptions in immunotherapy. N Engl J Med 2018;378:1158–9.
2. Chen T-T. Statistical issues and challenges in immuno-oncology. J Immunother Cancer 2013;1:18.
3. Rahman R, Ventz S, Fell G, Vanderbeek AM, Trippa L, Alexander BM. Divining responder populations from survival data. Ann Oncol. 2019 March 12. [Epub ahead of print].
4. Hernán MA. The hazards of hazard ratios. Epidemiol Camb Mass 2010;21: 13–5.
5. Trinquart L, Jacot J, Conner SC, Porcher R. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. J Clin Oncol 2016;34: 1813–9.
6. A'Hern RP. Restricted mean survival time: an obligatory end point for time-to-event analysis in cancer trials? J Clin Oncol 2016;34:3474–6.
7. Batson S, Greenall G, Hudson P. Review of the reporting of survival analyses within randomised controlled trials and the implications for meta-analysis. PLoS ONE 2016;11:e0154870.
8. Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan–Meier survival curves. BMC Med Res Methodol 2012;12:9.
9. Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika 1982;69:239–41.
10. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model [Internet]. New York: Springer-Verlag; 2000 [cited 2019 May 14]. Available from: https://www.springer.com/us/book/9780387987842.
11. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. Biometrika 1982;69:553–66.
12. Huang B, Kuan P-F. Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point. Pharm Stat 2018;17:202–13.
13. Chen T-T. Milestone survival: a potential intermediate endpoint for immune checkpoint inhibitors. J Natl Cancer Inst 2015;107.
14. Brahmer J, Reckamp KL, Baas P, Crinò L, Eberhardt WEE, Poddubskaya E, et al. Nivolumab versus docetaxel in advanced squamous-cell non–small-cell lung cancer. N Engl J Med 2015;373:123–35.
15. Borghaei H, Paz-Ares L, Horn L, Spigel DR, Steins M, Ready NE, et al. Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer. N Engl J Med 2015;373:1627–39.
16. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Pembrolizumab versus chemotherapy for PD-L1–positive non–small-cell lung cancer. N Engl J Med 2016;375:1823–33.
17. Kwon ED, Drake CG, Scher HI, Fizazi K, Bossi A, van den Eertwegh AJM, et al. Ipilimumab versus placebo after radiotherapy in patients with metastatic castration-resistant prostate cancer that had progressed after docetaxel chemotherapy (CA184-043): a multicentre, randomised, double-blind, phase 3 trial. Lancet Oncol 2014;15:700–12.
18. Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. J Am Stat Assoc 1992;87:942–51.
19. Spline-Based Tests in Survival Analysis on JSTOR [Internet]. [cited 2019 May 21]. Available from: https://www.jstor.org/stable/2532779?seq=1#page_scan_tab_contents.
20. Mick R, Chen T-T. Statistical challenges in the design of late-stage cancer immunotherapy studies. Cancer Immunol Res 2015;3:1292–8.
21. The Lancet Oncology null. Calling time on the immunotherapy gold rush. Lancet Oncol 2017;18:981.
22. Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Stat Med 1992;11:1871–9.
23. Kirmani SNUA, Gupta RC. On the proportional odds model in survival analysis. Ann Inst Stat Math 2001;53:203–16.
24. Pak K, Uno H, Kim DH, Tian L, Kane RC, Takeuchi M, 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.
25. Royston P, Parmar MKB. 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. Zhao L, Claggett B, Tian L, Uno H, Pfeffer MA, Solomon SD, et al. On the restricted mean survival time curve in survival analysis. Biometrics 2016; 72:215–21.
27. Horiguchi M, Cronin AM, Takeuchi M, Uno H. A flexible and coherent test/estimation procedure based on restricted mean survival times for censored time-to-event data in randomized clinical trials. Stat Med 2018;37: 2307–20.
28. Yang S, Prentice R. Improved logrank-type tests for survival data using adaptive weights. Biometrics 2010;66:30–8.
29. Uno H, Tian L, Claggett B, Wei LJ. A versatile test for equality of two survival functions based on weighted differences of Kaplan-Meier curves. Stat Med 2015;34:3680–95.
30. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted Cox regression. Stat Med 2009;28:2473–89.
31. Royston P, Parmar MKB. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. BMC Med Res Methodol 2016;16:16.
32. Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol 2014;32:2380–5.
33. Royston P, Parmar MKB. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. Trials 2014;15:314.
34. Grant S, Chen YQ, May S. Performance of goodness-of-fit tests for the Cox proportional hazards model with time-varying covariates. Lifetime Data Anal 2014;20:355–68.