BMC Medical Research Methodology

RESEARCH

Open Access

Beyond hazard ratios: appropriate statistical methods for quantifying the clinical effectiveness of immune-oncology therapies – the example of the Netherlands

Isaac Corro Ramos[1*], Venetia Qendri[2] and Maiwenn Al[2]

Abstract

Background The Dutch Committee for the Evaluation of Oncological Drugs evaluates the effectiveness of new oncological treatments. The committee compares survival endpoints to the so-called PASKWIL-2023 criteria for palliative treatments, which define if treatment effects are considered clinically relevant. A positive recommendation depends on whether the median overall survival (OS) is below or above 12 months in the comparator arm. If the former applies, an OS benefit of at least 12 weeks, and a hazard ratio (HR) smaller than 0.7 are required. If the latter applies, an OS or progression free survival (PFS) benefit of at least 16 weeks, and an HR smaller than 0.7 are required. Nonetheless, the median survival time may not be reached and the proportional hazards (PH) assumption, quantified by the HR, is likely violated for immuno-oncology (IO) therapies, deeming these criteria inappropriate.

Methods We conducted a systematic literature review to identify statistical methods used to represent the clinical effectiveness of IO therapies based on trial data. We searched MEDLINE and EMBASE databases from inception to August 31, 2022, limited to English papers. Methodological studies, randomized controlled trials, and discussion papers recognising key issues of survival data analysis of IO therapies were eligible for inclusion.

Results A total of 1,035 unique references were identified. After full paper screening, 17 publications were included in the review. Additionally, 43 papers were identified through 'snowballing'. We conclude that the current PASKWIL-2023 criteria are methodologically incorrect under non-PH. In that case, single summary statistics fail to capture the treatment effect and any measure should be interpreted in combination with the Kaplan-Meier curves. We recommend 'parameter-free' measures, such as the difference in restricted mean survival time, avoiding assumptions on the underlying survival.

Conclusions The HR is commonly used to assess treatment effectiveness, without investigating the validity of the PH assumption. This happens with the application of the PASKWIL-2023 criteria for palliative oncology treatments, which can only be valid under a PH setting. Under non-PH, alternative treatment effect measures are suggested. We propose

*Correspondence:
Isaac Corro Ramos
corroramos@imta.eur.nl

Full list of author information is available at the end of the article

a step-by-step approach supporting the choice of the most appropriate methods to quantify treatment effectiveness that can be used to redefine the PASKWIL-2023 criteria, or similar criteria in other clinical areas.

Keywords Immunotherapies, Oncology, Proportional hazards, Weighted log-rank test, Restricted mean survival time, Hazard ratio

Background

Over the past two decades, immuno-oncology (IO) has re-emerged as a promising approach to treat cancer, resulting in improved patient survival and quality of life [1]. The biological mechanism of action of novel IO agents, targeting the immune system to indirectly induce antitumor response, often poses unique statistical challenges in the determination of their effectiveness compared to conventional targeted therapies and/or chemotherapies, aiming to directly destroy cancer cells. One of the challenges is related to the statistical analysis of survival endpoints, such as overall survival (OS) or progression-free survival (PFS). Usually, in randomized clinical trials (RCTs) with survival endpoints, the ratio of the hazard functions of two different treatments is assumed to be constant over time. This assumption is commonly referred to as the proportional hazards (PH) assumption, and this constant ratio of the hazard functions is referred to as hazard ratio (HR). If this assumption is satisfied, this single HR is used to express the relative difference in survival benefit of two treatments. As a result of the IO mechanism of action, it is not uncommon to observe a delayed effect associated to IO therapies. This delayed effect is characterized by a late separation in the Kaplan-Meier (KM) curves, often combined with a plateau in the tail of the survival curve (often referred to as "cure"). As a consequence, the PH assumption is likely to be violated for IO therapies [2–4], since assuming PH implies immediate treatment effects (survival curves separate from the start) which are maintained over time in a constant fashion (no waning in treatment effects and no plateau in the survival curves).

Despite being commonly assumed; evidence suggests that the PH assumption is often violated. The study by Rahmadian et al. 2020 showed that from a total of 94 oncology clinical trials, the PH assumption was violated in 45 of them (48%); in particular, 9 out of 13 (69%) of the IO trials and 36 out of 81 (44%) of the non-IO trials [5]. Although the invalidity of the PH assumption seems to be more evident for IO therapies [6–10], it is noteworthy that many non-IO therapies do not comply with the PH assumption either [11–16]. In general, in the presence of a delayed separation of the survival curves or a survival plateau, a constant HR over time cannot be assumed. Estimating the HR in such cases may lead to biased and erroneous treatment effect estimates [17].

In the Netherlands, the Committee for the Evaluation of Oncological Drugs (cieBOM) evaluates a new oncological drug after the results of a randomized and comparative study have been published as a full paper in a peer-reviewed journal and after it has been approved by the European Medicines Agency. The cieBOM compares the endpoints of the study to the so-called PASKWIL-2023 criteria [18]. The name of the criteria is derived from the words **P**alliatief (palliative), **A**djuvant, **S**pecifieke bijwerkingen (specific side effects), **Kw**aliteit van leven (quality of life), **I**mpact van behandeling (impact of treatment), and **L**evel of evidence, which are mostly in Dutch language. According to the PASKWIL-2023 criteria for palliative oncology treatments, a positive recommendation regarding the effectiveness of a therapy depends on whether the median OS is below or above 12 months in the comparator arm. If the former applies, an OS benefit of at least 12 weeks, and an HR smaller than 0.7 are required. If the latter applies, an OS or PFS benefit of at least 16 weeks, and an HR smaller than 0.7 are required. Note that the thresholds applied reflect a minimum clinically relevant difference [18]. Whether the OS or PFS benefit is assessed depends on the primary endpoint of the randomized clinical trial [19]. A negative recommendation means the drug is not meeting the state-of-the-art in science and practice and will, therefore, not be included in the Dutch basic health insurance package (this package covers almost all medical care, e.g., general practitioners, specialists, hospital treatment and admission, devices, diagnostics, and most medical prescriptions). As explained above, assessing the effectiveness of IO therapies based on a single HR can be problematic in some cases. In addition to this, in some clinical studies the median survival time may not be reached and even if it is observed, it may not capture a delayed separation of survival curves or long-term survival benefits. Therefore, assessing the effectiveness of a new therapy based on the difference in median survival time can also be problematic. All these suggest that the current PASKWIL-2023 criteria may be inappropriate for determining the effectiveness and added value of IO therapies. To address the statistical challenges in measuring the relative effectiveness of IO therapies, alternative endpoints and statistical methods have been developed [4], including milestone analysis, restricted mean survival time, the weighted log-rank or weighted Kaplan-Meier tests.

This paper explores how the PASKWIL-2023 criteria to assess effectiveness of oncology treatments could be revised to accommodate the special characteristics of IO

therapies. A systematic literature review (SLR) was conducted for within-trial statistical analysis methods (as opposed to extrapolation-based methods that are often employed in health technology assessments). The review aimed to identify alternative statistical methods used to represent the effectiveness of IO therapies (e.g., with a delayed treatment effect) considering trial data. The most relevant statistical methods identified in the SLR are described separately. Finally, potential alternatives to the current PASKWIL-2023 criteria are proposed.

Methods

We conducted a systematic review aiming to identify alternative statistical methods used to represent the clinical effectiveness of IO therapies based on trial data.

Data sources

We searched MEDLINE and EMBASE databases. The reference lists from the included full manuscripts were also searched. The search terms were: 'immunotherapy'/exp AND 'cancer'/exp OR 'neoplasms'/exp OR 'cancer immunotherapy'/exp OR 'immuno oncology'/exp AND 'survival analysis'/exp OR 'survival benefit' OR 'nonproportional hazards' AND 'methodology'/exp OR 'methods'/exp, where /exp stands for "explode" and it is used to retrieve any subcategory terms included in the Medical Subject Headings (MeSH) terms. The database search was performed from inception to August 31, 2022, with papers only published in English. Additional papers were found through 'snowballing', by reviewing the reference list of the relevant primary studies identified through the database searches [20].

Inclusion criteria

Studies presenting statistical methods to estimate the clinical effectiveness of IO therapies were eligible for inclusion. Randomized controlled trials, and discussion papers, that recognised key issues with analysis of survival data of IO therapies and addressed these issues using alternative statistical options were also considered.

Data extraction

Two investigators (VQ and ICR) independently screened abstracts and identified articles using the predefined inclusion criteria. VQ, ICR and MA read the full manuscripts of the included studies. Disagreements in the included studies were resolved by consensus. ICR and VQ independently extracted the data using a standardized data extraction form (Additional file 1). Disagreements in the extracted data were again resolved by consensus.

We extracted data on the following: methods discussed, measure of effect size presented (Yes/No), HR-based methods (Yes/No), KM-based methods (Yes/No), methods of extrapolations (Yes/No), comparison of treatment effect methods (Yes/No).

Results

Overview of included studies

Our searches identified a total of 1,035 unique references. After initial screening of titles and abstracts, 38 references were considered to be potentially relevant. After full paper screening, 17 publications were included in the review [5, 21–36]. Additionally, 56 papers were identified through 'snowballing'. From these, 43 publications were included in the final review [2–4, 17, 37–75]. Figure 1 provides a flowchart of the review process. This figure was adapted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 flow diagram [76], and was created using the PRISMA2020 flow diagram R package [77]. The full extraction table can be seen in Additional file 1.

Statistical approaches

The included papers mostly covered three different statistical approaches: weighted log rank testing, testing based on Kaplan-Meier curves, and parametric extrapolation. The first two approaches concern testing statistically the difference between two sets of survival data, and in general they do not quantify a treatment effect. Based on the weighted log-rank test, the so-called average or weighted hazard ratio is often proposed to estimate treatment effect, whereas the difference (or ratio) in restricted mean survival time (RMST) is frequently used when KM methods are employed. Parametric extrapolation does not address statistical hypothesis testing but as explained below, can be used to overcome some statistical issues associated with the analysis of IO therapies data. These three approaches together with piece-wise HRs and milestone analysis are first discussed separately. Table 1 presents an overview of the most frequently reported measures of treatment effects, with some of their advantages and disadvantages. This section ends with a brief discussion of studies that present results comparing alternative methods to the common HR using patient-level data or simulated datasets.

Weighted log rank test and average hazard ratio

A substantial subset of the papers included in our review present variations of the log rank test (LRT), which is often used to test the hypothesis that two survival curves are the same. The LRT has optimal power when the PH assumption applies and the treatment effect is usually expressed by an HR, which can be estimated using the Cox regression model. However, for IO-therapies the PH assumption often does not hold, as survival curves for intervention and comparator may separate only after a certain time lag t (i.e., until t the

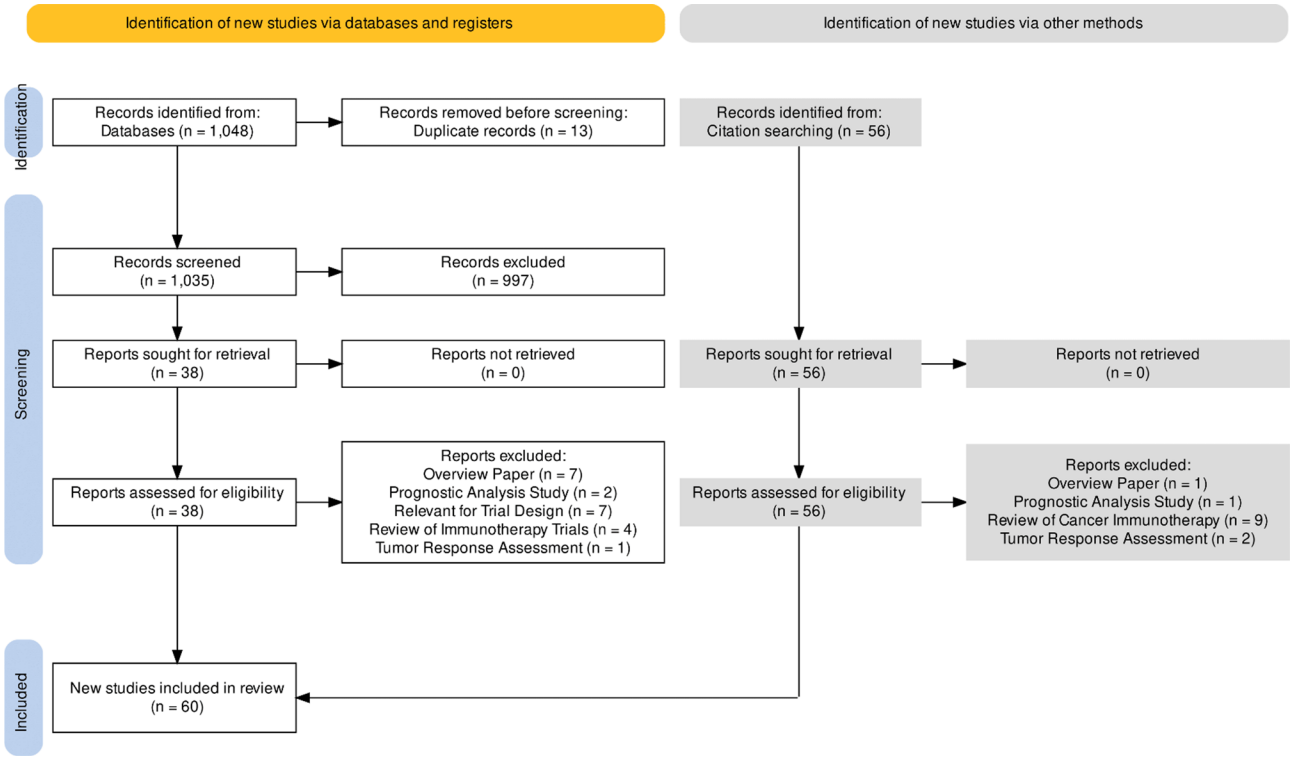

Fig. 1 Flow of papers through the review process

HR is 1, whereas thereafter the HR is either smaller or larger than 1). To deal with such situations, a weighted log rank test (WLRT) is suggested. In general, the purpose of the weights is to emphasize a certain part of the survival curve, whilst the standard LRT gives equal weight to all parts. Different WLRTs can be found in the literature, depending on the assumptions made to define the weights, which vary in complexity. A summary of the most frequently reported WRLT approaches can be found in Additional file 2.

The papers presenting WLRTs mostly focused on hypothesis testing and trial design and not on quantifying treatment effect. Yu et al. 2021 discussed how the weights derived for the WLRT can also be used to estimate an average HR (AHR), providing an estimate of overall treatment effect [36]. In principle, every different set of weights may result in a different estimate of the AHR. The idea of estimating an AHR when the PH assumption is not valid was already proposed by Kalbfleisch and Prentice back in 1981 [78], and further elaborated by Struthers and Kalbfleisch 1986 [69]. Rauch et al. 2018 provided a thorough overview on how an AHR can be derived [79]. They explored two approaches, one based on Kalbfleisch and Prentice 1981 and the other based on Schemper et al. 2009 [78, 80]. The paper provides the results of a simulation study exploring the estimated treatment effect comparing the standard Cox regression-based HR and several variations of the AHR. The R packages *coxphw* and *AHR* were used to estimate the AHR. The package *coxphw* implements the ideas of Schemper [81], and the package *AHR* implements the ideas of Kalbfleisch and Prentice; however, *AHR* has been removed from the CRAN repository [82]. A time-varying HR to quantify treatment effect (based on time dependent Cox model) was proposed in Lin and León 2017 [64], who also compared their method against a short- and long-term hazard ratios model previously developed by Yang and Prentice [60, 72]. Therneau et al. 2023 and Zhang et al. 2018 explained the implementation of time-dependent Cox regression models in R [83, 84].

Weighted differences in Kaplan-Meier curves and restricted mean survival

An alternative approach to test the hypothesis that two survival curves are the same is based on the weighted difference in Kaplan-Meier (WKM) estimates. Unlike the WLRT, this method does not rely on assumptions about the analytical form of the hazard functions. Different WKM statistics can be found in the literature with varying complexity. Details can be found in Additional file 3.

The most popular approach is based on the RMST. The RMST is defined as the mean survival time truncated at $\tau > 0$ and, therefore, it is simply the area under the survival curve $S(t)$ from 0 to τ. The difference (or ratio) in RMST is often used to estimate the added benefit of a treatment as an alternative to the HR or the difference in

Table 1 Comparison of most frequently reported measures of treatment effects

	Current PASKWIL-2023 criteria		Most common alternatives from literature				
	Single HR	**Median survival***	**Weighted HR**	**RMST****	**MST based on parametric modelling**	**Milestone analysis**	**Piecewise HR's**
Assessment of treatment effect	HR.	Incremental median PFS/OS.	Weighted (average) HR.	PFS/OS difference (area below KM curves).	PFS/OS difference (area below parametric curves).	Survival prob. at time $\tau > 0$.	Set of HR's applied to different time intervals.
Underlying statistical approach	Log-rank test (PH).	None.	Weighted LRT (varying weights).	Weighted KM tests ("parameter-free").	Assumptions about survival shape.	None.	Piecewise PH.
Minimal clinically relevant difference	HR < 0.70.	Difference > 12 or 16 weeks.	To be established.	To be established.	To be established.	To be established.	To be established.
Pros of method	Commonly used. Easy to estimate. Valid under PH.	Commonly used. Captures quick effects. Clinically meaningful. Assumption-free. Inference procedure for difference (ratio).	Valid under non-PH. Can accommodate different survival shapes (delayed effect, crossing hazards or long-term plateau).	Valid under non-PH. Clinically meaningful. Utilizes the whole shape of the KM curve. Assumption-free.	Predicts unobserved long-term benefits. Can accommodate different survival shapes (also non-PH). Well-established methods (in health economics).	Valid under non-PH. Easy to read of survival curve. Easy to interpret. Inference procedure for the difference. Assumption-free.	Easy to interpret. Can accommodate different survival shapes. Valid under non-PH.
Cons of method	Biased and difficult to interpret under non-PH. May require large study (HR depends on number of observed events). May focus on higher-risk population (many observed events needed).	Not always observed. Does not capture long-term benefits. Only reflects cumulative information at fixed (median) time. Does not reflect differences in profiles of the survival curves.	Same as HR. Results depend on chosen time (e.g., last HR). Subject to selection bias.	Results depend on chosen time. May be difficult to interpret. May focus on healthy population with low event rates. Rarely reported.	Needs assumptions about type of extrapolation (not only observed data). May introduce uncertainty. Not frequently used to assess clinical effectiveness.	Depends on milestone time. Imprecise if applied too early or too late. Only reflects cumulative information at fixed time. Does not reflect differences in profiles of the survival curves.	Same as HR and weighted HR.

Abbreviations Diff. = difference; HR=hazard ratio; KM=Kaplan-Meier; LRT=log-rank test; MST=mean survival time; OS=overall survival; PFS=progression-free survival; PH=proportional hazards; Prob. = probability; RMST=restricted mean survival time

* Can be generalised to any quantile/percentile. The median is one case. It can also be applied to other endpoints, even though OS/PFS are the most used

** Can be generalised to other measures based on KM curves, such as the restricted survival benefit or the area between two survival curves

median survival time (especially when the median survival time is not observed). When estimating the RMST, the cut-off point τ needs to be specified before data analysis to avoid bias. A fixed time point of specific clinical relevance can be used to define τ (e.g., x years), whereas Huang and Kuan 2018 suggest using either (1) the minimum of the largest observed event time of each treatment group, or (2) the minimum of the largest observed time (event or censoring) of each treatment group [59]. The R package *survRM2* provides different options to estimate the RMST [85].

The concept of the RMST was generalized to the area between two survival curves (ABS) and the restricted survival benefit (RSB) [39, 75]. The ABS is the accumulated absolute difference in treatment effects from time 0 to the end of follow-up. Thus, whilst the difference in RMST represents a relative difference in effects, ABS describes the absolute difference. When survival curves do not cross, ABS is equal to the difference of the two RMST. The RSB is the expected value of a function of the survival times under different treatments, including the difference in RMST, but can also be the restricted chance of a longer survival. These measures do not require the

PH assumption to hold and may thus be more meaningful than the HR in assessing the treatment efficacy for IO therapies. R code is available to estimate the ABS [39], whereas this does not seem to be the case for the RSB.

Parametric extrapolation

Parametric extrapolation of observed KM curves is often used to estimate survival beyond the observation period (usually over patients' lifetime). This is commonly used in health economic modelling, but less often in decision making about clinical effectiveness. Parametric extrapolation can potentially overcome the issues previously identified with IO therapies, since the methods described above can also be applied to extrapolated parametric survival curves. For example, parametric survival curves allow estimation of the HR (and validation of the PH assumption) or the RMST. An example of the latter is given in Royston and Parmar 2013, where the RMST is estimated assuming that the survival time follows a piecewise exponential distribution [45]. Also, given that extrapolations are used for a lifetime time horizon, the mean survival can be estimated, which is usually not possible with KM curves. However, parametric extrapolations entail uncertainty as it may be unclear what type of parametric curve fits better to the observed data, but more importantly, due to the immaturity of the observed data (e.g., OS data), it is in general unknown how these curves look like in the long-term.

Despite the usually wide range of shapes that are represented by standard parametric survival functions (e.g., exponential, Weibull, or generalised Gamma), a good fit to observed data is not guaranteed. In those situations, the so-called parametric spline models (polynomial functions with knots representing the point where pieces connect) might prove a better fit [43]. When a plateau appears in the survival curve, the so-called *cure* models can be used to improve the fit of the parametric curve and the validity of the extrapolated values. Five additional papers included in our review addressed the issue of estimating long-term effects of IO treatments [21, 22, 28, 37, 41]. Overall, these papers conclude that more complex survival models, such as mixture cure or spline-based models, seem to provide a better estimate of long-term survival associated to IO, compared to standard approaches. Additional details are provided in Additional file 4. R offers many resources related to parametric survival modelling such as the packages *survival* or *flexsurv* [86, 87].

Other methods

An alternative approach is the so-called milestone analysis [23]. Even though it can be argued that milestone analysis could be included in the procedures based on the (weighted) difference of KM curves presented above, it has received sufficient attention in the literature to be discussed separately [4, 23, 53, 56, 63, 88]. The test statistic for a milestone analysis at time $\tau > 0$ can be simply defined as the difference in survival probabilities (as estimated by KM curves) at time τ. An overview of the methods used to analyse survival data at fixed time points is provided in Klein at al. 2007 [89]. As it happened with the WKM methods described above, the milestone time τ at which the survival of the two groups is compared has to be chosen in advance. For details on how to properly select the milestone time, we refer to the section above. Note that some papers erroneously refer to milestone analysis as landmark analysis. A landmark analysis, however, is used for comparing survival between sub-groups of patients defined by response status [90]. Landmark analysis results in an unbiased estimate of survival probabilities conditional on sub-group membership of patients at a pre-specified time, the landmark time. Survival estimates from that time point are calculated conditional on patients' landmark responses. Therefore, patients who died before the landmark time, are excluded from the subsequent analysis.

Another approach sometimes mentioned is the estimation of so-called piecewise constant HRs. For example, in the paper by He et al. 2013 an approach is discussed where the point at which the HR changes is detected from the data [58]. Strictly speaking, the methods discussed as weighted log rank testing also assume piecewise constant HRs, as they assume that during the delayed treatment effect the $HR = 1$, and after that time the $HR \neq 0$. However, it has been pointed out that even in an RCT setting these piecewise HRs cannot be given a causal interpretation, since selection bias will occur in the estimation of the second (or later) HR due to conditioning on a post-treatment variable (i.e., being alive at the switch point) which is affected by treatment [88, 91]. Snapinn et al. 2022 introduced the concept of generalised hazard difference (GHD) as an alternative measure of treatment effect [57]. The GHD integrates the difference in milestone time survival and the difference in RMST into one single measure.

Comparisons of methods

Sixteen of the included papers compare the HR and one or more alternative measures of effectiveness in real and simulated datasets. Table 2 provides an overview of these 16 studies, and shows which methods were compared, whether real RCT or simulated datasets were used, the number of RCTs and simulated datasets on which the methods were tested, and the characteristics of the simulated data sets. A summary of the findings of each of these studies can be found in Additional file 5.

For some of the comparison studies it was possible to draw some conclusions; in these studies, it was found

Table 2 Measures of treatment effects included in comparison/simulation studies

Author	HR	WHR/AHR	Median	Diff. RMST	Milestone	Other effectiveness measures	Applied to # RCTs	Number simulated datasets	Characteristics simulated datasets
Connock et al. 2019 [1]	√	-	-	√	-	EMST	11	-	-
Huang et al. 2022 [2]	√	-	√	√	-	ABS	3	-	-
Huang and Kuan 2018 [3]	√	-	-	√	-	-	-	12	PH, late separation, crossing hazards.
Jiménez 2022 [4]	√	√	-	√	-	-	-	3	Delayed effect, crossing hazards, decreasing effect.
Lin and León 2017 [5]	√	√	-	-	-	Piecewise HR	-	30	Delayed treatment effect, reduced long-term treatment effect due to discontinuation.
Lin et al. 2020 [6]	√	√	-	√	-	Piecewise HR	2	9	No effect, delayed effect, diminishing effect, crossing hazards, delayed effect with converging tails.
Pak et al. 2017 [7]	√	-	√	√	-	-	1	-	-
Rauch et al. 2018 [8]	√	√	-	-	-	-	1	5	PH, increasing and decreasing HR.
Roychoudhury et al. 2021 [9]	√	√	-	√	-	Percentage net benefit	3	-	-
Royston and Parmar 2011 [10]	√	-	-	√	-	-	3	-	-
Royston and Parmar 2013 [11]	√	-	-	√	-	-	4	-	-
Snapinn et al. 2022 [12]	-	-	-	√	√	GHD	-	1	Delayed effect.
Trinquart et al. 2016 [13]	√	-	-	√		Ratio RMST	54	-	-
Uno et al. 2014 [14]	√	-	-	√	√	Percentile survival	3	-	-
Uno et al. 2015 [15]	√	-	-	√	√	Percentile survival	-	-	-
Zhang et al. 2021 [16]	-	-	-	√	-	Restricted prob. longer survival, restricted prob. longer survival by %, restricted net prob. longer survival	1	4	No effect, constant, early, and delayed effect.

Abbreviations ABS=area between survival curves; AHR=average hazard ratio; Diff. = difference; EMST=expected mean survival time; GHD=generalised hazard difference; HR=hazard ratio; PH=proportional hazards; Prob. = probability; RCT=randomised clinical trial; RMST=restricted mean survival time; WHR=weighted hazard ratio

that the direction of the effect was consistent between the estimated HR and the estimated RMST, as was the statistical significance. In other words, if an HR indicated that a new treatment improved survival, the same was observed based on the RMST and if an HR was statistically significant different from 1, the RMST was statistically significant different from 0.

However, in other comparison studies, the results were more diffuse, suggesting that no single measure will be preferred under non-PH.

An alternative to the PASKWIL-2023 criteria

Based on this review, we conclude that the current PASKWIL-2023 criteria are likely to be too simplistic and/or methodologically incorrect under non-PH. We propose a step-by-step approach more in line with the procedure described in Roychoudhury et al. 2021 [56]. The steps are summarised in Table 3.

The suggested approach starts with a visual assessment of the KM curves. Overlapping curves would suggest no additional treatment effect. If the median OS/PFS is not reached, the first PASKWIL-2023 criterion (i.e., OS/PFS gain of at least 12 or 16 weeks) cannot be applied. When curves separate early, it might be an indication of PH. A delayed separation, crossing curves or a long-term plateau would suggest non-PH, and then the second PASKWIL-2023 criterion (i.e., an HR smaller than 0.7) will most likely be invalid. The type of non-PH needs to be assessed based on the KM curves and/or prior clinical experience. Following the visual assessment of the KM

Table 3 Step-by-step approach to derive alternative PASKWIL-2023 criteria

Step 1 – Visual assessment of the KM curves	
1 A	Assess whether KM curves overlap (no effect) or deviate (possible effect). If the latter occurs, assess whether (1) curves separate early (might indicate PH) or not, (2) there is a delayed separation, (3) curves cross or (4) there is a long-term plateau.
1B	If the median OS/PFS is not reached, the first PASKWIL-2023 criterion (i.e., OS/PFS gain of at least 12 or 16 weeks) is not possible to assess.
1 C	If non-PH is suspected, the second PASKWIL criterion (i.e., an HR smaller than 0.7) will most likely be invalid. If non-PH is assumed, assess the type of non-PH that is to be expected (e.g., delayed effects, crossing hazards, long-term survival, etc.). This could be based on the KM curves and/or prior clinical experience.
Step 2 – Formal testing of PH assumption (nice to have)	
2 A	Test the PH assumption, e.g., log-log cumulative hazard plot or Schoenfeld residuals plot.
Step 3 – Quantify treatment effect	
3 A	If the PH assumption seems plausible, the current PASKWIL-2023 criteria (the HR – from Cox regression – or the difference in median survival) can be used as treatment effect measures.
3B	If there is strong evidence to assume that the PH assumption is violated, any single summary statistic will fail to capture the overall (time-varying) treatment effect. In this case, it is important to interpret any treatment effect measure in conjunction with what is observed in the KM curves. Ideally, the choice of the preferred treatment effect measures should be properly justified and guided by the outcomes obtained in the previous steps.
3 C	We would recommend 'parameter-free' measures such as the difference or ratio in RMST*, or the difference in milestone survival at time $\tau > 0$, because these do not make any assumption about the underlying mechanism of survival**.
3D	Alternative measures that are not 'parameter-free' that could also be used are for example the average HR or weighted HR.
3E	Any treatment effect measure should be reported with their corresponding measure of uncertainty (e.g., confidence intervals).

Abbreviations HR = hazard ratio; KM = Kaplan-Meier; OS = overall survival; PFS = progression-free survival; PH = proportional hazards; RMST = restricted mean survival time

*Scenarios where survival curves cross should be assessed with caution, paying extra attention to other factors such as the treatment effect before and after crossing, the timing of crossing and what measures are more appropriate to capture the overall benefit. The latter point could be assessed for example with more general version of the RMST (such as the area between two survival curves or the restricted survival benefit)

**Careful consideration must be given to the choice of time $\tau > 0$, which should be pre-specified in advance. It could be defined as the minimum of the maximum observed survival in intervention and control arms, but other choices are also possible

curves, the PH assumption should be evaluated in more detail by a log-log cumulative hazard plot or a Schoenfeld residuals plot before quantifying the treatment effect. If the PH assumption seems plausible, the second PASKWIL-2023 criterion can be considered appropriate. However, if there is clear evidence that the PH assumption is violated, the current HR criterion is incorrect. In this case, any single summary statistic will fail to capture the overall (time-varying) treatment effect. This is especially true for measures based on a single point on the survival curve such as the median survival time or milestone time survival rate, but also for summary measures based on interval time, such as a weighted HR, a RMST, or a generalized hazard ratio. Instead, a set of measures should be considered, in conjunction with the Kaplan-Meier curves. The choice of these preferred treatment effect measure should be properly justified and guided by the outcomes obtained in previous steps. In general, we would suggest 'parameter-free' measures, such as the difference in RMST, or the difference in milestone survival at time $\tau > 0$, because these do not require assumptions about the underlying survival mechanism. Scenarios where survival curves cross should be assessed with extra caution. Alternative not 'parameter-free' measures that could also be used are the average HR, the weighted HR, or the piecewise HR. To conform to best practices in statistics, these criteria should not be based on point estimates only, as this would ignore the uncertainty around such estimates. Confidence intervals should be included in the criteria as well. Finally, it worths mentioning that the suggested step-by-step approach cannot (yet) replace the current PASKWIL criteria, as it does not set any cut-off points in the criteria that can be applied under non-PH setting. Such cut-off points should be based on clinically meaningful differences, which need to be determined by the committee cieBOM in consultation with clinical experts in IO.

Discussion

This paper explored whether the current PASKWIL-2023 criteria used in the Netherlands to assess treatment effectiveness for palliative oncology treatments are appropriate in the context of new IO therapies. To this end, we conducted an SLR as described in Sect. 2 of this paper. Our review retrieved 60 papers.

Various papers were identified highlighting the invalidity of the PH assumption in studies involving IO therapies [3, 5, 26, 31, 32, 34–37, 49, 50, 54, 56, 59, 67, 71, 73]. This is the case when for example survival curves show a delayed separation or when they cross. In these situations, the standard LRT, commonly used to test for a difference in treatment effect, is not efficient, and therefore, the HR under PH will not capture the treatment benefits appropriately. Studies discussing issues related to non-PH go back to the 1980's and 1990's; thus, even before the use of IO therapies [42, 52, 62, 69, 78]. Despite this, the

HR is still commonly used to assess treatment effectiveness, without paying sufficient attention to the validity of the PH assumption. This happens for example in the Netherlands with the application of the PASKWIL-2023 criteria, based on which a positive recommendation on the effectiveness of a new oncology palliative therapy requires a gain in (median) OS or PFS of at least 12 or 16 weeks, or an HR smaller than 0.7 [10]. However, these criteria seem to be applied regardless the validity of the PH assumption.

When the PH does not hold, scientific literature suggests other measures of treatment effectiveness instead of the HR. The performance of those measures has been studied extensively, with advantages and disadvantages been widely discussed. Although there is no unanimous consensus regarding the most appropriate treatment effectiveness measure under non-PH, a substantial number of papers found in our review seem to discourage the use of *the* HR, even under a PH assumption [17, 47, 53, 54, 56, 63, 88]. Bartlett et al. 2020 recommended contrasts of survival probabilities at specified times (milestone analysis), contrasts of specified quantiles (e.g., median survival), or differences/ratios of RMST [88]. In the same line, Hellmann et al. 2016 proposed the increased use of survival estimates at specific milestones [63], and Uno and colleagues supported the use of "parameter-free" measures such as the RMST or the difference in median survival [17, 47]. Freidlin and Korn 2019 on the other hand, acknowledged that trying to estimate the treatment effect using a single measure of benefit is challenging, and concluded that more convincing evidence is needed before abandoning the LRT (and thus the HR under PH) as primary analyses in RCTs [92]. Other papers suggest that it is unlikely that a single effect measure can adequately capture the overall benefit of the treatment, given the potential complexity related to treatment effect changes over time [54, 56]. Based on the above, we concluded that the current PASKWIL-2023 approach is invalid under non-PH. We proposed a step-by-step procedure as described in Sect. 3.3, which is in line with the approach suggested in Roychoudhury et al. 2021 [56].

The underlying aim of our review was to identify studies presenting alternative statistical methods that can be used to assess the effectiveness of IO treatments whilst accounting for the special characteristics in the survival data of such treatments, including a potential long-term plateau or a delayed separation of the survival curves. Although new therapies in other disease areas may also encounter similar statistical challenges in showing a clinical benefit, our review focused on IO therapies because such cancer therapies are more prone to present survival data deviating from conventional survival analyses patterns [5]. Examples of these non-conventional survival patterns, including crossing survival curves, have been observed for example in pembrolizumab [6, 9], nivolumab [7, 10], or atezolizumab studies [8]. The European Society for Medical Oncology (ESMO) and the American Society of Clinical Oncology (ASCO) recognised the common challenges related to the measurement of the clinical benefit of IO therapies, and revised accordingly their frameworks that were developed to assess the magnitude of the clinical benefit of new and effective cancer therapies [93, 94]. Specifically, the ESMO Magnitude of Clinical Benefit Scale (ESMO-MCBS) tool suggested an upward adjustment in the scoring system of the treatment if there is a long-term plateau in the survival curve, and the OS advantage continues to be observed at 5 years (or 7 years for diseases with median survival > 24 months) [93]. The ASCO Framework for Assessing Value in Cancer Care awarded bonus points to the new treatment if there is a ≥ 50% improvement in the proportion of patients alive with the test regimen at the time point that is twice as high as the median OS or PFS of the comparator treatment (assuming > 20% surviving with standard care) [94].

Since our main interest was in IO therapies, our literature search used IO-related terms (i.e. immunotherapy' AND 'cancer' OR 'neoplasms' OR 'cancer immunotherapy' OR 'immuno oncology'). Because of this, a substantial number of general mathematical/statistical papers dealing with non-PH situations were not captured in the initial search process and were found through "snowballing" [2–4, 17, 37–75]. Some of these papers were deemed fundamental and made us acknowledge that our search terms may have not been optimal; however, our intention was to avoid an overload of irrelevant titles and abstracts. Moreover, since our review identified some recent overview papers [53, 54, 56, 57], it is likely that the impact of our omissions will have been negligible.

In our suggested step-by-step approach, we did not set any cut-off points in the proposed criteria since such cut-off points should be based on clinically meaningful differences, and therefore, need to be strongly guided by clinical experts in IO. Future research could be aimed at determining the difference or ratio in RMST or AHR that would be equivalent to the PASKWIL-2023 criterion of an HR < 0.7. Additionally, it should be explored what difference or ratio in RMST can be considered clinically relevant. It might also be of interest to explore the appropriateness of the current PASKWIL-2023 criteria for the evaluation of the IO therapies via a targeted literature search (TLR). The TLR can identify the IO therapies that have been accepted/rejected during the past years in the Netherlands and how many of these treatments showed a delayed and potentially durable treatment effect. For the identified IO therapies with a positive/negative cieBOM recommendation based on PASKWIL-2023 criteria,

further investigation could be conducted to see how our proposed step-by-step approach might have been implemented in comparison with the current PASKWIL-2023 criteria, or to assess how similar submissions have been assessed in other countries such as Belgium or the United Kingdom.

Conclusions

We recommend a step-by-step approach to assess the treatment effectiveness of palliative oncology therapies in the Netherlands. The currently used criteria can only be considered valid under a PH setting and/or when the observed survival data reach the median. A non-PH setting is not uncommon in the context of IO for oncology and immature survival data is also a frequent issue. Alternative treatment effect measures, such as the difference in RMST or milestone survival probabilities are suggested. Such "parameter-free" measures do not depend on assumptions about the underlying survival mechanism, are easy to explain and to estimate. The step-by-step procedure suggested in this paper, should be seen as a flexible tool supporting the choice of the most appropriate methods to quantify potential treatment effects. Ultimately, this procedure could be used to redefine new more comprehensive PASKWIL-2023 criteria, or other criteria in other clinical areas since the PH framework is not exclusive of IO therapies.

Abbreviations

ABS	Area between two survival curves
AHR	Average hazard ratio
ASCO	American Society of Clinical Oncology
cieBOM	Dutch Committee for the Evaluation of Oncological Drugs
ESMO	European Society for Medical Oncology
GHD	Generalised hazard difference
HR	Hazard ratio
IO	Immuno-oncology
KM	Kaplan-Meier
LRT	Log-rank test
MCBS	Magnitude of Clinical Benefit Scale
OS	Overall survival
PFS	Progression-free survival
PH	Proportional hazards
RCT	Randomised clinical trial
RMST	Restricted mean survival time
RSB	Restricted survival benefit
SLR	Systematic literature review
TLR	Targeted literature search
WKM	Weighted Kaplan-Meier
WLRT	Weighted log-rank test

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02373-5.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Acknowledgements

The authors would like to thank Ilse van Oostrum for the feedback provided throughout the process of writing this paper.

Author contributions

Isaac Corro Ramos had a major contribution in the design of the study, conducting the systematic literature review and writing the manuscript. Isaac Corro Ramos read and approved the final version of the manuscript.Venetia Qendri played a significant role in shaping the study design, conducting a systematic review of the existing literature, and authoring the manuscript. Venetia Qendri read and gave approval to the final version of the manuscript. Maiwenn Al substantially contributed to the study design, the systematic literature review and writing the manuscript. Maiwenn Al reviewed and provided approval for the final version of the manuscript.

Funding

This study was financially supported by AstraZeneca Netherlands. The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability

Search strings available at: https://doi.org/10.1079/searchRxiv.2023.00221 . https://doi.org/10.1079/searchRxiv.2023.00222.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Institute for Medical Technology Assessment (iMTA), Erasmus University Rotterdam, Burgemeester Oudlaan 50, Rotterdam 3062 PA, The Netherlands
[2]Erasmus School of Health Policy & Management (ESHPM), Erasmus University Rotterdam, Rotterdam, the Netherlands

Received: 17 June 2023 / Accepted: 16 October 2024
Published online: 30 October 2024

References

1. Esfahani K, Roudaia L, Buhlaiga N, Del Rincon SV, Papneja N, Miller WH. A review of Cancer Immunotherapy: from the past, to the Present, to the future. Curr Oncol. 2020;27(s2):87–97.
2. Chen TT. Statistical issues and challenges in immuno-oncology. J Immunother Cancer. 2013;1(1):18.
3. Mick R, Chen TT. Statistical challenges in the design of late-stage Cancer Immunotherapy studies. Cancer Immunol Res. 2015;3(12):1292–8.
4. Huang B. Some statistical considerations in the clinical development of cancer immunotherapies. Pharm Stat. 2018;17(1):49–60.
5. Rahmadian AP, Delos Santos S, Parshad S, Everest L, Cheung MC, Chan KK. Quantifying the Survival benefits of oncology drugs with a focus on Immunotherapy using restricted Mean Survival Time. J Natl Compr Cancer Netw JNCCN. 2020;18(3):278–85.
6. Bellmunt J, de Wit R, Vaughn DJ, Fradet Y, Lee JL, Fong L, et al. Pembrolizumab as Second-Line Therapy for Advanced Urothelial Carcinoma. N Engl J Med. 2017;376(11):1015–26.
7. 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(2):123–35.

8. Fehrenbacher L, Spira A, Ballinger M, Kowanetz M, Vansteenkiste J, Mazieres J, et al. Atezolizumab versus Docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. Lancet. 2016;387(10030):1837–46.
9. Herbst RS, Baas P, Kim DW, Felip E, Pérez-Gracia JL, Han JY, et al. Pembrolizumab versus Docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. Lancet. 2016;387(10027):1540–50.
10. Ferris RL, Blumenschein G, Fayette J, Guigay J, Colevas AD, Licitra L, et al. Nivolumab for Recurrent Squamous-Cell Carcinoma of the Head and Neck. N Engl J Med. 2016;375(19):1856–67.
11. Van Cutsem E, Köhne CH, Hitre E, Zaluski J, Chang Chien CR, Makhson A, et al. Cetuximab and Chemotherapy as initial treatment for metastatic colorectal Cancer. N Engl J Med. 2009;360(14):1408–17.
12. Demetri GD, von Mehren M, Jones RL, Hensley ML, Schuetze SM, Staddon A, et al. Efficacy and safety of Trabectedin or Dacarbazine for Metastatic Liposarcoma or Leiomyosarcoma after failure of conventional chemotherapy: results of a phase III randomized Multicenter Clinical Trial. J Clin Oncol. 2016;34(8):786–93.
13. Cicènas S, Geater SL, Petrov P, Hotko Y, Hooper G, Xia F, et al. Maintenance erlotinib versus erlotinib at disease progression in patients with advanced non-small-cell lung cancer who have not progressed following platinum-based chemotherapy (IUNO study). Lung Cancer. 2016;102:30–7.
14. Caplin ME, Pavel M, Ćwikła JB, Phan AT, Raderer M, Sedláčková E, et al. Lanreotide in metastatic enteropancreatic neuroendocrine tumors. N Engl J Med. 2014;371(3):224–33.
15. Attal M, Lauwers-Cances V, Marit G, Caillot D, Moreau P, Facon T, et al. Lenalidomide Maintenance after stem-cell transplantation for multiple myeloma. N Engl J Med. 2012;366(19):1782–91.
16. Eisenberger M, Hardy-Bessard AC, Kim CS, Géczi L, Ford D, Mourey L, et al. Phase III study comparing a Reduced Dose of Cabazitaxel (20 mg/m2) and the currently approved dose (25 mg/m2) in Postdocetaxel patients with metastatic castration-resistant prostate Cancer—PROSELICA. J Clin Oncol. 2017;35(28):3198–206.
17. 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(22):2380–5.
18. Nederlandse Vereniging voor Medische Oncologie (NVMO). NVMO. [cited 2023 Jun 6]. PASKWIL-criteria 2023. https://www.nvmo.org/over-de-adviezen/
19. Nederlandse Vereniging voor Medische Oncologie (NVMO). NVMO. [cited 2023 Mar 1]. Het beoordelen van subgroepen conform de PASKWIL-criteria van enkele eerder gepubliceerde rapporten. https://medischeoncologie.nl/artikelen/2018/februari/editie-1/beoordelenvansubgroepenconformpaskwilcriteriavanenkeleeerdergepubliceerderapporten
20. Badampudi D, Wohlin C, Petersen K. Experiences from using snowballing and database searches in systematic literature studies. In: Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering [Internet]. New York, NY, USA: Association for Computing Machinery; 2015 [cited 2024 Mar 5]. pp. 1–10. (EASE '15). https://doi.org/10.1145/2745802.2745818
21. Bansal A, Sullivan SD, Lin VW, Purdum AG, Navale L, Cheng P, et al. Estimating long-term survival for patients with relapsed or refractory large B-Cell lymphoma treated with Chimeric Antigen Receptor Therapy: a comparison of Standard and Mixture Cure models. Med Decis Mak. 2019;39(3):294–8.
22. Bullement A, Latimer NR, Bell Gorrod H. Survival extrapolation in Cancer Immunotherapy: a validation-based case study. Value Health. 2019;22(3):276–83.
23. Chen TT. Milestone survival: a potential Intermediate Endpoint for Immune Checkpoint inhibitors. JNCI J Natl Cancer Inst. 2015;107(9):djv156.
24. Chu C, Liu S, Rong A. Study design of single-arm phase II immunotherapy trials with long-term survivors and random delayed treatment effect. Pharm Stat. 2020;19(4):358–69.
25. Connock M, Armoiry X, Tsertsvadze A, Melendez-Torres GJ, Royle P, Andronis L, et al. Comparative survival benefit of currently licensed second or third line treatments for epidermal growth factor receptor (EGFR) and anaplastic lymphoma kinase (ALK) negative advanced or metastatic non-small cell lung cancer: a systematic review and secondary analysis of trials. BMC Cancer. 2019;19(1):392.
26. Ding X, Wu J. Designing cancer immunotherapy trials with delayed treatment effect using maximin efficiency robust statistics. Pharm Stat. 2020;19(4):424–35.
27. Ding X, Wu J. Cancer immunotherapy trial design with long-term survivors. Pharm Stat. 2021;20(1):117–28.
28. Grant TS, Burns D, Kiff C, Lee D. A Case Study examining the usefulness of cure modelling for the prediction of Survival based on data Maturity. PharmacoEconomics. 2020;38(4):385–95.
29. Liu S, Chu C, Rong A. Weighted log-rank test for time-to-event data in immunotherapy trials with random delayed treatment effect and cure rate. Pharm Stat. 2018;17(5):541–54.
30. Mukhopadhyay P, Huang W, Metcalfe P, Öhrn F, Jenner M, Stone A. Statistical and practical considerations in designing of immuno-oncology trials. J Biopharm Stat. 2020;30(6):1130–46.
31. 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(12):1692–6.
32. Vadgama S, Mann J, Bashir Z, Spooner C, Collins GP, Bullement A. Predicting Survival for Chimeric Antigen Receptor T-Cell therapy: a validation of Survival models using Follow-Up data from ZUMA-1. Value Health. 2022;25(6):1010–7.
33. Wang ZX, Wu HX, Xie L, Lin WH, Liang F, Li J, et al. Exploration of modified progression-free survival as a novel surrogate endpoint for overall survival in immuno-oncology trials. J Immunother Cancer. 2021;9(4):e002114.
34. Wu J, Wei J. Cancer Immunotherapy Trial Design with Random delayed treatment effect and cure rate. Stat Med. 2022;41(4):786–97.
35. Xu Z, Zhu B, Park Y. Design for immuno-oncology clinical trials enrolling both responders and nonresponders. Stat Med. 2020;39(27):3914–36.
36. Yu C, Huang X, Nian H, He P. A weighted log-rank test and associated effect estimator for cancer trials with delayed treatment effect. Pharm Stat. 2021;20(3):528–50.
37. Gibson E, Koblbauer I, Begum N, Dranitsaris G, Liew D, McEwan P, et al. Modelling the survival outcomes of Immuno-Oncology drugs in economic evaluations: a systematic Approach to Data Analysis and Extrapolation. PharmacoEconomics. 2017;35(12):1257–70.
38. Hasegawa T. Sample size determination for the weighted log-rank test with the Fleming–Harrington class of weights in cancer vaccine studies. Pharm Stat. 2014;13(2):128–35.
39. Huang X, Lyu J, Hou Y, Chen Z. A nonparametric statistical method for two crossing survival curves. Commun Stat - Simul Comput. 2022;51(9):5041–50.
40. Magirr D, Burman CF. Modestly weighted logrank tests. Stat Med. 2019;38(20):3782–90.
41. Ouwens MJNM, Mukhopadhyay P, Zhang Y, Huang M, Latimer N, Briggs A. Estimating lifetime benefits Associated with Immuno-Oncology therapies: challenges and approaches for overall survival extrapolations. PharmacoEconomics. 2019;37(9):1129–38.
42. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: a class of Distance tests for censored Survival Data. Biometrics. 1989;45(2):497–507.
43. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21(15):2175–97.
44. 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(1):16.
45. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol. 2013;13(1):152.
46. 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(15):1813–9.
47. Uno H, Wittes J, Fu H, Solomon SD, Claggett B, Tian L, et al. Alternatives to Hazard Ratios for comparing the efficacy or safety of therapies in Noninferiority studies. Ann Intern Med. 2015;163(2):127–34.
48. 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(28):3680–95.
49. Xu Z, Zhen B, Park Y, Zhu B. Designing therapeutic cancer vaccine trials with delayed treatment effect. Stat Med. 2017;36(4):592–605.
50. Xu Z, Park Y, Zhen B, Zhu B. Designing cancer immunotherapy trials with random treatment time-lag effect. Stat Med. 2018;37(30):4589–609.
51. Zhang D, Quan H. Power and sample size calculation for log-rank test with a Time lag in treatment effect. Stat Med. 2009;28(5):864–79.

52. Zucker DM, Lakatos E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. Biometrika. 1990;77(4):853–64.
53. Jiménez JL. Quantifying treatment differences in confirmatory trials under non-proportional hazards. J Appl Stat. 2022;49(2):466–84.
54. Lin RS, Lin J, Roychoudhury S, Anderson KM, Hu T, Huang B, et al. Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis. Stat Biopharm Res. 2020;12(2):187–98.
55. Rauch G, Brannath W, Brückner M, Kieser M. The average hazard ratio – a good Effect measure for time-to-event endpoints when the Proportional Hazard Assumption is violated? Methods Inf Med. 2018;57(03):089–100.
56. Roychoudhury S, Anderson KM, Ye J, Mukhopadhyay P. Robust design and analysis of clinical trials with nonproportional hazards: a Straw Man Guidance from a Cross-pharma Working Group. Stat Biopharm Res. 2021;0(0):1–15.
57. Snapinn S, Jiang Q, Ke C. Treatment effect measures under nonproportional hazards. Pharm Stat. 2023;22(1):181–93.
58. He P, Fang L, Su Z. A sequential testing approach to detecting multiple change points in the proportional hazards model. Stat Med. 2013;32(7):1239–45.
59. Huang B, Kuan PF. Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point: comparison of the RMST with the HR. Pharm Stat. 2018;17(3):202–13.
60. Yang S, Prentice R. Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. Biometrika. 2005;92(1):1–17.
61. Fine GD. Consequences of delayed Treatment effects on Analysis of Time-to-event endpoints. Drug Inf J. 2007;41(4):535–9.
62. Fleming TR, Harrington DP. A class of hypothesis tests for one and two sample censored survival data. Commun Stat - Theory Methods. 1981;10(8):763–94.
63. Hellmann MD, Kris MG, Rudin CM. Medians and milestones in describing the path to Cancer cures: telling tails. JAMA Oncol. 2016;2(2):167–8.
64. Lin RS, León LF. Estimation of treatment effects in weighted log-rank tests. Contemp Clin Trials Commun. 2017;8:147–55.
65. Péron J, Roy P, Ozenne B, Roche L, Buyse M. The net chance of a longer survival as a patient-oriented measure of Treatment Benefit in Randomized clinical trials. JAMA Oncol. 2016;2(7):901.
66. Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Stat Med. 2011;30(19):2409–21.
67. Saad ED, Zalcberg JR, Péron J, Coart E, Burzykowski T, Buyse M. Understanding and communicating measures of treatment effect on Survival: can we do better? JNCI J Natl Cancer Inst. 2018;110(3):232–40.
68. Shen Y, Cai J. Maximum of the Weighted Kaplan-Meier tests with application to Cancer Prevention and Screening trials. Biometrics. 2001;57(3):837–43.
69. Struthers CA, Kalbfleisch JD. Misspecified Proportional Hazard models. Biometrika. 1986;73(2):363–9.
70. Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. Biostat Oxf Engl. 2014;15(2):222–33.
71. Wei J, Wu J. Cancer Immunotherapy Trial Design with cure rate and delayed treatment effect. Stat Med. 2020;39(6):698–708.
72. Yang S, Prentice R. Improved Logrank-Type tests for Survival Data using adaptive weights. Biometrics. 2010;66(1):30–8.
73. Ye T, Yu M. A robust approach to sample size calculation in cancer immunotherapy trials with delayed treatment effect. Biometrics. 2018;74(4):1292–300.
74. 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(1):215–21.
75. Zhang S, LeBlanc ML, Zhao YQ. Restricted survival benefit with right-censored data. Biom J Biom Z. 2022;64(4):696–713.
76. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ Syst Rev 2021;10:89. https://doi.org/10.1186/s13643-021-01626-4.
77. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. *PRISMA*2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Syst Rev. 2022;18(2):e1230.
78. Kalbfleisch JD, Prentice RL. Estimation of the average hazard ratio. Biometrika. 1981;68(1):105–12.
79. Rauch G, Brannath W, Brückner M, Kieser M. The Average Hazard Ratio – A Good Effect Measure for Time-to-event Endpoints when the Proportional Hazard Assumption is Violated? Methods Inf Med [Internet]. 2018 May [cited 2023 Jan 31];57(03):089–100. http://www.thieme-connect.de/DOI/DOI?10.3414/ME17-01-0058
80. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted Cox regression. Stat Med. 2009;28(19):2473–89.
81. Dunkler D, Heinze G. Meinhard Ploner. R package coxphw: Weighted Estimation in Cox Regression [Internet]. 2020 [cited 2023 Feb 2]. https://cran.r-project.org/web/packages/coxphw/coxphw.pdf
82. Matthias Brueckner. R package AHR: Estimation and Testing of Average Hazard Ratios [Internet]. 2016 [cited 2023 Feb 2]. http://cran.nexr.com/web/packages/AHR/AHR.pdf
83. Therneau T, Crowson C, Atkinson E. Using Time Dependent covariates and Time dependent coefficients in the Cox Model. Surviv Vignettes. 2017;2(3):1–25.
84. Zhang Z, Reinikainen J, Adeleke KA, Pieterse ME, Groothuis-Oudshoorn CGM. Time-varying covariates and coefficients in Cox regression models. Ann Transl Med. 2018;6(7):121.
85. Hajime Uno L, Tian M, Horiguchi A, Cronin C, Battioui J, Bell. R package survRM2: Comparing Restricted Mean Survival Time [Internet]. 2022 [cited 2023 Jan 31]. https://cran.r-project.org/web/packages/survRM2/survRM2.pdf
86. Terry M, Therneau T, Lumley A, Elizabeth C, Cynthia. R package survival: Survival Analysis [Internet]. 2023 [cited 2023 Jan 31]. https://cran.r-project.org/web/packages/survival/survival.pdf
87. Jackson C, Metcalfe P, Amdahl,Matthew J, Warkentin T, Sweeting M, Kunzmann K. R package flexsurv: Flexible Parametric Survival and Multi-State Models [Internet]. 2022 [cited 2023 Jan 31]. https://cran.r-project.org/web/packages/flexsurv/flexsurv.pdf
88. Bartlett JW, Morris TP, Stensrud MJ, Daniel RM, Vansteelandt SK, Burman CF. The hazards of Period Specific and Weighted Hazard Ratios. Stat Biopharm Res. 2020;12(4):518–9.
89. Klein JP, Logan B, Harhoff M, Andersen PK. Analyzing survival curves at a fixed point in time. Stat Med. 2007;26(24):4505–19.
90. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. J Clin Oncol. 1983;1(11):710–9.
91. Hernán MA. The hazards of Hazard Ratios. Epidemiology. 2010;21(1):13.
92. Freidlin B, Korn EL. Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis? J Clin Oncol. 2019;37(35):3455–9.
93. Cherny NI, Dafni U, Bogaerts J, Latino NJ, Pentheroudakis G, Douillard JY, et al. ESMO-Magnitude of Clinical Benefit Scale version 1.1. Ann Oncol. 2017;28(10):2340–66.
94. Schnipper LE, Davidson NE, Wollins DS, Blayney DW, Dicker AP, Ganz PA, et al. Updating the American Society of Clinical Oncology Value Framework: revisions and reflections in response to comments received. J Clin Oncol. 2016;34(24):2925–34.

Publisher's note