Summary measures in non-inferiority clinical trials with a time-to-event outcome: an empirical comparison of power

Broer, S.D.L.; White, I.R.; Morris, T.P.; Weir, I.R.; Fiocco, M.; Quartagno, M.

BMC Medical Research Methodology

RESEARCH

Open Access

Summary measures in non-inferiority clinical trials with a time-to-event outcome: an empirical comparison of power

S. D. Lana Broer[1], Ian R. White[2], Tim P. Morris[2], Isabelle R. Weir[3], Marta Fiocco[4,5] and Matteo Quartagno[2*]

Abstract

Background Time-to-event data is commonly used in non-inferiority clinical trials. While the hazard ratio is a popular summary measure in this context, the difference in restricted mean survival time has been theoretically shown to increase power and interpretability. This study aimed to empirically compare the power of the hazard ratio, difference in survival and difference in restricted mean survival time for non-inferiority clinical trials with a time-to-event outcome recently published in key clinical journals.

Methods Sixty-five non-inferiority trials with a time-to-event outcome were included from two literature searches. Individual patient data were reconstructed and reanalysed. The hazard ratio, difference in survival and difference in restricted mean survival time were estimated under proportional hazards, using a Cox model for the hazard ratio and a flexible parametric survival model for the latter two summary measures. The latter measures were additionally estimated non-parametrically. Margin conversion was done using observed data in the control arm. Empirical power was defined as the proportion of trials that rejected the null hypothesis.

Results Difference in restricted mean survival time gave a potential power advantage over the hazard ratio with an empirical power increase of 7.7 (−5.4, 20.7) percentage points, and consistently outperformed difference in survival. Difference in survival was more powerful than the hazard ratio, but while difference in restricted mean survival time showed an empirical power advantage even when estimated non-parametrically, this was not generally true for difference in survival. Sub-group analyses consistently showed similar results. Results were more variable when converting margins under an exponential distribution, highlighting the importance of correct margin conversion.

Conclusion Our results empirically corroborate the theoretical advantage of difference in restricted mean survival time over the hazard ratio and difference in survival in non-inferiority clinical trials. This advantage is most apparent when estimation is done under proportional hazards. Choosing a relevant time point at which to evaluate the survival-based summary measures is an important aspect that should be carefully considered. We recommend incorporation of the difference in restricted mean survival time in the design and analysis of non-inferiority clinical trials when clinically justifiable. If appropriate, estimation under proportional hazards is preferable.

Keywords Power, Non-inferiority, Time-to-event data, Proportional hazards, Hazard ratio, Difference in survival, Restricted mean survival time

*Correspondence:
Matteo Quartagno
m.quartagno@ucl.ac.uk
Full list of author information is available at the end of the article

Background

Non-inferiority clinical trials are increasingly used to evaluate whether the efficacy of a new treatment is maintained as compared to the standard of care. New treatments may be preferred because of beneficial qualities not directly related to the primary outcome, such as fewer side effects or cheaper production [1, 2]. Non-inferiority trials often focus on time-to-event outcomes, such as overall or progression-free survival, with applications coming mostly from oncology and cardiovascular disease [3, 4]. The hazard ratio (HR), expressing the relative difference in hazard, is popular as a measure of how treatment effects may differ between groups. However, because the HR is a relative measure, it has been suggested that an absolute measure of survival should be additionally reported for clinical interpretability [5, 6]. This is supported by the fact that the unitless HR, especially when assuming this ratio to be constant over time as in the case of proportional hazards, provides no direct information on how the survival time is affected. Therefore, the usefulness of the HR for clinical decision making has been a topic of discussion [7, 8]. Moreover, estimation of the HR is often done under the strong assumption of proportional hazards, imposing a constant HR over time. Although time-varying HRs are possible, they are rarely implemented and further complicate interpretation. When the true hazards are non-proportional, the average HR imposed by assuming proportional hazards will be influenced by follow-up time. This raises the point that estimation of the HR is highly influenced by the (reported) follow-up time, highlighting that while the estimated HR may hold up until a certain time point, there is no guarantee it is truly generalisable [9].

Two alternative population-level summary measures for time-to-event outcomes are the absolute risk difference or difference in survival (DS) and the difference in restricted mean survival time (DRMST) [10]. The DS evaluates the survival in both treatment arms at a pre-specified time τ and takes the difference of the two estimates, thereby reporting only the survival at time τ without providing insight on survival information before and after. The restricted mean survival time equals the area under the survival curve until time τ, incorporating the behaviour of the survival curves until this time. It describes the expected event-free time for a patient followed until time τ [11]. As such, the DRMST, which is the difference between the restricted mean survival times in the two treatment arms, quantifies the absolute expected difference in survival time over a fixed time horizon and has an extra advantage over DS by making full usage of the survival information until the pre-specified time. An advantage of both the DS and DRMST is that they are absolute measures of survival and are directly clinically interpretable [6]. Additionally, the proportional hazards assumption is not a necessity in the estimation of these measures, although estimation under proportional hazards is possible [12].

For non-inferiority trials, beyond the aforementioned advantages, it has been suggested that these summary measures – and the DRMST in particular – can be more powerful and lead to smaller sample sizes than the HR [8, 10, 13]. This is due to the slightly different null hypotheses tested using different summary measures, as the non-inferiority margin can only be matched at the expected value of the distribution in the control arm [14]. In the study conducted by Weir and Trinquart [13], the DRMST was empirically shown to give a sample size advantage over the HR. In that study, DRMST was estimated using the Kaplan-Meier (KM) method, without enforcing proportional hazards. More recently, we showed through a simulation study that when the proportional hazards assumption holds, estimation of the DRMST under proportional hazards gives an additional power advantage [10].

The aim of this study was to understand whether the theoretical power advantage of the DRMST over the HR translates empirically, extending the results of [13] to estimation of the DRMST under proportional hazards. A literature search was conducted to identify non-inferiority clinical trials with a time-to-event outcome. The underlying individual patient data were reconstructed using the algorithm developed by Guyot et al. [15] and reanalysed to empirically compare the performance in power. To reanalyse the data, the non-inferiority margins were converted between measures using the observed data in the control arm to provide a fair comparison between the three measures. DS and DRMST were estimated both with a flexible parametric survival model under proportional hazards and using the non-parametric KM estimator; HR was estimated using the Cox proportional hazards model. The main performance measure to evaluate different methods was empirical power, defined as the proportion of 'successful' non-inferiority trials, i.e. trials for which the null hypothesis was rejected, for each summary measure and method.

Methods

Literature search

The literature search aimed to identify non-inferiority clinical trials with a primary time-to-event outcome. Inclusion criteria consisted of reporting of the non-inferiority margin, a KM curve for each trial arm, a risk table, and the expected event rate in the control group. Trials which were pooled, secondary or meta analyses, cluster-randomised or in a competing risk setting were excluded. We considered publications in The New England Journal

of Medicine, The Lancet, The British Medical Journal, Nature and JAMA, and their subsidiaries, published between January 1st 2021 and March 15th 2024. The search was conducted in PubMed. The complete literature search can be found in Appendix A. Articles were included chronologically, starting from 2024 and moving back in time, until the inclusion of 30 KM curves.

For each KM curve, the underlying individual patient data were reconstructed. Published KM curves were outlined using WebPlotDigitizer to obtain the corresponding coordinates [16]. The underlying trial data were then reconstructed using the algorithm developed by Guyot et al., which takes as input coordinates of the KM, as well as a predefined set of intervals with corresponding numbers at risk at the start of each interval. The total number of events were additionally included if available.

On top of the 30 reconstructed datasets as above, 35 datasets previously reconstructed by Weir and Trinquart were included in the analysis, leading to a total of 65 datasets.

Analysis

The summary measures of interest consisted of the HR, DS and DRMST. Let $S_C(t)$ and $S_A(t)$ denote the survival functions of the control and active arms, respectively, where t is any time point after baseline ($t_0 = 0$), and let $h_C(t)$ and $h_A(t)$ denote the hazard of the control and active arms, respectively. The HR then follows as [17]

$$HR(t) = \frac{h_A(t)}{h_C(t)},$$

where under proportional hazards the HR is constant over time, i.e. $HR(t) = HR$ for all $t > 0$. To define the DS and DRMST, choose a time $\tau > 0$ at which the summary measure is to be evaluated. The DS and DRMST at τ are then given by [7]

$$DS(\tau) = S_A(\tau) - S_C(\tau); \tag{1}$$

$$DRMST(\tau) = \int_0^\tau S_A(t)\,dt - \int_0^\tau S_C(t)\,dt. \tag{2}$$

For the primary analysis, proportional hazards were assumed for the estimation of all summary measures. The HR was estimated using a Cox proportional hazards regression model. The DRMST and the DS were estimated using a flexible parametric survival model under proportional hazards with two internal knots at the 33% and 67% quantiles, emulating what was done in Quartagno et al. [10]. Because the data were reconstructed without information on patient characteristics, no covariates were included in the analysis. The DRMST and DS were also estimated non-parametrically using the KM method, without any assumption on the proportionality of the hazards. The KM method was chosen over a flexible parametric non-proportional hazards model for fairer comparison with previous studies. Non-inferiority was tested against the non-inferiority margin of the corresponding summary measure by constructing the appropriate confidence interval, followed by a Z-test to acquire a p-value. As implemented in the dani package in R [18], confidence intervals for DS and DRMST under flexible parametric survival modelling were estimated using the delta method, Greenwood's formula was used to obtain confidence intervals for DS and DRMST under KM estimation, and standard Wald intervals were used for the HR. Significance was concluded based on the significance level of the accompanying trial. Each trial was tested for adherence to the proportional hazards assumption using the Grambsch–Therneau test against a significance level of 0.05.

Since non-inferiority margins are inherently tied to the summary measure they are defined on, it was necessary to convert the non-inferiority margin used in a given trial to a corresponding margin for each summary measure. To convert margins, it was assumed that the proportional hazards assumption held, so that

$$S_A^{MC}(t) = S_C^{MC}(t)^{HR},$$

leaving only $S_C^{MC}(t)$ to be estimated - here, the superscript ^{MC} is used to denote these survival functions to be those used for margin conversion. Assuming proportional hazards in this setting was deemed appropriate because it mimics how the margins would have been defined during trial design, therefore rendering non-adherence to proportional hazards irrelevant during margin conversion. Two approaches were used to estimate $S_C^{MC}(t)$: (i) an exponential distribution with an expected event rate as reported by the accompanying trial, and (ii) a flexible parametric regression model fit on the reconstructed data of the control arm using two internal knots placed at the 33% and 67% quantiles of the event times. To avoid misspecification of the distribution, the latter approach was deemed most appropriate and was used for the primary and subgroup analyses. After approximation of the control-arm survival function, i.e. $S_C^{MC}(t)$, and definition of τ, the non-inferiority margins were converted between each summary measure by plugging in the parameters in Eqs. (1) and (2), as appropriate, and solving for the last unknown parameter. We refer to Appendix J for further elaboration on margin conversion and worked out examples.

For the DS and DRMST, the choice of τ is of evident importance. In practice, it is recommended to choose τ based on clinical considerations (τ_{clin}). In the trials for

which the summary measure was chosen to be the HR, τ was not directly available. For these trials, τ was defined as the time at which the outcome was reported to be evaluated, if available – for example, if the outcome was 5-year progression-free survival, τ was set as 5 years. Otherwise, the maximum available follow-up time was used. For some trials, follow-up was reported for longer than the time at which the outcome was evaluated. In these cases, we additionally defined a maximal τ as the latest available follow-up time (τ_{max}) to explore how the choice of τ influences outcomes, with the expectation that reduction of τ increases DRMST power. For trials where follow-up was discontinued, or not reported, after the time at which the outcome was evaluated we have $\tau_{clin} = \tau_{max}$. In the conversion from a non-inferiority margin for the HR to a margin for the DS or DRMST, both τ_{clin} and τ_{max} were employed. The conversion from DS and DRMST margins to the HR margin considered only τ_{clin}, under the assumption that the margin was designed for this τ.

The empirical power was defined as the percentage of trials for which non-inferiority was concluded at the corresponding significance level. The summary measure with the highest percentage of trials concluding non-inferiority was deemed most empirically powerful. The difference in proportions was tested using the exact McNemar test; 95%-confidence intervals for the difference in proportions are reported by Wald intervals. Additional subgroup analyses were done for (i) trials without evidence of non-proportionality, by (ii) original summary measure, and by (iii) magnitude of the event risks. For the latter, we firstly estimated the median survival at τ_{clin} in the control arm for all trials, after which trials were categorised as a low/high event risk if the survival at τ_{clin} was above/below the median value across trials.

All analyses were performed in R, version 4.4.0 [19]. Flexible parametric survival models were fitted using flexsurv, version 2.3.2 [20]. The survival package, version 3.7.0, was used for KM and Cox proportional hazards model estimation [21]. Non-inferiority testing was done using the dani package, available from GitHub [18]. The reconstructed data and R code concerning the analysis can be found on GitHub [22]. While we pre-planned our analyses internally, because of the exploratory nature of this work we did not feel it was necessary to publish the pre-specified analysis plan.

Results

The literature search resulted in 106 published articles between January 1st 2021 and March 15th 2024; 69 manuscripts were considered before inclusion was concluded at 29 trials with 30 suitable time-to-event outcomes that had a published KM curve and corresponding non-inferiority margin (see Appendix D for the reconstructed curves). After addition of the 35 trials identified by Weir and Trinquart, a total of 65 datasets were considered for analysis. Characteristics of the trials can be found in Table 1; all 65 outcomes analysed were unfavourable.

Non-inferiority margins were defined equally often for the HR ($n = 15$) as for the DS ($n = 14$) in the newly identified trials, with only one trial defining a non-inferiority margin for the DRMST. This is in contrast with the distribution seen in the 35 trials identified by Weir and Trinquart, where the HR was more commonly used ($n = 25$ for the HR, $n = 10$ for the DS). Reconstructed summary measures for the 30 newly included trials can be found in Appendix F; we refer to Weir and Trinquart for more in-depth information on the 35 previously digitised trials. Converted margins for all 65 trials can be found in Appendix G. Margin conversion using flexible parametric regression did not converge for the NPC-CTVn trial when using the pre-defined quantiles, therefore the internal knots were manually placed at 2.7 and 3.2 months to achieve convergence. Evidence of non-proportionality was found for eight of the 65 trials.

The main analysis focused on the DRMST and DS estimated with a flexible parametric survival model under proportional hazards where the margin conversion was done with τ_{clin}, and showed a power advantage of the DRMST over both the DS and the HR (Table 2). An empirical power increase of 7.7 percentage points was observed between the DRMST and HR ($p = 0.06$), although the confidence interval included the null (95% -confidence interval: $-5.4, 20.7$). This advantage of the DRMST was maintained when the margin was converted using τ_{max} and under non-parametric estimation. A similar empirical power advantage over HR was observed for the DS, although non-parametric estimation of DS was found to be comparable to the HR and variable under the choice of τ.

Three (4.6%) trials presented a different conclusion between the HR and DS (Tang2022, ACTI, SIOP WT 2001). Two (3.08%) additional trials had disagreeing conclusions between the HR and DRMST (Oral (a), MERTH). Evidence of non-proportionality was found for one of these five trials (ACTI). These outcomes, which were observed when estimation was done under proportional hazards, held true regardless of which τ was used. Further investigation into the non-parametric estimation showed more discordance than under proportional hazards.

A negative relationship was found between the power and average p-values, such that, as expected, the trials with the highest empirical power had the lowest average p-values. Figure 1 shows the relationship between the p-values estimated using the different summary

Table 1 Characteristics of the included trials

	New included trials (*n* = 29)	**Trials included by [13] (*n* = 35)**
Journal		
NEJM	12 (41.4)	24 (68.6)
JAMA	9 (31.0)	4 (11.4)
Lancet	6 (20.7)	7 (20.0)
BMJ	1 (3.4)	0 (0.0)
Nature Medicine	1 (3.4)	0 (0.0)
Clinical application		
Oncology	14 (48.3)	10 (28.6)
Cardiovascular disease	9 (31.0)	12 (34.3)
Transplantation	2 (6.9)	0 (0.0)
Diabetes	1 (3.4)	4 (11.4)
Infectious disease	1 (3.4)	1 (2.9)
Rheumatoid Arthritis	1 (3.4)	0 (0.0)
Surgery	1 (3.4)	0 (0.0)
Asthma or COPD	0 (0.0)	6 (17.1)
HIV	0 (0.0)	1 (2.9)
Obesity	0 (0.0)	1 (2.9)
Analysis		
ITT	22 (75.9)	26 (74.0)
PP	4 (13.8)	4 (11.0)
mITT	2 (6.9)	5 (14.0)
Not reported	1 (3.4)	0 (0.0)
Randomization ratio		
1 : 1	25 (86.2)	33 (94.3)
1 : 2	3 (10.3)	1 (2.9)
1 : 3	1 (3.4)	1 (2.9)
Sample size (median (IQR))	913 (571, 1525)	1905 (733, 3330)
	Included curves (*n* = 30[a])	**Included curves by [13] (*n* = 35)**
Original margin		
HR	15 (50.0)	25 (71.4)
DS	14 (46.7)	10 (28.6)
DRMST	1 (3.3)	0 (0.0)
Concluded non-inferiority		
No	7 (23.3)	6 (17.1)
Yes	20 (66.7)	27 (77.1)
Yes, concluded superiority	3 (10.0)	2 (5.7)
Event risk (median (IQR))	0.88 (0.74, 0.94)	0.87 (0.71, 0.94)

[a] One trial reported two Kaplan-Meier curves with a non-inferiority margin

Statistics are presented as *n* (%) unless otherwise indicated

Abbreviations: *NEJM* New England Journal of Medicine, *JAMA* Journal of the American Medical Association, *BMJ* British Medical Journal, *COPD* chronic obstructive pulmonary disease, *HR* hazard ratio, *DS* difference in survival, *DRMST* difference in restricted mean survival time, *(m)ITT* (modified) intention to treat, *IQR* interquartile range

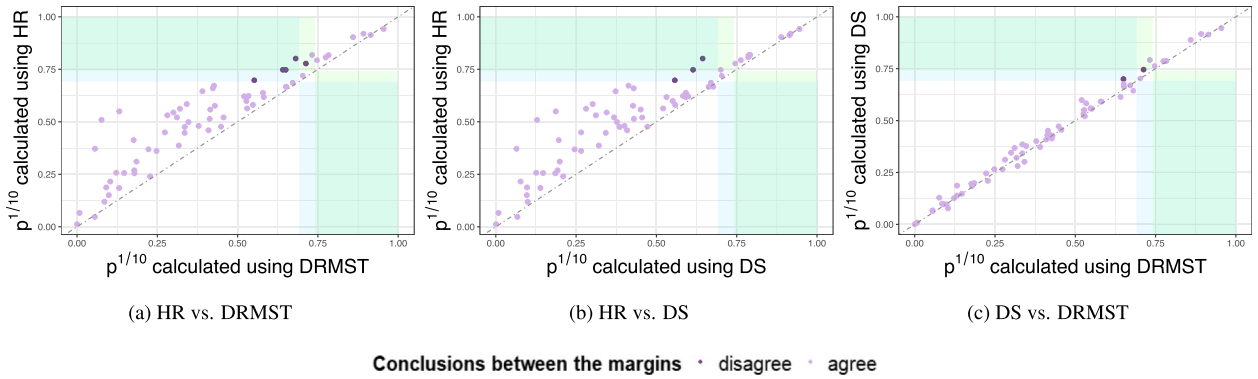
measures, with the dark coloured points indicating trials for which the summary measures were in disagreement. The green rectangles show the area in which the two margins would disagree for a significance level of 0.05, the blue rectangles indicate the area for which the margins would disagree with a significance level of 0.025. *P*-values were transformed to $p^{1/10}$ for visualisation purposes. The figure shows consistently lower *p*-values for the DRMST and DS, compared to the HR. All discrepancies between the summary measures disfavour the HR.

Table 2 Primary outcomes using margin conversion under flexible parametric regression

Summary measure	Model	τ	Empirical power	*P*-value characteristics			
				mean	*median*	*minimum*	*maximum*
DRMST	flexsurv (PH)	clinical	0.862	0.031	$1.33 \cdot 10^{-4}$	$2.33 \cdot 10^{-36}$	$6.30 \cdot 10^{-1}$
DRMST	flexsurv (PH)	maximum	0.862	0.031	$1.45 \cdot 10^{-4}$	$2.33 \cdot 10^{-36}$	$6.30 \cdot 10^{-1}$
DRMST	KM	maximum	0.862	0.032	$2.77 \cdot 10^{-4}$	$2.34 \cdot 10^{-31}$	$8.56 \cdot 10^{-1}$
DRMST	KM	clinical	0.846	0.033	$2.77 \cdot 10^{-4}$	$2.34 \cdot 10^{-31}$	$8.56 \cdot 10^{-1}$
DS	flexsurv (PH)	clinical	0.831	0.035	$1.30 \cdot 10^{-4}$	$2.22 \cdot 10^{-33}$	$5.70 \cdot 10^{-1}$
DS	flexsurv (PH)	maximum	0.831	0.037	$2.13 \cdot 10^{-4}$	$2.22 \cdot 10^{-33}$	$5.70 \cdot 10^{-1}$
HR	Cox	-	0.785	0.043	$2.51 \cdot 10^{-3}$	$6.05 \cdot 10^{-20}$	$5.51 \cdot 10^{-1}$
DS	KM	clinical	0.785	0.059	$1.38 \cdot 10^{-4}$	$9.85 \cdot 10^{-33}$	$6.44 \cdot 10^{-1}$
DS	KM	maximum	0.754	0.063	$7.21 \cdot 10^{-4}$	$9.85 \cdot 10^{-33}$	$6.44 \cdot 10^{-1}$

Standard errors for the empirical power ranged between 0.043 and 0.053

Abbreviations: *HR* hazard ratio, *DS* difference in survival, *DRMST* difference in restricted mean survival time, *PH* proportional hazards, *KM* Kaplan-Meier



(a) HR vs. DRMST (b) HR vs. DS (c) DS vs. DRMST

Fig. 1 *P*-values by summary measure, using a Box–Cox transformation [23] with $\lambda = \frac{1}{10}$

Margin conversion under the exponential distribution provided less strong evidence for a power gain when using DRMST; the DS estimated under proportional hazards outperformed the DRMST by one trial and came out as most powerful (Table 10, Appendix H). Similarly to the conversion using flexible parametric regression, the DRMST was robust between estimation methods and consistently had higher power than the HR. The DRMST regardless of estimation method, and DS estimated using the flexible parametric survival model performed comparably. Overall, a less clear pattern of behaviour in the relative performance of the summary measures was observed in this scenario. Further analysis of the *p*-values, as can be found in Fig. 6 in Appendix H, showed disagreement between measures in both directions for all three pairs of summary measures, with a generally wider spread of *p*-values, although lower *p*-values remained for the DRMST and DS as compared to the HR. Contradicting this, the average *p*-value was lowest for the HR, regardless of it having the lowest empirical power.

All subgroup analyses provided evidence of an increased empirical power when using the DRMST under proportional hazards (Table 10, Appendix H). The subgroup analysis in which trials with evidence of non-proportionality were excluded gave directly analogous results to the main analysis; this also holds for the low event risk group and the group which had an original non-inferiority margin on the HR scale, although an absolute increase in the power of the HR was observable. In case of a high event risk, all summary measures performed similarly: all five estimations performed under proportional hazards had identical power. The median (interquartile range) survival at τ_{clin} in the control group, which was used as the cut-off point between the low and high event risk groups, was 0.88 (interquartile range: 0.74, 0.94). Trials for which the non-inferiority margin was originally defined as a DS performed the worst on the HR scale, but equivalently for the DRMST and DS,

regardless of the estimation method. As in the primary analysis, while empirical power increases were observed in the subgroup analyses, 95%-confidence intervals around the differences in power all included the null.

Discussion

We found the DRMST estimated using a flexible parametric survival model under proportional hazards to be the most powerful summary measure for non-inferiority clinical trials with a time-to-event outcome. This remained consistent whether τ was defined clinically or based on the maximum follow-up time. Non-parametric estimation of the DRMST at τ_{max} was equivalent to flexible parametric estimation of DRMST in all analyses. Using τ_{clin} for the KM estimation of the DRMST reduced empirical power in most analyses, indicating that, for non-parametric estimation, using the longest available follow-up time is beneficial. Moreover, we found the estimation of DS under proportional hazards to outperform its non-parametrically estimated counterpart in all scenarios. The former performed slightly worse than the DRMST overall. Additionally, the HR never outperformed the DS under proportional hazards, although it frequently outperformed the non-parametrically estimated DS. Thereby, the DS was observed to benefit strongly from the proportional hazards framework. The HR was outperformed by the DRMST in all scenarios, with exception of the high event risk group, where the empirical power was identical.

Our estimates of empirical power when the margin was converted using the exponential distribution were less conclusive about the difference in performance between DS and DRMST, but showed that both outperformed the HR regardless of the estimation method. We hypothesise that the differences in empirical power between the two conversion methods are attributable to the trials for which the expected event rate deviated from the observed event rate. For these trials, the exponential distributions were an inaccurate approximation of the observed data in the control arm (see Appendix K). This leads to an incompatible conversion of the non-inferiority margin. The variability observed when using the exponential distribution as compared to the flexible parametric model for the margin conversion highlights the importance of correct margin matching between summary measures. Theoretically, using the expected event rate rather than the observed is preferable, but our simulations (yet to be published) show that incorrect specification of the expected survival curve leads to large reductions in power or inflation of type I error rate. Therefore, we gave preference to estimation of the survival curve using the observed data. While it has been shown that converting the margin based on observed data can inflate error rates with binary data [24], after exploratory simulations, we expect this choice to have little effect on the type I error rate in these settings.

The definition of the non-inferiority margin is an important aspect in both the comparison between summary measures and the design of non-inferiority trials [25]. Because the non-inferiority margin is the minimum non-acceptable boundary of the (1-α)·100%-confidence interval, where α is the significance level, the margin is inherently tied to the chosen summary measure, meaning that margins need to be correctly matched for a fair comparison. Incorrect conversion of a margin between measures could give a false benefit to one measure as a mere consequence of the margins being incomparable [26]. As such, accurate translation of one margin to another is of vital importance for a power comparison between summary measures. Tables 6 and 7 in Appendix suggest, though, that margins are on average different depending on if the original scale was the HR or DS. For example, the average non-inferiority margin when the original scale was HR (Table 6 in Appendix) is approximately 1.5, with an average DS of around five percentage points. However, when the original scale was DS (Table 7 in Appendix), we find an average margin of seven percentage points, with an average HR of 2.

The importance of properly defining the non-inferiority margin is further highlighted in guidelines published by the Food and Drug Administration (FDA) and European Medicines Agency (EMA) [27, 28]. FDA's 2016 guideline outlines key considerations in the design and analysis of non-inferiority trials. They stress that margin choices define the validity of the conclusions drawn from non-inferiority trials. They additionally emphasize that margins should be defined during the design phase based on prior knowledge. EMA's 2004 guideline, focused specifically on the non-inferiority margin, similarly remarks how statistical reasoning and clinical judgement should be the basis for margin definition. Notably, neither guideline addresses how to appropriately choose a summary measure. The FDA discusses measures for both absolute and relative risk differences, but focuses mainly on the hazard ratio with respect to time-to-event data, while the EMA guideline does not discuss summary measures.

While for the purpose of this study, it was important to convert the margins correctly, in order to make the comparisons as fair as possible, in reality when designing non-inferiority trials, margins should be designed on the original scale following clinical guidance, such that it has a meaningful interpretation for the chosen summary measure. This advice is given despite elongation of follow-up increasing power in the non-parametric setting. Since power gain was marginal, while clinical interpretability is vital, the recommendation remains to define the non-inferiority margin as the most clinically meaningful

in the first instance, and to only resort to considerations around power when more than one summary measure is considered clinically appropriate.

Of note, conversion of the margin is only necessary for (i) methodological comparisons of summary measures, as in this work, or (ii) situations where the summary measure was changed from the original plan after the design has been finalised. For the latter situation, we plan to publish guidelines on correctly matching the margins, where we plan to include a simulation comparison of various approaches.

Direct interpretation of the average p-values is complicated, because of a non-uniform and non-normal distribution of the observed p-values with a tendency towards zero with occasional outliers of high (> 0.4) p-values. Therefore, descriptive statistics were reported alongside the average p-values (also see Appendix I). Regardless, a clear trend was observed in the primary analysis: p-values of the DRMST and DS were consistently lower than the corresponding p-values of the HR. This was not observed in the comparison of the DS and DRMST. Additionally, the lowest average p-value was in line with the highest empirical power, for all analyses performed under the margin conversion with a flexible parametric model. In case of equivalence between summary measures in terms of power, the DRMST was generally favourable in terms of p-values. This gives an extra advantage to the DRMST, even under equivalent power.

Outcomes of the subgroup analyses were consistent with the primary analysis, strengthening the conclusion of a power advantage for the DRMST. Analysis solely on trials for which no evidence of non-proportionality was found were completely in line with the full analysis. As expected, under stratification by original summary measure, power performance increased for the corresponding summary measure. This again highlights the importance of correct margin conversion. Note that while the margin for the DRMST was calculated through the HR, even when the original margin was defined for the DS, the DRMST performs equivalently to the DS while the HR considerably drops in power for the analysis of the DS subgroup. No such behaviour was observed for the HR subgroup, with outcomes equivalent to the primary analysis.

The most noticeable difference between subgroups was observed between event risks. In the case of a low event risk, the DRMST clearly outperformed the DS and HR, but performance was similar for all summary measures when the event risk was high. This outcome is in line with the fact that estimation of the HR is dependent on the number of events, meaning that when the event risk is high – and more events occur – the HR is more informative and can be estimated more precisely. Thereby, performance of the HR in terms of power increases.

The main challenge in the use of the DS and DRMST is the choice of τ [26, 29]. Namely, τ should be chosen such that the follow-up period is long enough to evaluate the outcome reliably, while at the same time optimised as to not have unnecessarily long follow-up times. The optimal method for determining the appropriate value of τ remains unclear. In our study, little difference in performance between the two explored options of τ were observed. Because the maximum τ was defined differently from the clinical τ only for trials with additional follow-up over the pre-specified moment of evaluation, it held that $\tau_{max} \neq \tau_{clin}$ for only 18 (27.7%) trials. Therefore, τ_{max} equalled τ_{clin} for the majority of trials, leading to identical outcomes. This might have caused the marginal changes between different definitions of τ. Analysing both possibilities was done for exploration purposes, because equivalent margins between summary measures were not readily available. In line with previous recommendations [30], we recommend choosing τ as a clinically meaningful moment of evaluation.

Furthermore, we acknowledge that true follow-up times may have exceeded those reported, since it is common for trials to omit full follow-up to avoid long tails. This has some implications for all three summary measures. If the reported follow-up time was arbitrary such that enough data was presented without showing long tails with few events, our definition of τ (whether undefined at the moment of evaluation, or set as τ_{max} otherwise) holds little clinical meaning. This raises the question of how to appropriately choose a τ suitable for the situation at hand, as pointed out by Freidlin, Hu and Korn [26]. Omitting longer follow-up also affects estimates of the HR, because of possible deviations from proportional hazards and different behaviour of the event rate at different times after baseline.

In line with Weir and Trinquart [13], we found an empirical power advantage of the DRMST over the HR. As compared to their findings, we found a larger discrepancy between the HR and DRMST – i.e. more trials with a different conclusion between the HR and DRMST – possibly showing the additional benefit of estimation of DRMST under proportional hazards. However, difference between non-parametric and parametric estimation of DRMST only gave marginal differences, suggesting the stronger differences might just be due to chance. In their study, they found a different conclusion between the summary measures only for one trial. In this subset of the 35 trials, we found an additional two trials with opposing conclusions. Even with the non-inferiority margin as defined by Weir and Trinquart, who used the Weibull distribution for margin conversion, these discrepancies remain. Furthermore, our findings were in line with the simulation results of Quartagno et al. [10].

Our study has a number of limitations, the first one being the limited sample size of 65 trials. This was the number of

trials we could realistically include. The limited sample size is especially apparent in the subgroup analyses, where subgroups became as small as 24 trials. The difference in empirical power between HR and DRMST, while confirming the results of previous simulation studies, are not statistically significantly different from 0. However, the pattern of a power benefit when using DRMST across analyses confirms the plausibility that there is such an effect or even larger. With respect to the subgroup analysis for the trials without evidence of a violated proportional hazards assumption, it should be noted that trials were categorised solely on the Grambsch–Therneau test. As such, some trials portrayed clear evidence of non-proportionality (*e.g.* because of crossing Kaplan–Meier curves), but were not excluded in the subgroup analysis due to a non-significant Grambsch–Therneau test. This is further highlighted by the fact that we found additional trials to violate the proportional hazards assumption in the 35 previously reconstructed trials by Weir and Trinquart [13], and vice versa. Thus, one may wonder if, since the absence of evidence for non-proportionality is not evidence of proportionality, clinical considerations should be taken into account when determining adherence to the proportional hazards assumption.

Moreover, reconstructed data were used rather than the original; as a results, our data may have deferred from the actual underlying trial data. In doing so, all information on patient characteristics were lost, excluding the possibility of adjusted estimations. To ensure comparability, the reconstructed summary measure was compared with the original summary measure for all trials such that the reconstruction was within an acceptable range. If the original summary measure could not be reconstructed within a margin of 0.05 for the HR and DS, and 0.1 for the DRMST, the trial was excluded.

Lastly, as previously acknowledged, another limitation is the margin conversion. Since performing the margin conversion under an incorrect distribution may lead to large biases, ensuring the conversion was done correctly was a crucial aspect of this study. To ensure correct specification of the survival curve, we used observed data, thereby using the observed data twice - both in the margin conversion and later in the analyses. However, employing this approach avoids misspecification of the distribution, strengthening comparability between summary measures.

In the design of non-inferiority trials, margins should be defined directly for the chosen summary measure, without intermediate conversion between measures such that the margin is clinically meaningful. To this end, further exploration should be done on how to choose a τ that is both clinically relevant and optimises the length of follow-up. Additionally, further investigation into robustness against violation of the proportional hazards assumption is needed.

Conclusion

In non-inferiority clinical trials, the DRMST gives a power advantage over the popular HR as estimated with a Cox proportional hazards model. The DRMST and DS show power benefits when estimated under proportional hazards using a flexible parametric survival model, both compared to their non-parametrically estimated counterparts and the HR; in turn, the DRMST consistently outperformed the DS. Outcomes were variable under different methods of margin conversion, highlighting the importance of correct conversion of non-inferiority margins between summary measures for a fair comparison. In practice, margins should be designed for the appropriate summary measure with a clinically meaningful time τ at which the DS or DRMST is to be evaluated. When clinically justifiable, we recommend implementation of the DRMST in non-inferiority clinical trials with time-to-event outcomes, specifically with estimation under proportional hazards if appropriate.

Appendices

1 Literature search terms

("The New England journal of medicine"[Journal] OR "Lancet (London, England)"[Journal] OR "The Lancet. Oncology"[Journal] OR "The Lancet. Infectious diseases"[Journal] OR "The Lancet. Neurology"[Journal] OR "The Lancet. Global health"[Journal] OR "British medical journal"[Journal] OR "BMJ"[Journal] OR "BMJ (Clinical research ed.)"[Journal] OR "BMJ open"[Journal] OR "BMJ case reports"[Journal] OR "BMJ global health"[Journal] OR "BMJ quality safety"[Journal] OR "Nature"[Journal] OR "Nature medicine"[Journal] OR "Nature genetics"[Journal] OR "JAMA"[Journal] OR "JAMA network open"[Journal] OR "JAMA internal medicine"[Journal] OR "JAMA ophthalmology"[Journal] OR "JAMA surgery"[Journal])

AND

("non-inferiority" OR "non-inferior" OR "noninferiority" OR "noninferior" OR "non inferiority" OR "non inferior" OR "equivalence")

AND

("Kaplan-Meier" OR "Kaplan Meier" OR "cumulative incidence" OR "hazard ratio" OR "hazard" OR "hazards" OR "Cox" OR "time to" OR "survival" OR "median follow-up" OR "median follow up")

AND

("2021/01/01"[PDAT] : "2024/03/15"[PDAT])

NOT

("protocol"[Title] OR "systematic review"[Title] OR "systematic-review"[Title] OR "meta-analysis"[Title] OR "meta analysis"[Title])

2 Consort diagram of literature search

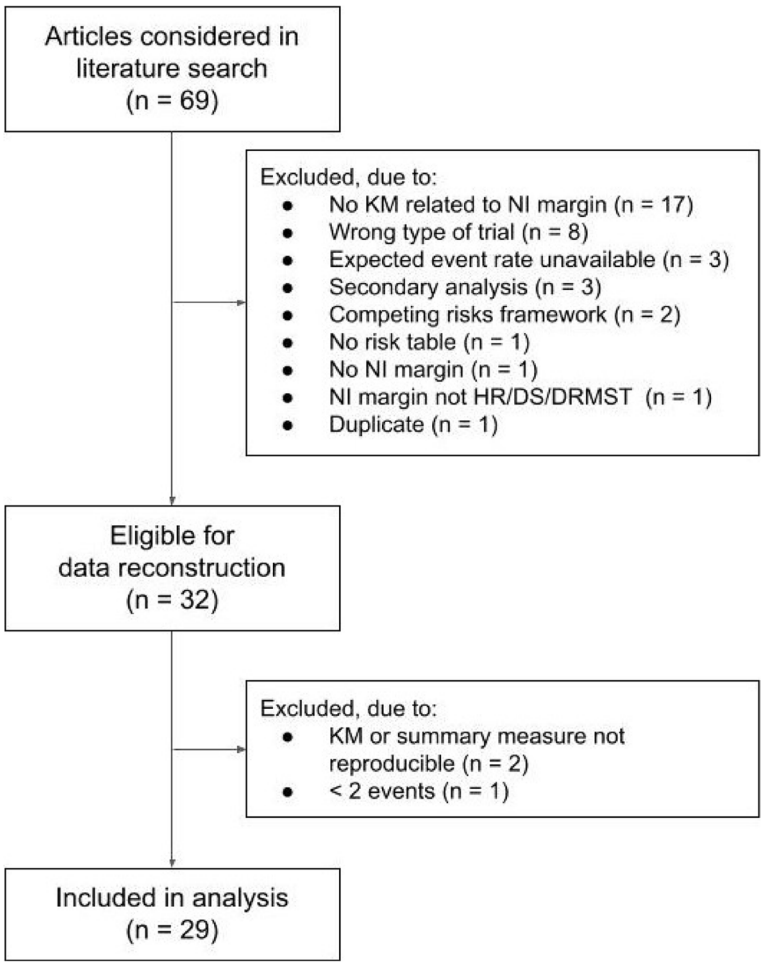


Fig. 2 Consort diagram of literature search

3 References to included trials

References to the 29 included trials. One trial (ORAL) reported two Kaplan-Meier curves with a corresponding non-inferiority margin and was therefore included twice, leading to a total of 30 datasets.

Table 3 References to included trials

Study ID	Reference
Altorki2023	Altorki N, Wang X, Kozono D, Watt C, Landrenau R, Wigle D, et al. Lobar or sublobar resection for peripheral stage IA non-small-cell lung cancer. New England Journal of Medicine. 2023;388(6):489-98.
ARIES-HM3	Mehra MR, Netuka I, Uriel N, Katz JN, Pagani FD, Jorde UP, et al. Aspirin and hemo- compatibility events with a left ventricular assist device in advanced heart failure: the ARIES-HM3 randomized clinical trial. JAMA. 2023;330(22):2171-81.
BioVasc	Diletti R, den Dekker WK, Bennett J, Schotborgh CE, van der Schaaf R, Sabaté M, et al. Immediate versus staged complete revascularisation in patients presenting with acute coronary syndrome and multivessel coronary disease (BIOVASC): a prospective, open-label, non-inferiority, randomised trial. The Lancet. 2023;401(10383):1172-82.

Study ID	Reference
CLASS-01	Huang C, Liu H, Hu Y, Sun Y, Su X, Cao H, et al. Laparoscopic vs open distal gastrectomy for locally advanced gastric cancer: five-year outcomes from the CLASS-01 randomized clinical trial. JAMA Surgery. 2022;157(1):9-17.
DYNAMIC	Tie J, Cohen JD, Lahouel K, Lo SN, Wang Y, Kosmider S, et al. Circulating tumor DNA analysis guiding adjuvant therapy in stage ll colon cancer. New England Journal of Medicine. 2022;386(24):2261-72.
Etoh2023	Etoh T, Ohyama T, Sakuramoto S, Tsuji T, Lee SW, Yoshida K, et al. Five-year survival outcomes of laparoscopy-assisted vs open distal gastrectomy for advanced gastric cancer: the JLSSG0901 randomized clinical trial. JAMA Surgery. 2023;158(5):445-54.
FAME 3	Fearon WF, Zimmermann FM, De Bruyne B, Piroth Z, van Straten AH, Szekely L, et al. Fractional flow reserve-guided PCI as compared with coronary bypass surgery. New England Journal of Medicine. 2022;386(2):128-37.
FOCUS	Yu J, Gao Y, Chen L, Wu D, Shen Q, Zhao Z, et al. Effect of S-1 plus oxaliplatin compared with fluorouracil, leucovorin plus oxaliplatin as perioperative chemotherapy for locally advanced, resectable gastric cancer: a randomized clinical trial. JAMA Network Open. 2022;5(2):e220426-6.
IMPORT HIGH	Coles CE, Haviland JS, Kirby AM, Griffin CL, Sydenham MA, Titley JC, et al. Dose-escalated simultaneous integrated boost radiotherapy in early breast cancer (IMPORT HIGH): a multicentre, phase 3, non-inferiority, open-label, randomised controlled trial. The Lancet. 2023;401(10394):2124-37.
INSURE	Pan Y, Meng X, Yuan B, Johnston SC, Li H, Bath PM, et al. Indobufen versus aspirin in patients with acute ischaemic stroke in China (INSURE): A randomised, double-blind, double-dummy, active control, non-inferiority trial. The Lancet Neurology. 2023;22(6):485-93.
INVICTUS-VKA	Connolly SJ, Karthikeyan G, Ntsekhe M, Haileamlak A, El Sayed A, El Ghamrawy A, et al. Rivaroxaban in rheumatic heart disease-associated atrial fibrillation. New England Journal of Medicine. 2022;387(11):978-88.
LIFE-BTK	Varcoe RL, DeRubertis BG, Kolluri R, Krishnan P, Metzger DC, Bonaca MP, et al. Drug-eluting resorbable scaffold versus angioplasty for infrapopliteal artery disease. New England Journal of Medicine. 2024;390(1):9-19.
Limaye2023	Limaye AP, Budde K, Humar A, Vincenti F, Kuypers DR, Carroll RP, et al. Letermovir vs valganciclovir for prophylaxis of cytomegalovirus in high-risk kidney transplant recipients: a randomized clinical trial. JAMA. 2023;330(1):33-42.
Mao2023	Mao YP, Wang SX, Gao TS, Zhang N, Liang XY, Xie FY, et al. Medial retropharyngeal nodal region sparing radiotherapy versus standard radiotherapy in patients with nasopharyngeal carcinoma: open label, non-inferiority, multicentre, randomised, phase 3 trial. BMJ. 2023;380.
NPC-CTVn	Tang LL, Huang CL, Zhang N, Jiang W, Wu YS, Huang SH, et al. Elective upper-neck versus whole-neck irradiation of the uninvolved neck in patients with nasopharyngeal carcinoma: an open-label, non-inferiority, multicentre, randomised phase 3 trial. The Lancet Oncology. 2022;23(4):479-90.
ODYSSEY	Turkova A, White E, Mujuru HA, Kekitiinwa AR, Kityo CM, Violari A, et al. Dolutegravir as first-or second-line treatment for HIV-1 infection in children. New England Journal of Medicine. 2021;385(27):2531-43.
ORAL[a]	Ytterberg SR, Bhatt DL, Mikuls TR, Koch GG, Fleischmann R, Rivas JL, et al. Cardiovascular and cancer risk with tofacitinib in rheumatoid arthritis. New England Journal of Medicine. 2022;386(4):316-26.
Plante2024	Plante M, Kwon JS, Ferguson S, Samouélian V, Ferron G, Maulard A, et al. Simple versus Radical Hysterectomy in Women with Low-Risk Cervical Cancer. New England Journal of Medicine. 2024;390(9):819-29.
POISE-3	Devereaux P, Marcucci M, Painter TW, Conen D, Lomivorotov V, Sessler DI, et al. Tranexamic acid in patients undergoing noncardiac surgery. New England Journal of Medicine. 2022;386(21):1986-97.
PROSPECT	Schrag D, Shi Q, Weiser MR, Gollub MJ, Saltz LB, Musher BL, et al. Preoperative Treatment of Locally Advanced Rectal Cancer. New England Journal of Medicine. 2023;389(4):322-34.

Study ID	Reference
Ruff2022	Ruff CT, Baron M, Im K, O'Donoghue ML, Fiedorek FT, Sabatine MS. Subcutaneous infusion of exenatide and cardiovascular outcomes in type 2 diabetes: a non-inferiority randomized controlled trial. Nature Medicine. 2022;28(1):89-95.
Saji2022	Saji H, Okada M, Tsuboi M, Nakajima R, Suzuki K, Aokage K, et al. Segmentectomy versus lobectomy in small-sized peripheral non-small-cell lung cancer (JCOG0802/WJOG4607L): a multicentre, open-label, phase 3, randomised, controlled, non-inferiority trial. The Lancet. 2022;399(10335):1607-17.
Schroder2023	Schroder JN, Patel CB, Devore AD, Bryner BS, Casalinova S, Shah A, et al. Trans- plantation outcomes with donor hearts after circulatory death. New England Journal of Medicine. 2023;388(23):2121-31.
SHARE	Min PK, Kang TS, Cho YH, Cheong SS, Kim BK, Kwon SW, et al. P2Y12 Inhibitor Monotherapy vs Dual Antiplatelet Therapy After Deployment of a Drug-Eluting Stent: The SHARE Randomized Clinical Trial. JAMA network open. 2024;7(3):e240877-7.
STAR	Brown JE, Royle KL, Gregory W, Ralph C, Maraveyas A, Din O, et al. Temporary treatment cessation versus continuation of first-line tyrosine kinase inhibitor in patients with advanced clear cell renal cell carcinoma (STAR): An open-label, non-inferiority, randomised, controlled, phase 2/3 trial. The Llancet Ooncology. 2023;24(3):213-27.
Tang2022	Tang LL, Guo R, Zhang N, Deng B, Chen L, Cheng ZB, et al. Effect of radiotherapy alone vs radiotherapy with concurrent chemoradiotherapy on survival without disease relapse in patients with low-risk nasopharyngeal carcinoma: a randomized clinical trial. JAMAama. 2022;328(8):728-36.
TRAVERSE	Lincoff AM, Bhasin S, Flevaris P, Mitchell LM, Basaria S, Boden WE, et al. Cardio- vascular safety of testosterone-replacement therapy. New England Journal of Medicine. 2023;389(2):107-17.
UK TAVI	Investigators TUTT. Effect of Transcatheter Aortic Valve Implantation vs Surgical Aortic Valve Replacement on All-Cause Mortality in Patients With Aortic Stenosis: A Randomized Clinical Trial. JAMA. 2022 05;327(19):1875-87.
Yuan2023	Yuan P, Kang Y, Ma F, Fan Y, Wang J, Wang X, et al. Effect of Epirubicin Plus Paclitaxel vs Epirubicin and Cyclophosphamide Followed by Paclitaxel on Disease-Free Survival Among Patients With Operable ERBB2-Negative and Lymph Node-Positive Breast Can- cer: A Randomized Clinical Trial. JAMA Network Open. 2023;6(2):e230122-2.

[a] Two curves were included from this trial

4 Reconstructed Kaplan-Meier curves

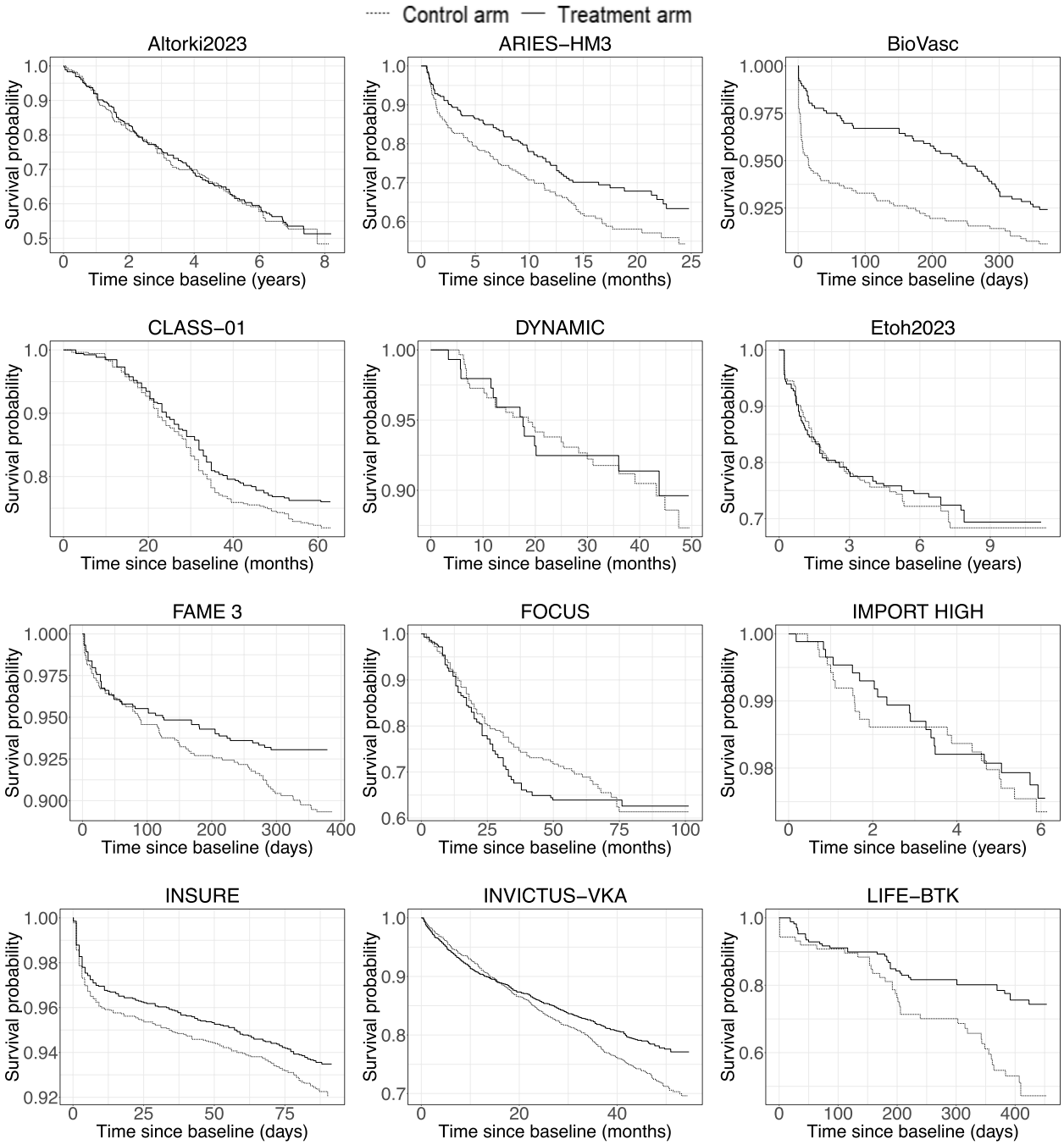


Fig. 3 A selection (Curves 1-12) of reconstructed Kaplan-Meier curves for trials included in this study

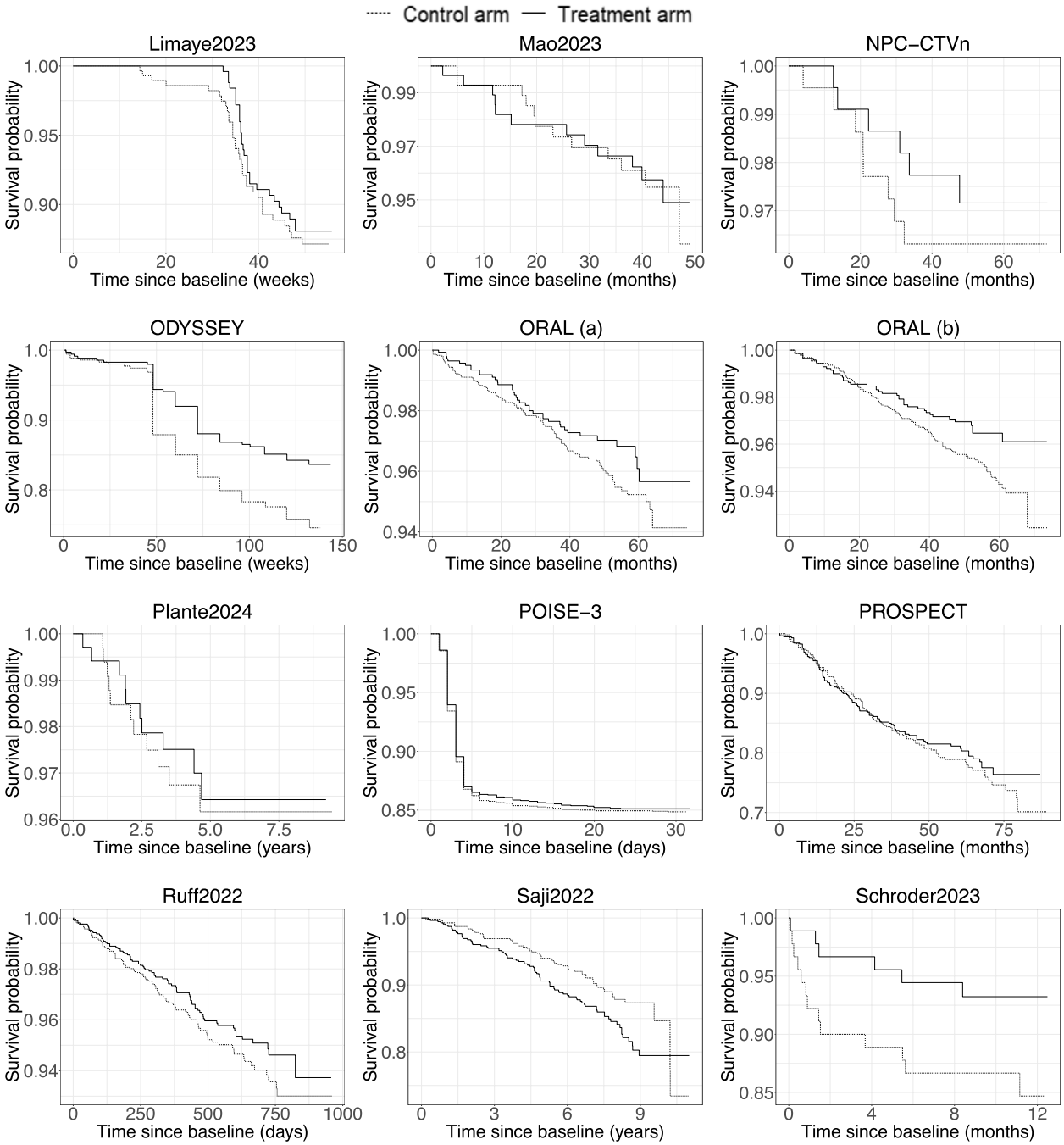


Fig. 4 A selection (Curves 13-24) of reconstructed Kaplan-Meier curves for trials included in this study

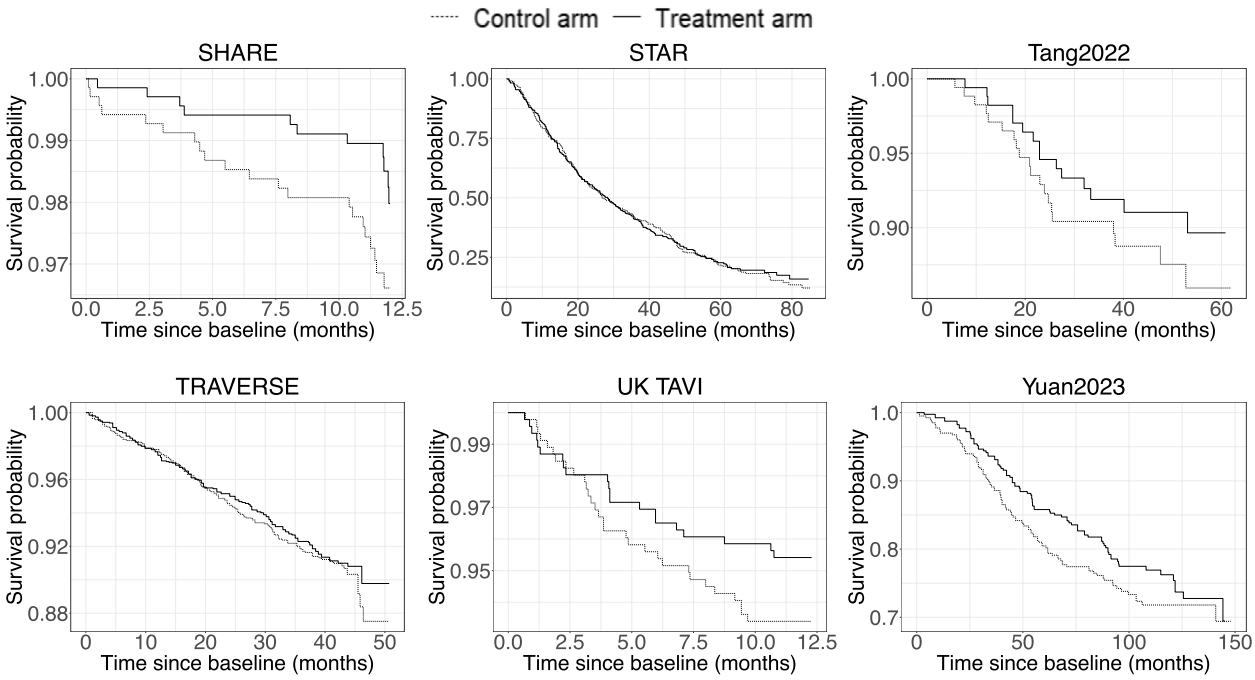


Fig. 5 A selection (Curves 25-30) of reconstructed Kaplan-Meier curves for trials included in this study

5 Reported significance level and power

Table 4 Reported significance level and power

Study ID	Significance level	One/twosided	Power
Altorki2023	0.05	onesided	0.8
ARIES-HM3	0.025	onesided	0.8
BioVasc	0.05	twosided	0.8
CLASS-01	0.05	onesided	0.9
DYNAMIC	0.05	twosided	0.8
Etoh2023	0.05	onesided	0.75
FAME 3	0.025	onesided	0.9
FOCUS	0.025	onesided	0.8
IMPORT HIGH	0.025	onesided	0.8
INSURE	0.025	onesided	0.8
INVICTUS-VKA	0.025	onesided	0.86
LIFE-BTK	0.025	onesided	0.84
LIFE-BTK	0.025	onesided	0.84
Limaye2023	0.05	twosided	0.9
Mao2023	0.025	onesided	0.8
NPC-CTVn	0.025	onesided	0.8
ODYSSEY	0.05	twosided	0.9
ORAL	0.05	twosided	0.8
Plante2024	0.05	onesided	0.85
POISE-3	0.025	onesided	0.98
PROSPECT	0.098	twosided	0.85
Ruff2022	0.046	twosided	0.9
Saji2022	0.05	onesided	0.8
Schroder2023	0.1	twosided	0.8
SHARE	0.05	twosided	0.8
STAR	0.05	twosided	0.8
Tang2022	0.05	onesided	0.8
TRAVERSE	0.05	twosided	0.9
UK TAVI	0.025	onesided	0.8
Yuan2023	0.05	onesided	0.8

6 Reconstructed effect sizes

Table 5 Reconstructed effect size estimates and evidence of non-proportionality

Study ID	Summary measure	Effect estimate		Time horizon		Adjusted	NI Concluded	Evidence of non-proportionality
		Reported	*Reconstructed*	*Time*	*Unit*			
Altorki2023	HR	1.01	1.02	7	y	Yes	Yes	No
BioVasc	HR	0.78	0.78	365	d	No	Yes	Yes
Etoh2023	HR	0.96	0.94	5	y	Yes	Yes	No
FAME 3	HR	1.5	1.55	365	d	Yes	No	No
INSURE	HR	1.23	1.22	90	d	Yes	No	No
INVICTUS-VKA	HR	1.25	1.25	54	m	Yes	No	Yes
ORAL (a)	HR	1.33	1.33	72	m	No	No	No
ORAL (b)	HR	1.48	1.48	72	m	No	No	No
POISE-3	HR	1.02	1.02	30	d	Yes	No	No
PROSPECT	HR	0.92	0.90	60	m	Yes	Yes	No
Ruff2022	HR	1.21	1.21	900	d	Yes	Yes	No
Saji2022	HR	0.663	0.71	5	y	Yes	Yes	No
STAR[a]	HR	1.03	1.01	24	m	Yes	No	No
TRAVERSE	HR	0.96	0.92	48	m	No	Yes	No
Yuan2023	HR	0.82	0.82	60	m	No	Yes	Yes
ARIES-HM3	DS	-0.06	-0.0685	12	m	No	Yes	No
DYNAMIC	DS	-0.011	-0.0228	24	m	No	Yes	No
FOCUS	DS	-0.074	-0.0780	36	m	No	Yes	No
IMPORT HIGH	DS	-0.001	-0.0010	5	y	No	Yes	No
LIFE-BTK	DS	-0.3	-0.2539	365	d	No	Yes, superiority	No
Limaye2023	DS	-0.014	-0.0094	52	w	Yes	Yes	No
Mao2023	DS	-0.002	-0.0011	36	m	Yes	Yes	No
NPC-CTVn	DS	-0.014	-0.0143	36	m	No	Yes	No
ODYSSEY	DS	-0.08	-0.0851	96	w	Yes	Yes, superiority	No
Plante2023	DS	-0.0035	-0.0037	3	y	No	Yes	No
Schroder2023	DS	-0.03	-0.0778	6	m	Yes	No	No
SHARE	DS	-0.0136	-0.0137	12	m	No	Yes	No
Tang2022	DS	-0.014	-0.0148	36	m	No	Yes	No
UK TAVI	DS	-0.02	-0.0044	12	m	Yes	Yes	No
CLASS-01	DRMST	1.17	1.07	60	m	No	Yes	No

[a] Inverse of the reported effect estimate is given here for uniformity.

Abbreviations: *NI* non-inferiority, *HR* hazard ratio, *DS* difference in survival, *DRMST* difference in restricted mean survival time, *d* day, *w* week, *m* month, *y* year

7 Converted non-inferiority margins

Table 6 Converted non-inferiority margins, original summary measure: hazard ratio

Study ID	Original margin		Conversion under flexible parametric model		Conversion under exponential distribution			τ_{clin}	
	Type	*Margin*	*DS*	*DRMST*	λ	*DS*	*DRMST*	*Time*	*Unit*
Altorki2023	HR	1.306	-0.09	-0.42	0.0862	-0.09	-0.41	7	y
BioVasc	HR	1.39	-0.03	-10.23	0.0003	-0.04	-7.56	365	d
Etoh2023	HR	1.31	-0.07	-0.24	0.0862	-0.08	-0.24	5	y
FAME 3	HR	1.65	-0.04	-11.86	0.0003	-0.06	-11.39	365	d
INSURE	HR	1.25	-0.02	-0.98	0.0010	-0.02	-0.99	90	d
INVICTUS-VKA	HR	1.46	-0.11	-3.46	0.0017	-0.04	-1.05	54	m
ORAL (a)	HR	1.8	-0.03	-1.23	0.0009	-0.21	-11.37	72	m
ORAL (b)	HR	1.8	-0.03	-1.28	0.0009	-0.21	-11.37	72	m
PROSPECT	HR	1.29	-0.05	-1.86	0.0050	-0.08	-3.01	60	m
POISE-3	HR	1.125	-0.02	-0.44	0.0084	-0.02	-0.25	30	d
Ruff2022	HR	1.8	-0.05	-22.59	0.0000	-0.01	-5.27	900	d
Saji2022	HR	1.54	-0.05	-0.10	0.0211	-0.05	-0.13	5	y
STAR[a]	HR	1.232	-0.07	-1.00	0.0302	-0.07	-1.20	24	m
TRAVERSE	HR	1.5	-0.05	-1.22	0.0013	-0.03	-0.69	48	m
Yuan2023	HR	1.30	-0.05	-1.44	0.0031	-0.05	-1.46	60	m
AUSTRI	HR	2	-0.01	-0.02	0.0013	-0.01	-0.02	6	m
CHORUS	HR	1.18	-0.06	-1.71	0.0193	-0.06	-1.38	36	m
COMPARZ	HR	1.25	-0.07	-1.63	0.0630	-0.07	-1.59	24	m
ELIXA	HR	1.3	-0.02	-0.15	0.0088	-0.03	-0.18	12	m
EXAMINE	HR	1.3	-0.04	-0.81	0.0052	-0.04	-0.62	30	m
FIREandICE	HR	1.43	-0.11	-21.70	0.0010	-0.10	-21.09	365	d
FLAME	HR	1.15	-0.04	-2.30	0.0006	0.00	-0.12	52	w
HOKUSAI-VTE	HR	1.5	-0.02	-4.03	0.0001	-0.01	-2.71	365	d
Hussain-NEJM	HR	1.2	-0.06	-0.43	0.2374	-0.04	-0.48	9	y
IBIS-II DCIS	HR	1.25	-0.01	-0.02	0.0161	-0.02	-0.05	5	y
JASPAC 01	HR	1.25	-0.08	-0.16	0.3406	-0.08	-0.19	3	y
MERTH	HR	1.7	-0.05	-0.09	0.1054	-0.14	-0.25	3	y
NBCVOT	HR	1.4	-0.01	-0.55	0.0003	-0.01	-0.61	104	w
PARTNER2	HR	1.2	-0.04	-0.68	0.0149	-0.05	-0.66	24	m
Peters-NEJM	HR	2	-0.01	-0.71	0.0000	-0.01	-0.68	182.5	d
PET-NECK	HR	1.5	-0.08	-1.04	0.0120	-0.10	-1.36	24	m
PROTECT AF	HR	2	-0.14	-4.55	0.0051	-0.02	-7.73	60	m
REDUCE	HR	1.515	-0.13	-18.90	0.0011	-0.15	-18.66	182.5	d
RE-SONATE	HR	2.85	-0.02	-0.22	0.0038	-0.04	-0.33	18	m
Rummel-Lancet	HR	1.32	-0.10	-2.35	0.0193	-0.10	-2.38	36	m
SIMPLE	HR	1.5	-0.03	-0.05	0.0513	-0.05	-0.07	2.5	y
SUSTAIN6	HR	1.8	-0.06	-3.50	0.0004	-0.03	-1.60	104	w
TECOS	HR	1.3	-0.04	-1.04	0.0023	-0.03	-0.74	48	m
VESTRI	HR	2.675	-0.01	-0.03	0.0012	-0.01	-0.03	6	m
WISDOM	HR	1.2	-0.06	-2.42	0.0015	-0.01	-2.02	52	w

[a] Inverse of the reported non-inferiority margin is given here for uniformity.

Abbreviations: *HR* hazard ratio, *DS* difference in survival, *DRMST* difference in restricted mean survival time, *d* day, *w* week, *m* month, *y* year

Table 7 Converted non-inferiority margins, original summary measure: difference in survival

Study ID	Original margin		Conversion under flexible parametric model		Conversion under exponential distribution			τ_{clin}	
	Type	*Margin*	*HR*	*DRMST*	λ	*HR*	*DRMST*	*Time*	*Unit*
ARIES-HM3	DS	-0.1	1.42	-0.86	0.0285	1.44	-0.69	12	m
DYNAMIC	DS	-0.085	2.32	-0.98	0.0073	1.61	-1.10	24	m
FOCUS	DS	-0.08	1.33	-1.54	0.0107	1.32	-1.68	36	m
IMPORT HIGH	DS	-0.03	2.49	-0.08	0.0103	1.63	-0.08	5	y
LIFE-BTK	DS	-0.1	1.35	-21.78	0.0012	1.39	-21.81	365	d
Limaye2023	DS	-0.1	1.91	-1.73	0.0036	1.69	-2.83	52	w
Mao2023	DS	-0.08	3.37	-1.50	0.0017	2.44	-1.49	36	m
NPC-CTVn	DS	-0.08	3.73	-1.42	0.0008	3.83	-1.48	36	m
ODYSSEY	DS	-0.1	1.58	-4.33	0.0021	1.66	-5.25	96	w
Plante2024	DS	-0.04	2.89	-0.05	0.0136	2.04	-0.06	3	y
Schroder2023	DS	-0.2	2.92	-0.94	0.0121	4.34	-0.64	6	m
SHARE	DS	-0.03	1.94	-0.16	0.0043	1.63	-0.18	12	m
Tang2022	DS	-0.1	2.38	-1.51	0.0029	2.12	-1.90	36	m
UK TAVI	DS	-0.05	1.81	-0.37	0.0065	1.71	-0.31	12	m
ACTI	DS	-0.03	1.90	-8.93	0.0001	1.69	-5.59	365	d
ARROW	DS	-0.055	1.51	-3.09	0.0010	1.67	-2.76	96.2	w
BEST	DS	-0.04	1.56	-0.06	0.0639	1.36	-0.04	2	y
EXCEL	DS	-0.042	1.32	-1.25	0.0032	1.41	-0.79	36	m
Johnson-NEJM	DS	-0.05	1.39	-1.07	0.0014	2.05	-0.92	36	m
LEADERS FREE	DS	-0.032	1.28	-8.50	0.0002	1.42	-6.04	365	d
OPTIMIZE	DS	-0.027	1.48	-6.98	0.0003	1.32	-5.11	365	d
RAPID	DS	-0.07	2.40	-1.85	0.0014	2.49	-1.30	36	m
SIOP WT 2001	DS	-0.1	2.53	-1.39	0.0063	1.82	-1.29	24	m
US Core Valve[a]	DS	-0.075	1.46	-0.62	0.0186	1.44	-0.49	12	m

[a] The margin reported here differs from the margin reported by Weir and Trinquart [13].

Abbreviations: *HR* hazard ratio, *DS* difference in survival, *DRMST* difference in restricted mean survival time, *d* day, *w* week, *m* month, *y* year

Table 8 Converted non-inferiority margins, original summary measure: difference in restricted mean survival time

Study ID	Original margin		Conversion under flexible parametric model		Conversion under exponential distribution			τ_{clin}	
	Type	*Margin*	*HR*	*DS*	λ	*HR*	*DS*	*Time*	*Unit*
CLASS-01	DRMST	-10	2.53	-0.28	0.0090	2.04	-0.25	60	m

Abbreviations: *HR* hazard ratio, *DS* difference in survival, *DRMST* difference in restricted mean survival time, *m* month

Table 9 Converted non-inferiority margins for trials where $\tau_{clin} \neq \tau_{max}$

Study ID	Original margin		Conversion under flexible parametric model		Conversion under exponential distribution			τ_{max}	
	Type	*Margin*	*DS*	*DRMST*	λ	*DS*	*DRMST*	*Time*	*Unit*
Etoh2023	HR	1.31	-0.08	-0.71	0.0862	-0.10	-0.85	5	y
PROSPECT	HR	1.29	-0.07	-3.67	0.0084	-0.09	-5.60	60	m
Saji2022	HR	1.54	-0.10	-0.56	0.0211	-0.09	-0.57	5	y
STAR[a]	HR	1.232	-0.05	-5.15	0.0302	-0.03	-4.84	84	m
Yuan2023	HR	1.30	-0.07	-7.06	0.0031	-0.08	-7.23	60	m
CHORUS	HR	1.18	-0.02	-4.29	0.0193	-0.04	-4.98	102	m
COMPARZ	HR	1.25	-0.04	-2.49	0.0630	-0.04	-2.40	39	m
ELIXA	HR	1.3	-0.04	-1.12	0.0088	-0.07	-1.61	40	m
FIREandICE	HR	1.43	-0.13	-100.73	0.0010	-0.13	-99.51	1000	d
Hussain-NEJM	HR	1.2	-0.04	-0.76	0.2374	-0.01	-0.63	15	y
IBIS-II DCIS	HR	1.25	-0.02	-0.09	0.0161	-0.03	-0.18	10	y
JASPAC 01	HR	1.25	-0.07	-0.32	0.3406	-0.06	-0.34	5	y
NBCVOT	HR	1.4	-0.01	-1.13	0.0003	-0.02	-1.37	156	w
Peters-NEJM	HR	2	-0.01	-0.90	0.0000	-0.01	-0.90	210	d
PET-NECK	HR	1.5	-0.11	-4.53	0.0120	-0.15	-6.09	60	m
RE-SONATE	HR	2.85	-0.05	-0.85	0.0011	-0.07	-1.23	35	m
Rummel-Lancet	HR	1.32	-0.08	-7.62	0.0193	-0.07	-7.60	93	m
SIMPLE	HR	1.5	-0.04	-0.09	0.0513	-0.07	-0.15	4	y

[a] Inverse of the reported effect estimate is given here for uniformity.

Abbreviations: *HR* hazard ratio, *DS* difference in survival, *DRMST* difference in restricted mean survival time, *d* day, *w* week, *m* month, *y* year

8 Complete overview of outcomes

Outcomes of all sub-analyses

Table 10 Estimated empirical power for all analyses

Summary measure	Model	τ	Margin conversion		Subgroups				
			Flexsurv	*Exponential*	*PH*	*Original margin*		*Event risk*	
						HR	DS	Low	High
			$n = 65$	$n = 65$	$n = 57$	$n = 40$	$n = 24$	$n = 33$	$n = 32$
DRMST	flexsurv (PH)	clinical	**0.862**	0.831	**0.842**	**0.875**	**0.833**	**0.818**	**0.906**
DRMST	flexsurv (PH)	maximum	**0.862**	0.815	**0.842**	**0.875**		**0.818**	**0.906**
DRMST	KM	clinical	0.846	0.831	0.825	0.850	**0.833**	**0.818**	0.875
DRMST	KM	maximum	**0.862**	0.831	**0.842**	**0.875**		**0.818**	**0.906**
DS	flexsurv (PH)	clinical	0.831	**0.846**	0.807	0.825	**0.833**	0.758	**0.906**
DS	flexsurv (PH)	maximum	0.831	0.815	0.807	0.825		0.758	**0.906**
DS	KM	clinical	0.785	0.800	0.772	0.750	**0.833**	0.758	0.812
DS	KM	maximum	0.754	0.769	0.754	0.700		0.727	0.781
HR	Cox	-	0.785	0.754	0.772	0.825	0.708	0.667	**0.906**

Bold-faced values indicate the highest estimated empirical power in the column.

Abbreviations: *HR* hazard ratio, *DS* difference in survival, *DRMST* difference in restricted mean survival time, *PH* proportional hazards, *KM* Kaplan-Meier

Table 11 Average *p*-values for all analyses

Summary measure	Model	τ	Margin conversion		Subgroups				
			Flexsurv	*Exponential*	*PH*	*Original margin*		*Event risk*	
						HR	DS	Low	High
			$n = 65$	$n = 65$	$n = 57$	$n = 40$	$n = 24$	$n = 33$	$n = 32$
DRMST	flexsurv (PH)	clinical	**0.031**	0.067	**0.035**	0.022	**0.047**	**0.053**	**0.008**
DRMST	flexsurv (PH)	maximum	**0.031**	0.069	**0.035**	0.022		0.054	**0.008**
DRMST	KM	clinical	0.033	0.063	0.037	0.019	0.057	**0.053**	0.012
DRMST	KM	maximum	0.032	0.065	0.036	**0.018**		**0.053**	0.011
DS	flexsurv (PH)	clinical	0.035	0.049	0.040	0.027	0.049	0.060	0.009
DS	flexsurv (PH)	maximum	0.037	0.056	0.041	0.030		0.060	0.012
DS	KM	clinical	0.059	0.062	0.063	0.067	0.049	0.085	0.033
DS	KM	maximum	0.063	0.065	0.065	0.072		0.093	0.031
HR	Cox	-	0.043	**0.047**	0.047	0.033	0.061	0.073	0.012

Bold-faced values indicate the lowest average *p*-value in the column.

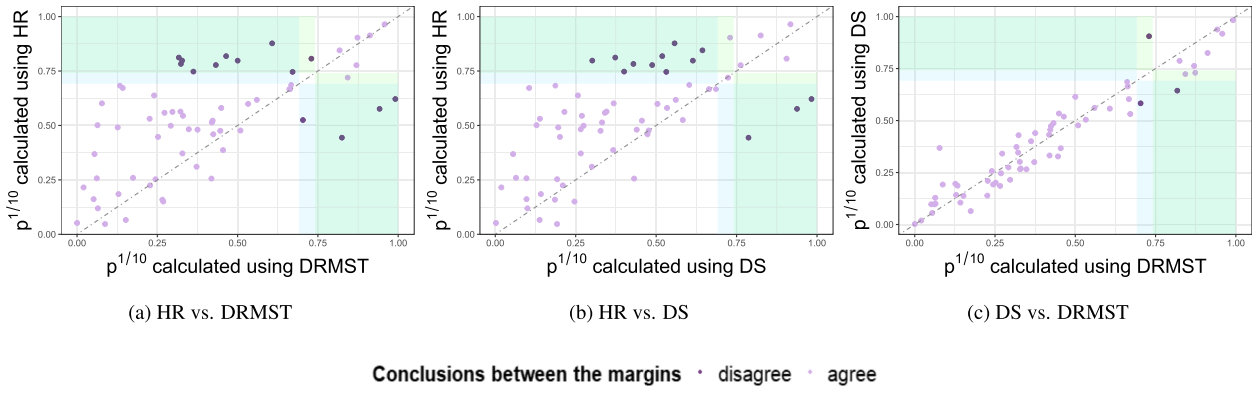
Abbreviations: *HR* hazard ratio, *DS* difference in survival, *DRMST* difference in restricted mean survival time, *PH* proportional hazards, *KM* Kaplan-Meier

Primary outcomes of margin conversion under the exponential distribution

Table 12 Descriptive statistics of *p*-values for primary analyses done with margin conversion under the exponential distribution

Summary measure	Model	τ	*P*-value characteristics			
			mean	*median*	*minimum*	*maximum*
DRMST	flexsurv (PH)	clinical	0.067	$3.91 \cdot 10^{-5}$	$2.33 \cdot 10^{-36}$	$9.15 \cdot 10^{-1}$
DRMST	flexsurv (PH)	maximum	0.069	$3.91 \cdot 10^{-5}$	$2.33 \cdot 10^{-36}$	$9.15 \cdot 10^{-1}$
DRMST	KM	clinical	0.063	$2.44 \cdot 10^{-4}$	$2.34 \cdot 10^{-31}$	$9.05 \cdot 10^{-1}$
DRMST	KM	maximum	0.065	$1.35 \cdot 10^{-4}$	$2.34 \cdot 10^{-31}$	$9.05 \cdot 10^{-1}$
DS	flexsurv (PH)	clinical	0.049	$4.40 \cdot 10^{-5}$	$2.77 \cdot 10^{-27}$	$8.41 \cdot 10^{-1}$
DS	flexsurv (PH)	maximum	0.056	$4.93 \cdot 10^{-5}$	$2.77 \cdot 10^{-27}$	$8.41 \cdot 10^{-1}$
DS	KM	clinical	0.062	$1.07 \cdot 10^{-4}$	$9.28 \cdot 10^{-27}$	$9.87 \cdot 10^{-1}$
DS	KM	maximum	0.065	$2.19 \cdot 10^{-4}$	$9.28 \cdot 10^{-27}$	$9.87 \cdot 10^{-1}$
HR	Cox	-	0.047	$1.77 \cdot 10^{-3}$	$1.64 \cdot 10^{-60}$	$6.98 \cdot 10^{-1}$

Abbreviations: *HR* hazard ratio, *DS* difference in survival, *DRMST* difference in restricted mean survival time, *PH* proportional hazards, *KM* Kaplan-Meier



(a) HR vs. DRMST (b) HR vs. DS (c) DS vs. DRMST

Fig. 6 *P*-values per summary measure (conversion under exponential distribution), using a Box-Cox transformation with $\lambda = \frac{1}{10}$

9 *P*-value distributions for primary outcomes

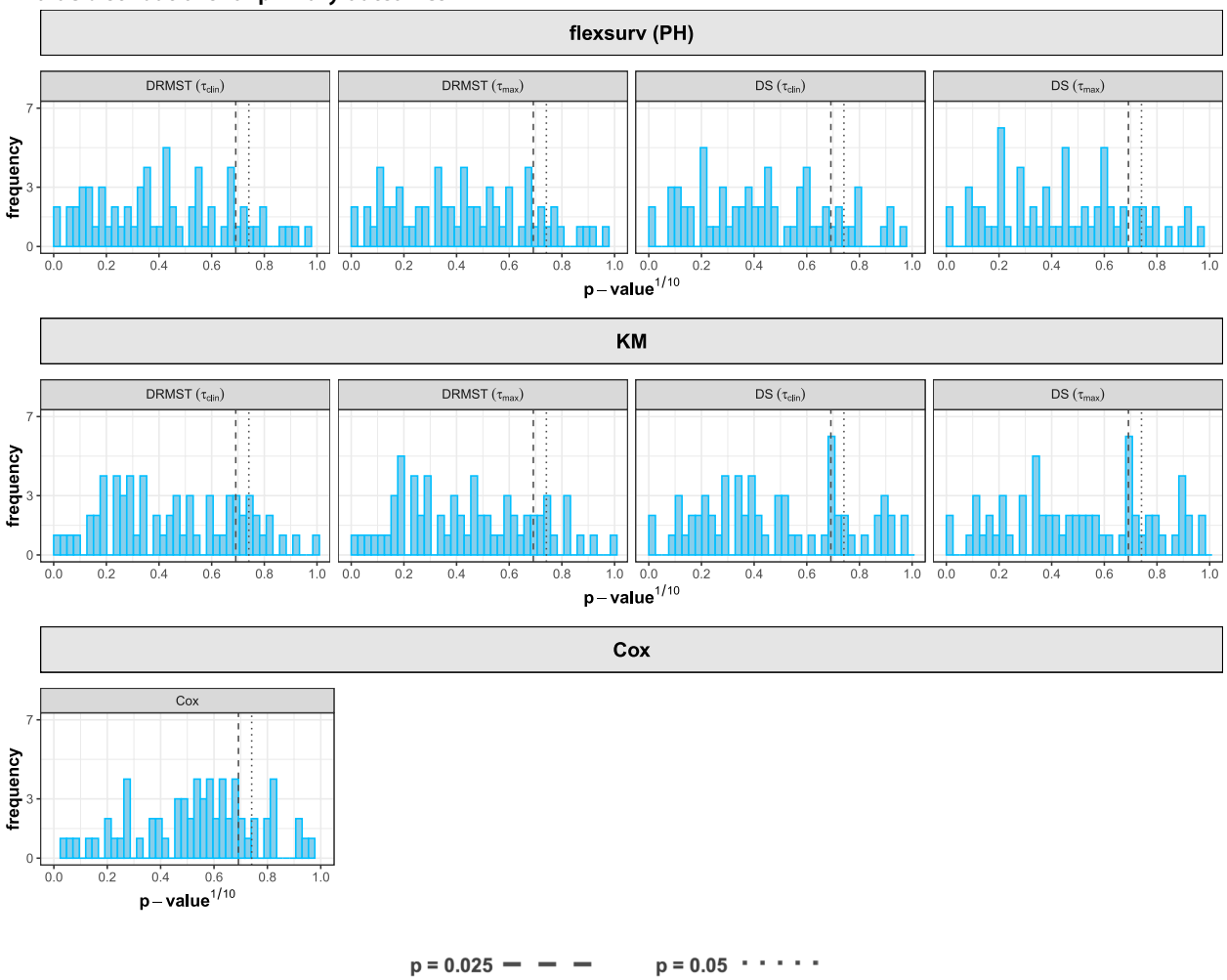


Fig. 7 *P*-value distributions (margin conversion under flexible parametric regression), using a Box-Cox transformation with $\lambda = \frac{1}{10}$

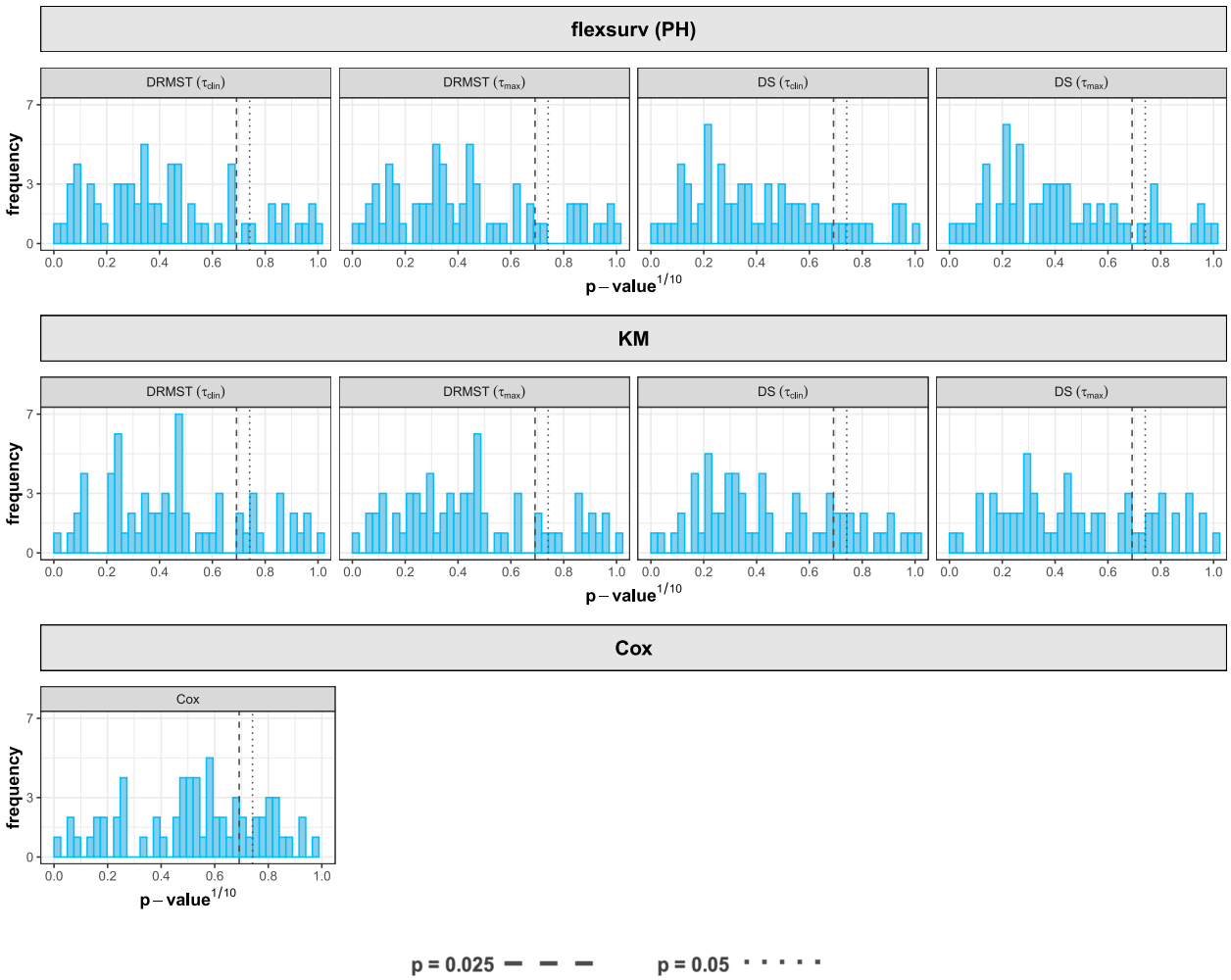


Fig. 8 *P*-value distributions (margin conversion under exponential distribution), using a Box-Cox transformation with $\lambda = \frac{1}{10}$

10 Margin conversion: further details and examples

This appendix further elaborates on how margin conversion was performed, with in particular two worked out examples that illustrate our methodology practically. Firstly, we note that the estimated survival functions used for margin conversion are not the survival functions used for the analyses. Therefore, we denote the survival functions used for margin conversion with the superscript ^{MC}.

Using the proportional hazards assumption, we can substitute $S_A^{MC}(\tau)$ with $S_C^{MC}(\tau)$ and rewrite Eqs. (1) and (2) as

$$DS(\tau) = \left(S_C^{MC}(\tau)\right)^{HR} - S_C^{MC}(\tau); \tag{3}$$

$$DRMST(\tau) = \int_0^\tau \left(S_C^{MC}(t)\right)^{HR} dt - \int_0^\tau S_C^{MC}(t)\, dt \tag{4}$$

To convert margins, we therefore only need to estimate the survival function of the control arm. Using the known parameters, i.e. the non-inferiority margin, τ and an estimation of the survival function, we can solve for the two unknown margins. In the case where the non-inferiority margin was defined as a HR, it is straightforward to obtain the DS and DRMST margins from the above equations. To obtain the HR margin from a DS, we use that since $DS(\tau) = \left(S_C^{MC}(\tau)\right)^{HR} - S_C^{MC}(\tau)$, it immediately follows that

$$DS(\tau) + S_C^{MC}(\tau) = \left(S_C^{MC}(\tau)\right)^{HR},$$

and therefore

$$HR = \log_{S_C^{MC}(\tau)}(\mathrm{DS}(\tau) + \mathrm{S_C^{MC}}(\tau)). \tag{5}$$

Using Eqs. (3), (4) and (5), we can convert all margins between the three different measures.

In our study, we compared two types of margin conversion. The first approach assumes an exponential distribution for $S_C^{MC}(\tau)$, while the second approximates $S_C^{MC}(\tau)$ using flexible parametric regression. To illustrate how to convert margins, we work through two examples using an exponential distribution for the survival function. Margin conversion using the flexible parametric approach is done analogously, with the only difference being the estimation of $S_C^{MC}(\tau)$. Because of identical methodology and the impracticality of writing out the estimated flexible parametric survival function, we will not further illustrate an example of the approach under flexible parametric estimation.

Additionally, we note that conversion from the DRMST margin to the HR and DS margins require numerical optimisation methods. To obtain the HR margin, we let `R` numerically solve Eq. (4) for the HR. This estimate is then used to also calculate the DS margin using the methodology illustrated in example 1 below.

Example 1: conversion from an HR margin to the DS and DRMST margins

In this example, we show how to obtain the DS and DRMST margins for trials that defined the non-inferiority margin on the HR scale. We will use the Altorki2023 trial for this demonstration [31]. The non-inferiority margin for this trial was 1.306 years with $\tau = 7$ years. From the manuscript, we obtain the event rate $\lambda = 0.0862$. The DS at 7 years then follows as,

$$\begin{aligned} DS(\tau) &= S_A^{MC}(\tau) - S_C^{MC}(\tau) \\ &= S_C^{MC}(\tau)^{HR} - S_C^{MC}(\tau) \\ &= \exp(-\lambda \cdot \tau \cdot \mathrm{HR}) - \exp(-\lambda \cdot \tau) \\ &= \exp(-0.0862 \cdot 7 \cdot 1.306) - \exp(-0.0862 \cdot 7) \\ &= -0.0922. \end{aligned}$$

We obtain the DRMST(τ) in a similar fashion,

$$\begin{aligned} DRMST(\tau) &= \int_0^\tau S_A^{MC}(t) - S_C^{MC}(t)\, dt \\ &= \int_0^\tau S_C^{MC}(t)^{HR} - S_C^{MC}(t)\, dt \\ &= \int_0^\tau \exp(-\lambda \cdot \mathrm{t} \cdot \mathrm{HR}) - \exp(-\lambda \cdot \mathrm{t})\, \mathrm{dt} \\ &= \int_0^7 \exp(-0.0862 \cdot \mathrm{t} \cdot 1.306) - \exp(-0.0862 \cdot \mathrm{t})\, \mathrm{dt} \\ &= -0.4123 \text{ years}. \end{aligned}$$

Example 2: conversion from a DS margin to the HR and DRMST margins

Next, using the ARIES-HM3 trial, we illustrate how to obtain the HR and DRMST margins for trials that defined the non-inferiority margin as a DS. They defined their DS margin as -0.1 at $\tau = 12$ months, with an event rate of $\lambda = 0.0285$. The corresponding HR margin can then be calculated as

$$\begin{aligned} HR &= \log_{S_C^{MC}(\tau)}\left(\mathrm{DS}(\tau) + \mathrm{S_C^{MC}}(\tau)\right) \\ &= \log_{\exp(-\lambda\cdot\tau)}(\mathrm{DS}(\tau) + \exp(-\lambda \cdot \tau)) \\ &= \log_{\exp(-0.0285\cdot 12)}(-0.1 + \exp(-0.0285 \cdot 12)) \\ &= 1.4436. \end{aligned}$$

To obtain the DRMST margin, it is easiest to use the newly calculated HR margin. Note that in the computations, we used the non-rounded HR margins (as far as the software allows this). The DRMST margin then follows as,

$$
\begin{aligned}
DRMST(\tau) &= \int_0^{\tau} S_A^{MC}(t) - S_C^{MC}(t)\, dt \\
&= \int_0^{\tau} S_C^{MC}(t)^{HR} - S_C^{MC}(t)\, dt \\
&= \int_0^{\tau} \exp(-\lambda \cdot \mathrm{t} \cdot \mathrm{HR}) - \exp(-\lambda \cdot \mathrm{t})\, \mathrm{dt} \\
&= \int_0^{12} \exp(-0.0285 \cdot \mathrm{t} \cdot 1.443642) - \exp(-0.0285 \cdot \mathrm{t})\, \mathrm{dt} \\
&= -0.6927 \text{ months.}
\end{aligned}
$$

11 Best and worst fitting exponential distributions

This appendix demonstrates how misspecification of the exponential distribution leads to incorrect margin conversion. Each plot contains the Kaplan-Meier curve based on the reconstructed data (i.e. the 'real' survival curve), accompanied by the survival curve as estimated by (i) an exponential distribution, and (ii) flexible parametric regression. The rate parameter of the exponential distribution was based on the reported event rate. Figure 9 shows the three trials for which the exponential distribution was most accurately specified, while Fig. 10 shows the three trials for which the exponential distribution was most inaccurately specified.

Since margin conversion was based on the estimated survival curves, incorrect specification of the survival curve leads to margins which are not comparable between summary measures. Comparison between Figs. 9 and 10 demonstrate the robustness of the flexible parametric regression as compared to an exponential distribution, leading to the former having been our preferred method for margin conversion.

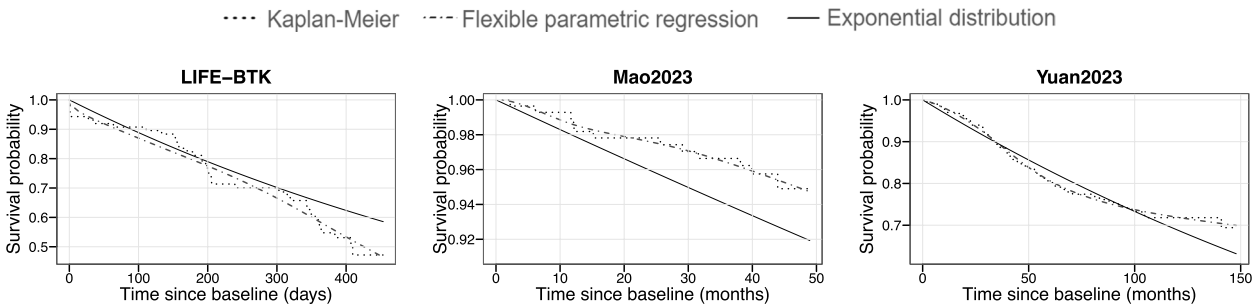


Fig. 9 Estimated survival curves for the trials with the most similar converted margins

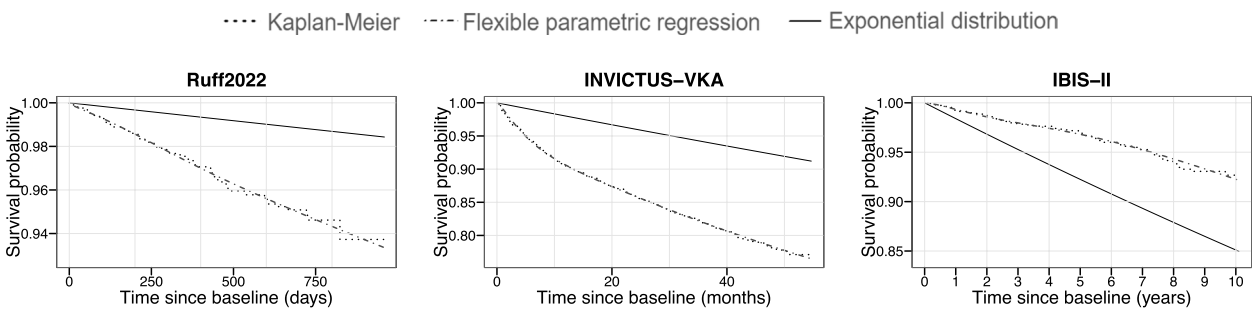


Fig. 10 Estimated survival curves for the trials with the most different converted margins

Abbreviations

DRMST	Difference in restricted mean survival time
DS	Difference in survival
HR	Hazard ratio
KM	Kaplan-Meier

Acknowledgements
Not applicable.

Authors'contributions
SDLB - data collection, analysis, first draft and final manuscript preparation. IaRW - conceptualisation, methodology, revising and editing of manuscript. TPM - conceptualisation, methodology, revising and editing of manuscript. IsRW - data collection revising of manuscript. MF - revising and editing of manuscript. MQ - conceptualisation, methodology, revising and editing of manuscript.

Funding
MQ, IaRW and TPM were supported by the Medical Research Council Programme MC_UU_00004/09.

Data availability
The reconstructed datasets supporting the conclusions of this article, as well as all code, are available in the GitHub repository, https://github.com/sdlbroer/DRMST_empirical_power.

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]Leiden University, Leiden, the Netherlands. [2]MRC Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, London, UK. [3]Edwards Lifesciences, Boston, Massachusetts, USA. [4]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands. [5]Mathematical Institute, Leiden, the Netherlands.

Received: 6 March 2025 Accepted: 21 April 2025
Published online: 24 May 2025

References

1. Cuzick J, Sasieni P. Interpreting the results of noninferiority trials–a review. Br J Cancer. 2022;127(10):1755–9.
2. Tweed CD, Quartagno M, Clements MN, Turner RM, Nunn AJ, Dunn DT, et al. Exploring different objectives in non-inferiority trials. BMJ. 2024;385.
3. Collett D. Modelling survival data in medical research. Chapman and Hall/CRC; 2023.
4. Royle KL, Meads D, Visser-Rogers JK, White IR, Cairns DA. How is overall survival assessed in randomised clinical trials in cancer and are subsequent treatment lines considered? A systematic review. Trials. 2023;24(1):708.
5. Blagoev KB, Wilkerson J, Fojo T. Hazard ratios in cancer clinical trials–a primer. Nat Rev Clin Oncol. 2012;9(3):178–83.
6. Kim DH, Uno H, Wei LJ. Restricted mean survival time as a measure to interpret clinical trial results. JAMA Cardiol. 2017;2(11):1179–80.
7. 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;13:1–15.
8. Tian L, Fu H, Ruberg SJ, Uno H, Wei LJ. Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. Biometrics. 2018;74(2):694–702.
9. Hernán MA. The hazards of hazard ratios. Epidemiology. 2010;21(1):13–5.
10. Quartagno M, Morris TP, Gilbert DC, Langley RE, Nankivell MG, Parmar MK, et al. A comparison of different population-level summary measures for randomised trials with time-to-event outcomes, with a focus on non-inferiority trials. Clin Trials. 2023;20(6):594–602.
11. 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.
12. Royston P, Parmar MK. 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.
13. Weir IR, Trinquart L. Design of non-inferiority randomized trials using the difference in restricted mean survival times. Clin Trials. 2018;15(5):499–508.
14. Quartagno M, Morris TP, White IR. Why restricted mean survival time methods are especially useful for non-inferiority trials. Clin Trials. 2021;18(6):743–5.
15. Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol. 2012;12:1–13.
16. Rohatgi A. Webplotdigitizer: Version 4.5. 2020. https://appsautomerisio/wpd4/. Accessed 15 May 2024.
17. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer. 2003;89(2):232–8.
18. Quartagno M. dani: Design and Analysis of Non-Inferiority trials. 2022. https://CRAN.R-project.org/package=dani. Accessed 15 May 2024.
19. R Core Team. R: A Language and Environment for Statistical Computing. Vienna; 2021. https://www.R-project.org/. Accessed 15 May 2024.
20. Jackson C. flexsurv: A Platform for Parametric Survival Modeling in R. J Stat Softw. 2016;70(8):1–33.
21. Therneau TM. A Package for Survival Analysis in R. 2024. R package version 3.7-0. https://CRAN.R-project.org/package=survival. Accessed 15 May 2024.
22. Broer SDL. DRMST empirical power. 2025. https://github.com/sdlbroer/DRMST_empirical_power. Accessed 15 May 2024.
23. Box GEP, Cox DR. An Analysis of Transformations. J R Stat Soc Ser B Stat Methodol. 1964;26(2):211–243. http://dx.doi.org/10.1111/j.2517-6161.1964.tb00553.x. Accessed 15 May 2024.
24. Li Z, Quartagno M, Böhringer S, van Geloven N. Choosing and changing the analysis scale in non-inferiority trials with a binary outcome. Clin Trials. 2022;19(1):14–21.
25. Mauri L, D'Agostino RB Sr. Challenges in the design and interpretation of noninferiority trials. N Engl J Med. 2017;377(14):1357–67.
26. Freidlin B, Hu C, Korn EL. Are restricted mean survival time methods especially useful for noninferiority trials? Clin Trials. 2021;18(2):188–96.
27. Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. Non-Inferiority Clinical Trials to Establish Effectiveness. Food and Drug Administration; 2016. https://www.fda.gov/media/78504/download. Accessed 6 Apr 2025.
28. for medicinal products for human use (CHMP) C. Guideline on the choice of the non-inferiority margin. European Medicines Agency; 2005. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-choice-non-inferiority-margin_en.pdf. Accessed 6 Apr 2025.
29. Tian L, Jin H, Uno H, Lu Y, Huang B, Anderson KM, et al. On the empirical choice of the time window for restricted mean survival time. Biometrics. 2020;76(4):1157–66.
30. Eaton A, Therneau T, Le-Rademacher J. Designing clinical trials with (restricted) mean survival time endpoint: practical considerations. Clin Trials. 2020;17(3):285–94.
31. Altorki N, Wang X, Kozono D, Watt C, Landrenau R, Wigle D, et al. Lobar or sublobar resection for peripheral stage IA non-small-cell lung cancer. N Engl J Med. 2023;388(6):489–98.

Publisher's Note