BJC British Journal of Cancer

www.nature.com/bjc

ARTICLE

Clinical Study

Assessment of proportional hazard assumption in aggregate data: a systematic review on statistical methodology in clinical trials using time-to-event endpoint

Eliana Rulli[1], Francesca Ghilotti[1,2], Elena Biagioli[1], Luca Porcu[1], Mirko Marabese[3], Maurizio D'Incalci[4], Rino Bellocco[2,5] and Valter Torri[1]

BACKGROUND: The evaluation of the proportional hazards (PH) assumption in survival analysis is an important issue when Hazard Ratio (HR) is chosen as summary measure. The aim is to assess the appropriateness of statistical methods based on the PH assumption in oncological trials.

METHODS: We selected 58 randomised controlled trials comparing at least two pharmacological treatments with a time-to-event as primary endpoint in advanced non-small-cell lung cancer. Data from Kaplan–Meier curves were used to calculate the relative hazard at each time point and the Restricted Mean Survival Time (RMST). The PH assumption was assessed with a fixed-effect meta-regression.

RESULTS: In 19% of the trials, there was evidence of non-PH. Comparison of treatments with different mechanisms of action was associated ($P = 0.006$) with violation of the PH assumption. In all the superiority trials where non-PH was detected, the conclusions using the RMST corresponded to that based on the Cox model, although the magnitude of the effect given by the HR was systematically greater than the one from the RMST ratio.

CONCLUSION: As drugs with new mechanisms of action are being increasingly employed, particular attention should be paid on the statistical methods used to compare different types of agents.

British Journal of Cancer (2018) 119:1456–1463; https://doi.org/10.1038/s41416-018-0302-8 Presentation:This work has been presented at the 2015 Italian Stata Users Group meeting in Florence on 12 November 2015.

INTRODUCTION

In many clinical and observational studies, especially in oncology, the quantity of main interest is the length of time before an event occurs. In oncology, especially in phase III trials, the outcomes of interest are death, progression or relapse of the disease. In this setting, a time-to-event endpoint is used and survival analysis is performed to analyse the data.

Different methods can be used for survival analysis. Among the non-parametric methods, Kaplan–Meier (KM) estimator is the most common.[1] The Log-rank test can be used to evaluate whether KM curves are statistically different. The Cox proportional hazards (PH) model[2] is the most common approach[3–5] to detect and estimate the effect of several risk factors on survival. The measure of association estimated by the Cox PH model is the hazard ratio (HR), which is, with two treatment groups, the ratio of the hazard of the outcome of interest in the treated to the control group. The hazard rate represents the instantaneous risk of the event of interest occurrence. The Cox model does not require any parametric assumptions about the shape of the baseline hazard function, but relies on the proportionality of the hazards, so the HR is assumed constant over time. When the PH assumption fails, the HR estimated by the Cox model depends on follow-up time[6,7] and it has also been seen that, under these circumstances, the Log-rank test has a lower power.[8]

When the PH assumption is not met, the effectiveness estimates are likely to be not representative for the whole intervention period and the effect of this bias, in case of meta-analyses, will be carry over to the analyses of aggregate data. Although survival curves convergence and crossings are common in medical research,[9,10] too little attention is paid to this issue and the statistical test of PH assumption is rarely reported in clinical trial publications. It has been reported that crossing survival curves are common when one of the treatment compared offers a short-term benefit, but no long-term advantages[8] or when the treatments being compared have different biological mechanisms of action, or differently responsive sub-populations are included.[11] When individual patient data are available, there are several options for assessing the PH assumption.[12,13] However, no ready to use methods exist to test the validity of PH assumption when only aggregate data are available. Testing the PH assumption on aggregate data is therefore needed to guarantee the validity of meta-analysis results.

[1]Laboratory of Methodology for Clinical Research, Oncology Department, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy; [2]Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, Italy; [3]Laboratory of Molecular Pharmacology, Oncology Department, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy; [4]Laboratory of Cancer Pharmacology, Oncology Department, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan, Italy and [5]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
Correspondence: Eliana Rulli (eliana.rulli@marionegri.it)

Received: 23 February 2018 Revised: 18 September 2018 Accepted: 21 September 2018
Published online: 13 November 2018

Several methods exist to analyse time-to-event endpoints when the HR is not an adequate summary statistic for treatment effect.[14,15] The Restricted Mean Survival Time (RMST) is an alternative to the HR.[11,14] The RMST does not need any model assumptions, such as the hazards to be proportional, and is readily interpretable as 'life expectancy' between the time of randomisation and a relevant time point.

We conducted a systematic review of randomised phase II and III clinical trials comparing different types of pharmacological treatment in patients with advanced non-small cell lung cancer (NSCLC) with the aim of assessing the appropriateness of survival analysis based on the PH assumption. A method to test the PH assumption on aggregate data was proposed and factors influencing it were also investigated.

In the absence of PH, the Cox model results were compared to those based on the RMST to investigate the robustness of the conclusions drawn by looking at the biased average HR estimate.

The choice of focusing our analysis on the oncological area was driven by the publication in the last years of some important trials where the PH assumption clearly failed (e.g. IPASS[16] and ICON7[17]).

We focused on NSCLC since this is the most common cause of cancer deaths worldwide[18] and although recent preclinical studies have improved the knowledge of the molecular mechanism governing the cancer cell, the majority of oncologic patients still do not benefit from new clinical therapy. NSCLC therefore represents a research area where the development of drugs with new mechanisms of action is very active.

METHODS

Eligibility criteria

All phase II and III randomised controlled trials (RCTs) published from January 2004 to January 2015 comparing two or more systemic therapies in patients with advanced NSCLC were considered. Studies using a time-to-event primary endpoint, such as overall survival (OS), progression-free survival (PFS) and time-to-progression (TTP) were eligible. We excluded studies assessing different strategies for the same agent (i.e. different schedules or doses), using placebo, radiotherapy or surgery alone as comparator. Studies terminated early were also excluded.

Search strategy

We searched Medline and Embase databases using search terms including non-small cell lung cancer, randomised clinical trials, chemotherapy, vascular endothelial growth factor receptor inhibitor, tyrosine-kinase inhibitor and a list of approved drugs for NSCLC. The complete search strategy is available in Supplementary Table S1. Two reviewers independently evaluated both the titles and the abstracts to ensure eligibility. Full manuscripts of potentially eligible trials were read to identify studies to be included. A third reviewer solved disagreements.

Data extraction

Two reviewers, using a data extraction form, independently recorded the study design, patient and treatment characteristics, methodological and statistical features. As above, a third reviewer, who referred back to the original article, solved any differences in data extraction.

Treatments and comparisons

According to their mechanism of action, treatments were classified in four categories:[1] conventional therapy (drugs causing DNA damage or inhibition of DNA synthesis),[2] biologics (antibodies against growth and angiogenic factors),[3] tyrosine-kinase inhibitor (TKI) and[4] miscellaneous group. The latter includes inhibitors of kinase other than tyrosine-kinase (i.e. protein kinase alpha and cand PI-3 kinase), metalloproteinase inhibitors, inhibitors of proapoptotic proteins and Toll-like receptor-9 agonist.

Studies were classified as those comparing treatments belonging to different categories or those comparing treatments within the same category.

Statistical methods

Only papers reporting KM curves with patients at risk at more than two time points and separately for each treatment group were included in the statistical analysis; the papers not fulfilling this condition were only described. For each study analysed, the number of event-free patients at each time point and the respective survival probabilities were extracted from the published KM curves. Then, during each interval, we estimated the number of patients at risk, the number of censored patients and the number of events. These items were used to estimate the ln (HR) and its variance according to the methodology proposed by Williamson.[19] The PH assumption was tested graphically using a plot of the log cumulative hazard, where the logarithm of time is plotted against the estimated log cumulative hazard calculated as ln[-ln (S(t))].[20] If the curves for the two treatment groups were approximately parallel, the PH assumption was deemed reasonable. The graphical method is useful for visualising clear departures from the PH assumption, but due to the arbitrary assessment of these plots, a formal test was then applied. For each study, the relationship between the ln(HR) estimate at each time point and the follow-up time was described by a forest plot. To formally assess PH assumption a fixed-effect meta-regression within each study was conducted. The ln(HR) at each time point was the outcome of interest and the follow-up time was included in the model as the only covariate. It has been previously showed that, within each study, the HR(t) for $t = 1,\ldots,p$ are independent[21] and therefore log(HR) estimates at different time points were treated as being obtained from different studies, as it is the case in the classic setting of meta-regression. If the ln(HR) was found to change over the follow-up time, as a result of a statistically significant estimate in meta-regression, the PH hypothesis was rejected. For the articles with co-primary endpoints, only one of them was included in the analysis. The endpoint with a statistically significant result was chosen; if none reported a significant result, the one with most patients at risk at the beginning of the follow-up was considered. For articles with more than two treatment arms, the PH assumption was tested separately for each comparison. The RMST was obtained for each study calculating the area under the KM curve for each interval using the method of trapezoids. The RMST was calculated up to last available follow-up time, defined as the greatest time point with still patients at risk reported under the KM curve in each arm. The follow-up time chosen was the same for all the treatment arms within study, unless one of the two treatment arms being compared reaches a value of zero in the survival curve. In this case the RMST of the other arm was calculated beyond the time point in which the first curve reaches the 0, until its last available time point. The RMST variance was calculated according to Klein et al.[22] Differences in RMST between arms, the ratio and the relative Z test to assess the statistical significance of the difference were computed. Fisher's exact test was used to investigate the association between absence of PH and study characteristics such as treatment comparisons (different categories vs. same category), endpoint (OS vs. other) and study results (significant results vs. non-significant). The two-sided significance level was set at 0.05 to test associations with study characteristics and to test RMST difference, and at 0.10 for meta-regression analysis. We used the web-based tool WebPlotDigitizer available online at https://automeris.io/WebPlotDigitizer/ to extract data from the KM curves.

Meta-regression was conducted in Stata (version 13) with the *vwls* variance-weighted least squares command.[23]

RESULTS

The databases search identified 1078 records. Of these, 882 were excluded on the basis of the abstract evaluation, and the full text was obtained for the remaining 196. One additional paper was identified from the references of selected articles and added to the list of articles to be screened because potentially eligible. Full-text review led to the exclusion of 82 studies, not meeting the inclusion criteria: 32 did not have a time-to-event endpoint, 22 had a different aim, in 11 the control group received placebo or no therapy, for 8 the trial was terminated prematurely (3 trials were terminated due to low recruitment, 2 were stopped for futility, 2 for a high rate of unexpected mortality and toxicity, 1 for new evidences external from the trial), 2 were study protocols, 5 were updates of other articles included with no additional information, for 2 the full text was not available. Since 52 of the 115 articles we selected for data extraction did not report the number of patients at risk and 5 gave only two time points, survival data were extracted from the remaining 58 articles (Supplementary Figure S1).[16,24–80] Four of these had co-primary endpoints[47,62,69,76] and four had three treatment arms.[26,64,77,81] Details of the 58 trials regarding settings, interventions and methodological characteristics are shown in Supplementary Tables S2 and S3.

Study characteristics

Table 1 summarises the 115 studies in the review, according to their inclusion/exclusion from the statistical analysis. For 56 (49%) studies, OS was the primary endpoint, while for 52 (45%) PFS was used. The median number of randomised patients was 332. It is worth noting that 12 (10%) trials randomised fewer than 100 participants. Out of 115 articles, 102 (89%) reported the sample size. Among these, only 70 (69%) articles reported the number of events needed to reach the desired statistical power. The remaining 32 (31%) articles only reported the total number of patients to be included. Of the 70 (61%) articles reporting the number of events required, only 49 (70%) described the number of events reached. In particular 40 (82%) of these reported a number of events observed greater than 95% of those planned, while 9 (18%) did not reach the 95% of events required. Only four (3%) out of 115 studies reported that the PH assumption was tested and no evidence of failure of PH was seen in any of these studies. In the NVALT-10 study[82] the PH assumption was assessed from scaled Schoenfeld residuals. In the TAILOR study[32] the PH assumption was verified using graphic plots of Schoenfeld residuals over time, and by adding time-dependent variables in the model to test their statistical significance. In the FASTACT-2 trial[79] the PH assumption was assessed graphically by plotting log–log survival functions for the two treatment groups. In the last trial[83] the PH assumption was informally assessed by simply looking at the survival functions. In a further study[41] the absence of PH was mentioned but the method used to test it was not described. From 115 studies, 128 comparisons were obtained. According to our classification, 53 (41%) involved treatments with the same mechanism of action while 75 (59%) compared treatments with different mechanisms (Table 2).

PH assumption assessment

The median number of time points with the corresponding number of patients at risk reported under the KM curves was six (Interquartile range (IQR) 4–9). The median decrease in the number of patients at risk from the start of the observation period to the first time point reported was 34% (IQR 18–49%). According to our meta-regression analysis, in 12 (19%) out of 62 treatment comparisons the PH assumption was violated.[16,25,26,38,41,49,52,66,71,72,75,78]

Supplementary Figures S2, S3 and S4 report all the log–log plots and the forest plots of these 12 articles. For illustrative purposes Fig. 1 reports two studies, in the first the PH assumption was violated,[25] while in the second the PH assumption was not rejected.[42]

Comparisons of treatments with a different mechanisms of action were significantly associated ($P = 0.006$) with violation of PH assumption (Table 3). Ten (83%) studies in which the assumption was violated and 27 (54%) in which the PH was met, had PFS as primary endpoint, but this difference did not

Table 1. Characteristics of the studies included in the review

	Excluded from statistical analysis ($N = 57$)		Included in statistical analysis ($N = 58$)		Total ($N = 115$)	
	N	%	*N*	%	*N*	%
Phase						
I–II	1	2	0	0	1	1
II	19	33	19	33	38	33
II–III	1	2	1	2	2	2
III	36	63	38	65	74	64
Type of study—centre						
Multicentre	43	75	57	98	100	87
Single centre	14	25	1	2	15	13
Blinding						
Yes	10	18	22	38	32	28
No	47	82	36	62	83	72
Primary Endpoint						
OS	32	56	24	41	56	49
PFS	20	35	32	55	52	45
TTF	1	2	0	0	1	1
TTP	4	7	2	4	6	5
Proportionality assessed						
Yes	2[a]	4	2	3	4	3
No	55	96	56	97	111	97
Sample size calculation						
Number of events reported	25	44	45	78	70	61
Only number of patients reported	23	40	9	15	32	28
Not provided	9	16	4	7	13	11
Reached >95% of target events						
Yes	14	82	26	81	40	82
No	3	18	6	19	9	18
Number of patients analysed						
Median	302		379		332	
IQR	154–440		175–772		168–595	
Minimum–Maximum	48–1725		60–1433		48–1725	

[a]In one study the proportionality assumption was informally assessed by looking at the survival functions

N number, *OS* overall survival, *PFS* progression-free survival, *TTF* time to failure, *TTP* time to progression, *IQR* interquartile range

reach statistical significance ($P = 0.101$). Among the nine superiority trials in which absence of PH was detected, seven (78%) gave statistically significant results ($P = 0.069$). Among the seven non-inferiority trials, four (57%) satisfied the PH assumption. In the latter, the experimental treatment was non-inferior to the control arm. In the other three (43%), the PH assumption was rejected and the experimental arm could not be considered non-inferior to the control ($P = 0.029$).

Restricted mean survival time

HR, RMST (difference and ratio) and the relative statistical tests for the nine superiority trials in which the PH assumption was rejected, are reported in Table 4. In all these studies, results with the RMST difference corresponded to the conclusions drawn by the authors, based on Cox models as far as the statistical significance is concerned. Though the HR and the RMST ratio have different meanings in quantifying differences between arms, the two measures of the magnitude of treatment effect are on the same relative scale and can be easily compared. When the HR is plotted against the RMST ratio (Figure S5), it can be observed that the magnitude of the treatment effect given by the HR is systematically greater than the RMST ratio. The tendency does not seem related to the PH assumption violation as shown in the two regression lines in Figure S5.

Table 2. Treatment comparisons investigated by the articles included in the review

	Excluded from statistical analysis		Included in statistical analysis		Total	
	N	%	*N*	%	*N*	%
Same treatment comparison	33	50	20	32	53	41
1 vs 1	30	91	15	75	45	85
3 vs 3	3	9	5	25	8	15
Different treatments comparison	33	50	42	68	75	59
1 vs 3	5	15	13	31	18	24
1 vs 1 + 2	6	18	9	21	15	20
1 vs 1 + 3	12	37	14	34	26	35
1 vs 1 + 4	6	18	0	0	6	8
1 vs 1 + 2 + 3	1	3	0	0	1	1
2 vs 1 + 2	0	0	2	5	2	3
2 vs 2 + 3	0	0	1	2	1	1
3 vs 1 + 3	2	6	0	0	2	3
3 vs 2 + 3	0	0	3	7	3	4
3 vs 4	1	3	0	0	1	1
Total	66	100	62	100	128	100

1 conventional therapy (drugs causing DNA damage or inhibition of DNA synthesis), *2* biologics (antibodies against growth and angiogenic factors), *3* tyrosine-kinase inhibitor (TKI), *4* miscellaneous group

Table 3. Association between proportional hazard assumption results and study characteristics

	PH assumption violated		*P* Fisher's exact test
	No	Yes	
Treatments			
Same treatment comparison	20 (40%)	0 (0%)	
Different treatments comparison	30 (60%)	12 (100%)	.006
Primary endpoint			
OS	23 (46%)	2 (17%)	
PFS/TTP/TTF	27 (54%)	10 (83%)	.101
Superiority trial			
Positive result (superiority demostrated)	19 (41%)	7 (78%)	
Negative result (superiority not demostrated)	27 (59%)	2 (22%)	.069
Non-inferiority trial			
Positive result (non-inferiority demonstrated)	4 (100%)	0 (0%)	
Negative result (non-inferiority not demonstrated)	0 (0%)	3 (100%)	.029

PH proportional hazard, *OS* overall survival, *PFS* progression-free survival, *TTP* time to progression, *TTF* time to failure

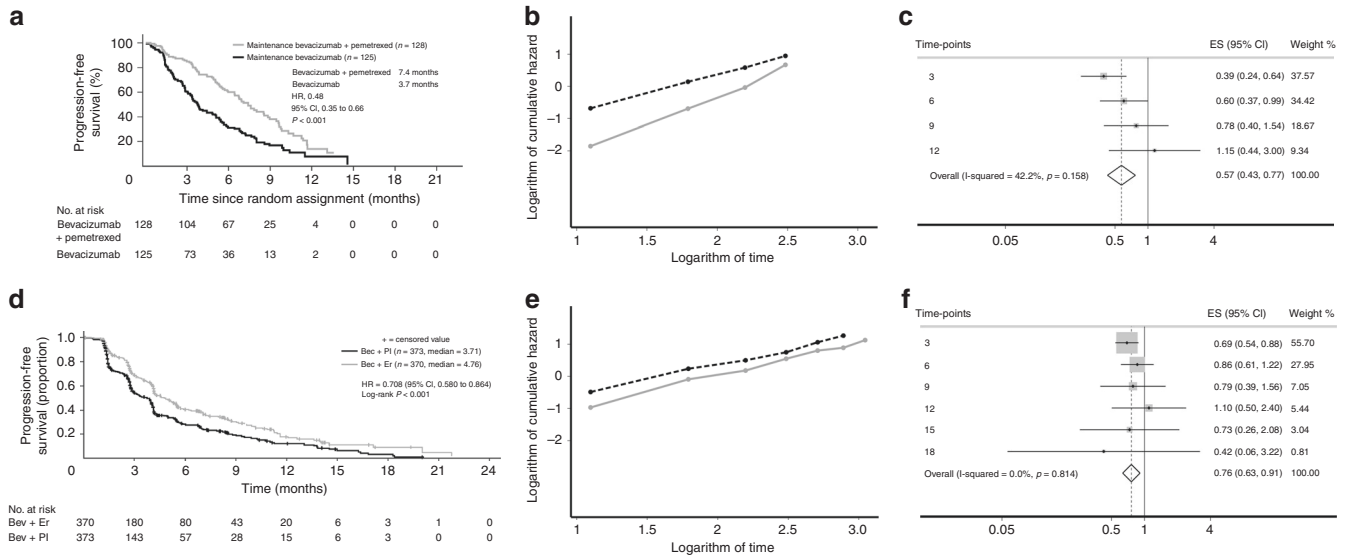

Fig. 1 **a**–**c** Example in which proportional hazard assumption is violated:[25] **a** Published KM curves; **b** Log–log plot; **c** Forest plot. **d**–**f** example in which proportional hazard assumption is verified:[42] **d** Published KM curves; **e** Log–log plot; **f** Forest plot

Table 4. Comparison of the RMST results and the HR results in studies with PH assumption violated

RMST results						HR results	
Study	Control arm (months)	Experimental arm (months)	Difference (months)	*P* test Z	Ratio[a] 95%CI	HR	*P*
Belani[26]	9.51	10.15	0.65	0.730	0.94 (0.65–1.36)	0.89	0.360
Reck[66]	3.42	4.18	0.76	0.018	0.82 (0.69–0.97)	0.79	0.002
Lee[49]	3.70	4.92	1.23	0.150	0.76 (0.51–1.10)	0.73	0.040[b]
Janne[41]	8.82	9.65	0.82	0.610	0.91(0.65–1.30)	0.80	0.210[b]
Barlesi[25]	4.88	7.24	2.37	<0.001	0.67 (0.53–0.86)	0.48	<0.001
Shaw[72]	5.95	9.27	3.33	0.004	0.64 (0.47–0.88)	0.49	<0.001
Seto[71]	11.35	16.48	5.13	<0.001	0.69 (0.55–0.87)	0.54	0.002
Solomon[75]	8.00	14.13	6.13	<0.001	0.57 (0.45–0.72)	0.45	<0.001
Wu[78]	5.63	12.29	6.66	<0.001	0.46 (0.37–0.57)	0.28	<0.001

RMST restricted mean survival times, *HR* hazard ratio, *KM* Kaplan–Meier
[a]Calculated as ratio between RMST in the control and RMST in the experimental arm
[b]One-sided, 95%CI, 95% confidence intervals

DISCUSSION

An important finding that emerged from this analysis was the significant association between the kind of treatments compared and the absence of PH. New oncological treatments were frequently compared to a conventional therapy with a different mechanism of action. This is frequently reflected in a different course of the disease progression and might explain why, when treatments with different mechanisms are compared, the hazards are not proportional. As drugs with different mechanisms of action are increasingly investigated, due attention must be paid to the statistical methods used in these circumstances.

From our review stands out that only 3% of the studies analysed reported or mentioned the PH test. In these studies, the conclusion drawn by authors were concordant with the result from meta-regression testing. In one study,[41] where it was found that hazards were nonproportional, they still decided to report and interpret HR without discussing possible implications. The proposed method showed that in 12 out of 56 treatment comparisons (19%) the PH assumption was not satisfied. This means that the statistical methods applied to analyse the treatment effect might be not adequate. When the PH assumption fails, the RMST difference can be used as a primary endpoint as it does not require hazards to be proportional. Even when the PH assumption appears to be satisfied, RMST may be a useful secondary measure because it gives a different, but complementary information.

We acknowledge that RMST also has some limitations. Since it depends on the time point chosen to calculate it, an inappropriate choice may give misleading results. A further limit, in the study design setting, concerns the within-group variances hypothesis for sample size calculation. However, when comparing conclusions based on the Cox model to those based on the RMST difference no discrepancy was observed.

Trinquart et al.[84] analysed 54 RCTs, reconstructing individual patient data and calculating both the HR and the RMST ratio for each trial. The results obtained with these two measures were consistent and this behaviour was independent of the presence or absence of PH. Despite the agreement between RMST difference and HR on the statistical significance of the treatment effect, they provided empirical evidence that the treatment effect measure based on RMST yielded more conservative estimates than the ones based on HR. We found the same behaviour, with HR systematically overestimating the treatment effect compared to the RMST ratio.

In our review, the comparisons of treatments with different mechanisms of action involved conventional therapies compared with TKI or biological therapies. In the last few years, immunotherapy has emerged as a promising therapeutic strategy and has radically transformed the therapeutic landscape for NSCLC.[85,86] In this setting too the problem of non-PH can be expected since immunotherapy efficacy translates into long-term survival and delayed clinical effects.[87] When conventional therapy is compared with immunotherapy, in case of non-PH an underestimation of treatment effect can be expected when HR is used to measure it. When the PH assumption is unmet, based on our findings the Cox model still seems to bring to the right conclusions. It is important to note, however, that the estimate obtained becomes time-dependent and might not appropriately describe the phenomenon investigated.

Here we propose a method for testing the PH assumption when only aggregate data are available. The suggested method relies on published KM curves and, after data have been extracted, meta-regression is used to assess the PH assumption by testing for a linear trend of HRs with time. Systematic reviews and meta-analyses of well-designed and executed RCTs have the potential to provide the highest levels of evidence to support diagnostic and therapeutic decisions. To guarantee the accuracy of the meta-analysis results, however, it is important to assess the PH assumption for each considered trial and to evaluate the impact of the inclusion of trials not satisfying the assumption on the meta-analysis results. Even if the authors of the original study have not checked whether the PH assumption is satisfied or not, the method proposed here can be applied. Since a pooled analysis of studies with non-PH can produce an over- or underestimation of the efficacy comparison, this tool could be very useful when the aim is to conduct a meta-analysis. In case of non-PH detection in one or more than one study to be included in the meta-analysis different approaches might be adopted: e.g. one possibility is to include all the studies in the meta-analysis and run sensitivity analyses excluding non-PH studies; otherwise one can use alternative summary measures beyond HR which do not require the proportionality of the hazards such as survival at time *t* or RMST.

The main limit of the method is that it depends on the quality of published KM curves. Moreover, to estimate the number of events and censored in each interval an assumption must be made about the censoring mechanism. Censoring is assumed to be constant within each interval. The smaller the interval, the more likely the

assumption will be met. Once again, the size of the interval depends on the number of time points reported under the KM curve. If these are relatively few, an appropriate assessment of the PH assumption is not possible, leading to low statistical power of the test. It is worth investigating how many time points are needed to draw appropriate conclusions with regard to the PH assumption. Future developments include comparisons of the conclusions obtained with individual patient data using the consolidated methods and those from aggregate data using meta-regression.

Guyot et al.[88] recently proposed a method that allows simulating (or approximating) patient-level data based on KM curves. The advantage of using simulated patient-level data is that once simulated patient-level data are obtained, standard methods to assess the PH assumption can be used. This probably leads to a better chance of detecting the departure from the PH assumption, if any, compared to our method which could be less sensitive. Even if our method and that proposed by Guyot were presented as tools for testing PH assessment from published data, the aim and the context of application can be different. The method proposed by Guyot is very helpful and necessary when the aim of the meta-analysis is to pool survival data between studies or to analyse data with different survival models than the ones used in the published papers. The performance of this method relies on the information reported in the original papers. The authors stated that if the total number of events and the numbers at risk other than at time zero are not provided, the algorithm may produce poor results. Furthermore, it is a time-consuming process: it takes about half an hour to obtain the initial input data for one KM curve (i.e. one hour for each study in case of a two arm trial). Our aim was to propose a simpler method to test PH assumption from aggregate data when performing a meta-analysis. In this setting, it is very important to have an easy and quick to use tool. The meta-regression can be considered a valid alternative to the method proposed by Guyot because it does not require paid software and the time required to complete the process is less. However, a formal comparison between the two methods would be useful to investigate their properties in different settings.

A further limit of our work is that we applied our method in only 45% of eligible papers. This was because of the missing information on patients at risk at different time points. Parmar et al.[21] proposed methods to estimate not only the number of censored and the number of events in each interval, but also the number of patients at risk. These methods, however, require further assumptions resulting in estimates even more approximated. For this reason, we decided to minimise the number of assumptions to be made, including only papers with all information available.

The conclusions from the meta-regression were compared to the results from the consolidated graphic method in which the logarithm of the time is plotted against the estimated log cumulative hazard. This was done in order to analyse the appropriateness of conclusions drawn according to meta-regression results. When the sample size is small, this method may lack power to detect deviations from PH; while for large sample sizes, hypothesis tests may be over sensitive to slight deviations from this assumption. In our sample, the results obtained with the meta-regression were in line with the log–log plots in all the studies. Despite the subjectivity of the graphic method, in all the 12 studies in which the PH assumption was rejected the graph agrees with the test results, indeed the curves were clearly not parallel. Moreover, for the four trials in which the authors tested the PH assumption, the conclusions with meta-regression were the same as the ones reported in the original papers.

Although further investigations are needed, the results we observed do suggest that meta-regression is a valid method for testing the PH assumption when only aggregate data are available.

AUTHORS' CONTRIBUTION

Study conception and design: E.R., V.T. Acquisition of data: E.R., F.G., E.B., M.M. Analysis and interpretation of data: E.R., F.G., E.B., L.P., M.D'.I., R.B. Drafting of manuscript: E.R., M.M. Critical revision; E.R., F.G., E.B., L.P., M.M., M.D'.I., R.B., V.T.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41416-018-0302-8.

Competing interests: The authors declare no competing interests.

Funding: None.

REFERENCES

1. Kaplan, E. M. P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).
2. Cox, D. Regression models and life-tables. *J. Royal Stat. Soc. Ser B (Method)* **34**, 187–220 (1972).
3. Breslow, N. Analysis of survival data under the proportional hazards model. *Int. Stat. Rev.* **43**, 45–57 (1975).
4. Breslow, N. Statistical methods for censored survival data. *Environ. Health Perspect.* **32**, 181–192 (1979).
5. Blagoev, K. B., Wilkerson, J. & Fojo, T. Hazard ratios in cancer clinical trials—a primer. *Nat. Rev. Clin. Oncol.* **9**, 178–183 (2012).
6. Prentice, R. L., Pettinger, M. & Anderson, G. L. Statistical issues arising in the Women's Health Initiative. *Biometrics* **61**, 899–911 (2005).
7. Hernan, M. A. The hazards of hazard ratios. *Epidemiology* **21**, 13–15 (2010).
8. Li, H., Han, D., Hou, Y., Chen, H. & Chen, Z. Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS ONE* **10**, e0116774 (2015).
9. Zewdu A. Survival Curve Convergences and Crossings: How frequent are they in medical research? A Study of five medical journals. https://www.duo.uio.no/handle/10852/30384. accessed 30 october 2018.
10. Kristiansen, I. PRM39 survival curve convergences and crossing: a threat to validity of meta-analysis? *Value Health* **15**, A652 (2012).
11. Royston, P. & Parmar, M. K. 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.* **13**, 152 (2013).
12. Grambsch, P. T. T. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526 (1994).
13. Stablein, D. M., Carter, W. H. Jr. & Novak, J. W. Analysis of survival data with nonproportional hazard functions. *Control Clin. Trials* **2**, 149–159 (1981).
14. Royston, P. & Parmar, M. K. 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.* **30**, 2409–2421 (2011).
15. Royston, P. & Parmar, M. K. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. *Trials* **15**, 314 (2014).
16. Mok, T. S. et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N. Engl. J. Med.* **361**, 947–957 (2009).
17. Oza, A. M. et al. Standard chemotherapy with or without bevacizumab for women with newly diagnosed ovarian cancer (ICON7): overall survival results of a phase 3 randomised trial. *Lancet Oncol.* **16**, 928–936 (2015).
18. GLOBOCAN 2012 v1.0, *Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11*. (International Agency for Research on Cancer, Lyon, France, 2013).
19. Williamson, P. R., Smith, C. T., Hutton, J. L. & Marson, A. G. Aggregate data meta-analysis with time-to-event outcomes. *Stat. Med.* **21**, 3337–3351 (2002).
20. Kleinbaum D. K. M. Survival Analysis—A self-learning text. *Statistics for Biology and Health*. Third ed. (Springer, New York, 2012).
21. Parmar, M. K., Torri, V. & Stewart, L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat. Med.* **17**, 2815–2834 (1998).
22. Klein J. M. M. L. *Survival analysis: techniques for censored and truncated data*. (Springer-Verlag, New York, 2003).
23. Harbord, R. M. H. & Julian, P. T. Meta-regression in Stata. *Stata J.* **8**, 493–519 (2008).

24. Ardizzoni, A. et al. Pemetrexed versus pemetrexed and carboplatin as second-line chemotherapy in advanced non-small-cell lung cancer: results of the GOIRC 02-2006 randomized phase II study and pooled analysis with the NVALT7 trial. *J. Clin. Oncol.* **30**, 4501–4507 (2012).
25. Barlesi, F. et al. Randomized phase III trial of maintenance bevacizumab with or without pemetrexed after first-line induction with bevacizumab, cisplatin, and pemetrexed in advanced nonsquamous non-small-cell lung cancer: AVAPERL (MO22089). *J. Clin. Oncol.* **31**, 3004–3011 (2013).
26. Belani, C. P. et al. Randomized phase II study of pemetrexed/cisplatin with or without axitinib for non-squamous non-small-cell lung cancer. *BMC Cancer* **14**, 290 (2014).
27. Chouaid, C., Nathan, F., Pemberton, K. & Morris, T. A phase II, randomized, multicenter study to assess the efficacy, safety, and tolerability of zibotentan (ZD4054) in combination with pemetrexed in patients with advanced non-small cell lung cancer. *Cancer Chemother. Pharmacol.* **67**, 1203–1208 (2011).
28. de Boer, R. H. et al. Vandetanib plus pemetrexed for the second-line treatment of advanced non-small-cell lung cancer: a randomized, double-blind phase III trial. *J. Clin. Oncol.* **29**, 1067–1074 (2011).
29. Dittrich, C. et al. A randomised phase II study of pemetrexed versus pemetrexed + erlotinib as second-line treatment for locally advanced or metastatic non-squamous non-small cell lung cancer. *Eur. J. Cancer* **50**, 1571–1580 (2014).
30. Edelman, M. J. et al. Randomized phase II study of ixabepilone or paclitaxel plus carboplatin in patients with non-small-cell lung cancer prospectively stratified by beta-3 tubulin status. *J. Clin. Oncol.* **31**, 1990–1996 (2013).
31. Flotten, O. et al. Vinorelbine and gemcitabine vs vinorelbine and carboplatin as first-line treatment of advanced NSCLC. A phase III randomised controlled trial by the Norwegian Lung Cancer Study Group. *Br. J. Cancer* **107**, 442–447 (2012).
32. Garassino, M. C. et al. Erlotinib versus docetaxel as second-line treatment of patients with advanced non-small-cell lung cancer and wild-type EGFR tumours (TAILOR): a randomised controlled trial. *Lancet Oncol.* **14**, 981–988 (2013).
33. Garon, E. B. et al. Ramucirumab plus docetaxel versus placebo plus docetaxel for second-line treatment of stage IV non-small-cell lung cancer after disease progression on platinum-based therapy (REVEL): a multicentre, double-blind, randomised phase 3 trial. *Lancet* **384**, 665–673 (2014).
34. Gridelli, C. et al. Single-agent pemetrexed or sequential pemetrexed/gemcitabine as front-line treatment of advanced non-small cell lung cancer in elderly patients or patients ineligible for platinum-based chemotherapy: a multicenter, randomized, phase II trial. *J. Thorac. Oncol.* **2**, 221–229 (2007).
35. Gridelli, C. et al. Phase II randomized study of vandetanib plus gemcitabine or gemcitabine plus placebo as first-line treatment of advanced non-small-cell lung cancer in elderly patients. *J. Thorac. Oncol.* **9**, 733–737 (2014).
36. Groen, H. J. et al. A randomized, double-blind, phase II study of erlotinib with or without sunitinib for the second-line treatment of metastatic non-small-cell lung cancer (NSCLC). *Ann. Oncol.* **24**, 2382–2389 (2013).
37. Hanna, N. et al. Randomized phase III trial of pemetrexed versus docetaxel in patients with non-small-cell lung cancer previously treated with chemotherapy. *J. Clin. Oncol.* **22**, 1589–1597 (2004).
38. Heigener, D. F. et al. Open, randomized, multi-center phase II study comparing efficacy and tolerability of Erlotinib vs. Carboplatin/Vinorelbin in elderly patients (> 70 years of age) with untreated non-small cell lung cancer. *Lung Cancer* **84**, 62–66 (2014).
39. Herbst, R. S. et al. Efficacy of bevacizumab plus erlotinib versus erlotinib alone in advanced non-small-cell lung cancer after failure of standard first-line chemotherapy (BeTa): a double-blind, placebo-controlled, phase 3 trial. *Lancet* **377**, 1846–1854 (2011).
40. Herbst, R. S. et al. Vandetanib plus docetaxel versus docetaxel as second-line treatment for patients with advanced non-small-cell lung cancer (ZODIAC): a double-blind, randomised, phase 3 trial. *Lancet Oncol.* **11**, 619–626 (2010).
41. Janne, P. A. et al. Selumetinib plus docetaxel for KRAS-mutant advanced non-small-cell lung cancer: a randomised, multicentre, placebo-controlled, phase 2 study. *Lancet Oncol.* **14**, 38–47 (2013).
42. Johnson, B. E. et al. ATLAS: randomized, double-blind, placebo-controlled, phase IIIB trial comparing bevacizumab therapy with or without erlotinib, after completion of chemotherapy, with bevacizumab for first-line treatment of advanced non-small-cell lung cancer. *J. Clin. Oncol.* **31**, 3926–3934 (2013).
43. Karampeazis, A. et al. Pemetrexed versus erlotinib in pretreated patients with advanced non-small cell lung cancer: a Hellenic Oncology Research Group (HORG) randomized phase 3 study. *Cancer* **119**, 2754–2764 (2013).
44. Kawaguchi, T. et al. Randomized phase III trial of erlotinib versus docetaxel as second- or third-line therapy in patients with advanced non-small-cell lung cancer: Docetaxel and Erlotinib Lung Cancer Trial (DELTA). *J. Clin. Oncol.* **32**, 1902–1908 (2014).
45. Kelly, K. et al. Randomized phase 2b study of pralatrexate versus erlotinib in patients with stage IIIB/IV non-small-cell lung cancer (NSCLC) after failure of prior platinum-based therapy. *J. Thorac. Oncol.* **7**, 1041–1048 (2012).
46. Kim, E. S. et al. Docetaxel or pemetrexed with or without cetuximab in recurrent or progressive non-small-cell lung cancer after platinum-based therapy: a phase 3, open-label, randomised trial. *Lancet Oncol.* **14**, 1326–1336 (2013).
47. Kim, E. S. et al. Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (INTEREST): a randomised phase III trial. *Lancet* **372**, 1809–1818 (2008).
48. Kubota, K. et al. Vinorelbine plus gemcitabine followed by docetaxel versus carboplatin plus paclitaxel in patients with advanced non-small-cell lung cancer: a randomised, open-label, phase III study. *Lancet Oncol.* **9**, 1135–1142 (2008).
49. Lee, D. H. et al. Randomized Phase III trial of gefitinib versus docetaxel in non-small cell lung cancer patients who have previously received platinum-based chemotherapy. *Clin. Cancer Res.* **16**, 1307–1314 (2010).
50. Li, N. et al. Pemetrexed-carboplatin adjuvant chemotherapy with or without gefitinib in resected stage IIIA-N2 non-small cell lung cancer harbouring EGFR mutations: a randomized, phase II study. *Ann. Surg. Oncol.* **21**, 2091–2096 (2014).
51. Lilenbaum, R. C. et al. Single-agent versus combination chemotherapy in advanced non-small-cell lung cancer: the cancer and leukemia group B (study 9730). *J. Clin. Oncol.* **23**, 190–196 (2005).
52. Maruyama, R. et al. Phase III study, V-15-32, of gefitinib versus docetaxel in previously treated Japanese patients with non-small-cell lung cancer. *J. Clin. Oncol.* **26**, 4244–4252 (2008).
53. Natale, R. B. et al. Phase III trial of vandetanib compared with erlotinib in patients with previously treated advanced non-small-cell lung cancer. *J. Clin. Oncol.* **29**, 1059–1066 (2011).
54. Niho, S. et al. Randomized phase II study of first-line carboplatin-paclitaxel with or without bevacizumab in Japanese patients with advanced non-squamous non-small-cell lung cancer. *Lung Cancer* **76**, 362–367 (2012).
55. Okamoto, I. et al. Phase III trial comparing oral S-1 plus carboplatin with paclitaxel plus carboplatin in chemotherapy-naive patients with advanced non-small-cell lung cancer: results of a west Japan oncology group study. *J. Clin. Oncol.* **28**, 5240–5246 (2010).
56. Paccagnella, A. et al. Adding gemcitabine to paclitaxel/carboplatin combination increases survival in advanced non-small-cell lung cancer: results of a phase II-III study. *J. Clin. Oncol.* **24**, 681–687 (2006).
57. Patel, J. D. et al. PointBreak: a randomized phase III study of pemetrexed plus carboplatin and bevacizumab followed by maintenance pemetrexed and bevacizumab versus paclitaxel plus carboplatin and bevacizumab followed by maintenance bevacizumab in patients with stage IIIB or IV nonsquamous non-small-cell lung cancer. *J. Clin. Oncol.* **31**, 4349–4357 (2013).
58. Paz-Ares, L. G. et al. Phase III, randomized, double-blind, placebo-controlled trial of gemcitabine/cisplatin alone or with sorafenib for the first-line treatment of advanced, nonsquamous non-small-cell lung cancer. *J. Clin. Oncol.* **30**, 3084–3092 (2012).
59. Paz-Ares, L. et al. A randomized phase 2 study of paclitaxel and carboplatin with or without conatumumab for first-line treatment of advanced non-small-cell lung cancer. *J. Thorac. Oncol.* **8**, 329–337 (2013).
60. Pirker, R. et al. Cetuximab plus chemotherapy in patients with advanced non-small-cell lung cancer (FLEX): an open-label randomised phase III trial. *Lancet* **373**, 1525–1531 (2009).
61. Quoix, E. et al. Carboplatin and weekly paclitaxel doublet chemotherapy compared with monotherapy in elderly patients with advanced non-small-cell lung cancer: IFCT-0501 randomised, phase 3 trial. *Lancet* **378**, 1079–1088 (2011).
62. Ramalingam, S. S. et al. Dacomitinib versus erlotinib in patients with advanced-stage, previously treated non-small-cell lung cancer (ARCHER 1009): a randomised, double-blind, phase 3 trial. *Lancet Oncol.* **15**, 1369–1378 (2014).
63. Ramlau, R. et al. Aflibercept and Docetaxel versus Docetaxel alone after platinum failure in patients with advanced or metastatic non-small-cell lung cancer: a randomized, controlled phase III trial. *J. Clin. Oncol.* **30**, 3640–3647 (2012).
64. Reck, M. et al. Phase III trial of cisplatin plus gemcitabine with either placebo or bevacizumab as first-line therapy for nonsquamous non-small-cell lung cancer: AVAil. *J. Clin. Oncol.* **27**, 1227–1234 (2009).
65. Reck, M. et al. A randomized, double-blind, placebo-controlled phase 2 study of tigatuzumab (CS-1008) in combination with carboplatin/paclitaxel in patients with chemotherapy-naive metastatic/unresectable non-small cell lung cancer. *Lung Cancer* **82**, 441–448 (2013).
66. Reck, M. et al. Docetaxel plus nintedanib versus docetaxel plus placebo in patients with previously treated non-small-cell lung cancer (LUME-Lung 1): a phase 3, double-blind, randomised controlled trial. *Lancet Oncol.* **15**, 143–155 (2014).
67. Rosell, R. et al. Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol.* **13**, 239–246 (2012).
68. Rudd, R. M. et al. Gemcitabine plus carboplatin versus mitomycin, ifosfamide, and cisplatin in patients with stage IIIB or IV non-small-cell lung cancer: a phase III

randomized study of the London Lung Cancer Group. *J. Clin. Oncol.* **23**, 142–153 (2005).
69. Scagliotti, G. V. et al. International, randomized, placebo-controlled, double-blind phase III study of motesanib plus carboplatin/paclitaxel in patients with advanced nonsquamous non-small-cell lung cancer: MONET1. *J. Clin. Oncol.* **30**, 2829–2836 (2012).
70. Scagliotti, G. V. et al. Sunitinib plus erlotinib versus placebo plus erlotinib in patients with previously treated advanced non-small-cell lung cancer: a phase III trial. *J. Clin. Oncol.* **30**, 2070–2078 (2012).
71. Seto, T. et al. Erlotinib alone or with bevacizumab as first-line therapy in patients with advanced non-squamous non-small-cell lung cancer harbouring EGFR mutations (JO25567): an open-label, randomised, multicentre, phase 2 study. *Lancet Oncol.* **15**, 1236–1244 (2014).
72. Shaw, A. T. et al. Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *N. Engl. J. Med.* **368**, 2385–2394 (2013).
73. Shi, Y. et al. Icotinib versus gefitinib in previously treated advanced non-small-cell lung cancer (ICOGEN): a randomised, double-blind phase 3 non-inferiority trial. *Lancet Oncol.* **14**, 953–961 (2013).
74. Smit, E. F. et al. Randomized phase II and pharmacogenetic study of pemetrexed compared with pemetrexed plus carboplatin in pretreated patients with advanced non-small-cell lung cancer. *J. Clin. Oncol.* **27**, 2038–2045 (2009).
75. Solomon, B. J. et al. First-line crizotinib versus chemotherapy in ALK-positive lung cancer. *N. Engl. J. Med.* **371**, 2167–2177 (2014).
76. Spigel, D. R. et al. Randomized phase II trial of Onartuzumab in combination with erlotinib in patients with advanced non-small-cell lung cancer. *J. Clin. Oncol.* **31**, 4105–4114 (2013).
77. Treat, J. A. et al. A randomized, phase III multicenter trial of gemcitabine in combination with carboplatin or paclitaxel versus paclitaxel plus carboplatin in patients with advanced or metastatic non-small-cell lung cancer. *Ann. Oncol.* **21**, 540–547 (2010).
78. Wu, Y. L. et al. Afatinib versus cisplatin plus gemcitabine for first-line treatment of Asian patients with advanced non-small-cell lung cancer harbouring EGFR mutations (LUX-Lung 6): an open-label, randomised phase 3 trial. *Lancet Oncol.* **15**, 213–222 (2014).
79. Wu, Y. L. et al. Intercalated combination of chemotherapy and erlotinib for patients with advanced stage non-small-cell lung cancer (FASTACT-2): a randomised, double-blind trial. *Lancet Oncol.* **14**, 777–786 (2013).
80. Zukin, M. et al. Randomized phase III trial of single-agent pemetrexed versus carboplatin and pemetrexed in patients with advanced non-small-cell lung cancer and Eastern Cooperative Oncology Group performance status of 2. *J. Clin. Oncol.* **31**, 2849–2853 (2013).
81. Paz-Ares, L. G. et al. PARAMOUNT: final overall survival results of the phase III study of maintenance pemetrexed versus placebo immediately after induction treatment with pemetrexed plus cisplatin for advanced nonsquamous non-small-cell lung cancer. *J. Clin. Oncol.* **31**, 2895–2902 (2013).
82. Aerts, J. G. et al. A randomized phase II study comparing erlotinib versus erlotinib with alternating chemotherapy in relapsed non-small-cell lung cancer patients: the NVALT-10 study. *Ann. Oncol.* **24**, 2860–2865 (2013).
83. Georgoulias, V. et al. Docetaxel versus docetaxel plus cisplatin as front-line treatment of patients with advanced non-small-cell lung cancer: a randomized, multicenter phase III trial. *J. Clin. Oncol.* **22**, 2602–2609 (2004).
84. Trinquart, L., Jacot, J., Conner, S. C. & 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.* **34**, 1813–1819 (2016).
85. Brahmer, J. et al. Nivolumab versus Docetaxel in advanced squamous-cell non-small-cell lung cancer. *N. Engl. J. Med.* **373**, 123–135 (2015).
86. Borghaei, H. et al. Nivolumab versus Docetaxel in advanced nonsquamous non-small-cell lung cancer. *N. Engl. J. Med.* **373**, 1627–1639 (2015).
87. Chen, T. T. Statistical issues and challenges in immuno-oncology. *J. Immunother. Cancer* **1**, 18 (2013).
88. Guyot, P., Ades, A. E., Ouwens, M. J. N. M. & Welton, N. J. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan–Meier survival curves. *BMC Med. Res. Methodol.* **12**, 9 (2012).