ESMO Open Cancer Horizons

Detailed statistical assessment of the characteristics of the ESMO Magnitude of Clinical Benefit Scale (ESMO-MCBS) threshold rules

Urania Dafni,[1,2] Dimitris Karlis,[3] Xanthi Pedeli,[2] Jan Bogaerts,[4] George Pentheroudakis,[5] Josep Tabernero,[6] Christoph C Zielinski,[7] Martine J Piccart,[8] Elisabeth G E de Vries,[9] Nicola Jane Latino,[10] Jean-Yves Douillard,[10] Nathan I Cherny[11]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/esmoopen-2017-000216).

To cite: Dafni U, Karlis D, Pedeli X, *et al*. Detailed statistical assessment of the characteristics of the ESMO Magnitude of Clinical Benefit Scale (ESMO-MCBS) threshold rules. *ESMO Open* 2017;**2**:e000216. doi:10.1136/esmoopen-2017-000216

Received 9 May 2017
Revised 6 July 2017
Accepted 20 July 2017

For numbered affiliations see end of article.

Correspondence to
Prof Urania Dafni; udafni@nurs.uoa.gr

ABSTRACT

Background The European Society for Medical Oncology (ESMO) has developed the ESMO Magnitude of Clinical Benefit Scale (ESMO-MCBS), a tool to assess the magnitude of clinical benefit from new cancer therapies. Grading is guided by a dual rule comparing the relative benefit (RB) and the absolute benefit (AB) achieved by the therapy to prespecified threshold values. The ESMO-MCBS v1.0 dual rule evaluates the RB of an experimental treatment based on the lower limit of the 95%CI (LL95%CI) for the hazard ratio (HR) along with an AB threshold. This dual rule addresses two goals: inclusiveness: not unfairly penalising experimental treatments from trials designed with adequate power targeting clinically meaningful relative benefit; and discernment: penalising trials designed to detect a small inconsequential benefit.

Methods Based on 50 000 simulations of plausible trial scenarios, the sensitivity and specificity of the LL95%CI rule and the ESMO-MCBS dual rule, the robustness of their characteristics for reasonable power and range of targeted and true HRs, are examined. The per cent acceptance of maximal preliminary grade is compared with other dual rules based on point estimate (PE) thresholds for RB.

Results For particularly small or particularly large studies, the observed benefit needs to be relatively big for the ESMO-MCBS dual rule to be satisfied and the maximal grade awarded. Compared with approaches that evaluate RB using the PE thresholds, simulations demonstrate that the MCBS approach better exhibits the desired behaviour achieving the goals of both inclusiveness and discernment.

Conclusions RB assessment using the LL95%CI for HR rather than a PE threshold has two advantages: it diminishes the probability of excluding big benefit positive studies from achieving due credit and, when combined with the AB assessment, it increases the probability of downgrading a trial with a statistically significant but clinically insignificant observed benefit.

Key message

Extensive simulations of plausible trial scenarios demonstrate that the ESMO-Magnitude of Clinical Benefit Scale (ESMO-MCBS) dual rule, incorporating the lower limit of the 95% CI (LL95%CI) threshold for HR, addresses both goals of inclusiveness and discernment effectively, and more so than a dual rule using point estimate (PE) thresholds. The ESMO-MCBS avoids excluding a substantial proportion of big benefit positive studies from achieving due credit, as would be the case if the PE were to be used, and results in downgrading more trials with clinically insignificant observed benefit.

INTRODUCTION

The European Society for Medical Oncology (ESMO) has developed the ESMO-Magnitude of Clinical Benefit Scale (ESMO-MCBS) to evaluate high-quality clinical trial results, recognising the need for a standardised approach for grading the magnitude of clinical benefit derived from new therapeutic approaches.[1] The ESMO-MCBS is a reproducible, semiquantitative tool for grading clinical benefit, thereby intending to prioritise therapies that should be rapidly accessible to all European citizens. The ESMO-MCBS version 1.0 is especially designed to evaluate comparative outcome studies in solid cancers.

In ESMO-MCBS v1.0, separate forms were developed for the curative and non-curative setting—forms 1 and 2. In the non-curative setting, a three-step process is implemented (figure 1). The strength of the randomised evidence is established at the first step, and in the second step, based on the quantitative component of the scale, a preliminary grade is assigned. It incorporates a dual rule that evaluates both the observed relative benefit (RB), that is, the observed hazard ratio (HR), and the observed absolute benefit (AB) on a time-to-event outcome (progression-free survival (PFS) and overall survival (OS)) achieved by the treatment. The dual rule consists of the following two components:

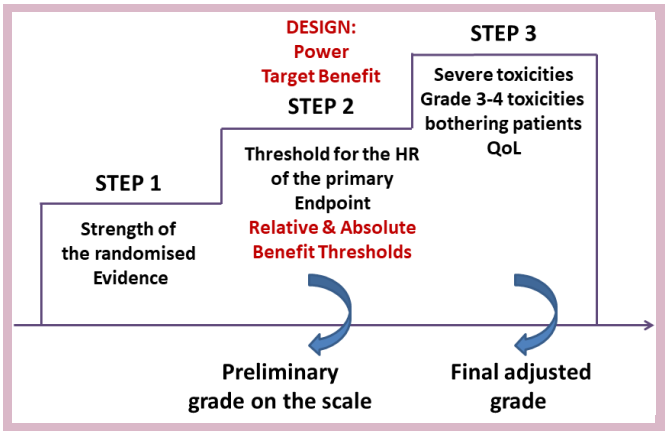


Figure 1 The three critical evaluation steps of ESMO-MCBS.
ESMO, European Society for Medical Oncology; MCBS, Magnitude of Clinical Benefit Scale; QoL, quality of life.

1. *RB rule* : the lower limit (LL) of the 95% CI for the HR is compared with specified threshold values (PFS: LL ≤0.65; OS: LL ≤0.65 or LL ≤0.70 for median control ≤12 months or >12 months, respectively)
2. *AB rule* : the observed absolute difference in median treatment outcomes is compared with the *minimum clinically significant AB*.

Furthermore, when the primary outcome is OS, the preliminary score can also be obtained based on an alternative rule evaluating the magnitude of observed long-term beneficial effect to prespecified threshold values.

In the third step, the grade can be adjusted to reflect the toxicity and quality of life (QoL) outcomes of the investigative treatment (qualitative component).

The ESMO-MCBS approach to RB evaluation highlights the importance of the CI for the HR, which is estimated from the observed data and presents a range of values in which the true value for HR may lie.[2] The RB criterion uses the observed HR along with its precision by setting a threshold at the LL of the corresponding 95% CI. A magnitude of RB included in the interval of HR is a plausible true value for the treatment benefit and should be recognised as such. Congruence with meaningful gains drives the choice of the LL of the CI of HR (LL95%CI rule) as a critical statistic, rather than the point estimate (PE rule). The LL95%CI rule is by design more lenient for higher true benefit, than the corresponding PE rule, minimising the false negative results, that is, maximising the sensitivity of the first component of the dual rule.

As a counterbalance, the role of the AB rule is to guarantee that the relevant *minimum clinically significant AB* is observed. This required AB gets progressively smaller, the smaller the median control time-to-event is (table 1). Thus, the dual rule, combining the RB and AB rules, addresses two requirements: not penalising treatments that have effects that are plausibly congruent with the desired magnitude of RB, while penalising treatments that provide only a trivial observed AB.

This approach has been questioned, and concern has been expressed that it is excessively lenient especially in the setting of lower powered studies while at the same time discouraging high powered studies with narrow 95% CIs.[3–9]

To address these concerns, we have carried out extensive simulations to explore the behaviour of the ESMO-MCBS dual rule using the LL95%CI and compare it with the approach using PE within the context of comparative outcome studies. Here we report, for the non-curative setting, the results of simulations of plausible trial scenarios, each replicated 50 000 times.

Table 1 The ESMO-MCBS dual rule: implemented combined thresholds for the HR and the minimum absolute benefit (AB) gain for OS and PFS that could be considered as deserving the maximal preliminary grade.

Non-curative setting				
Maximal preliminary grade	**Primary endpoint (time-to-event)**	**Criteria**		
	OS - median (control)	HR*		AB gain
4	≤12 months	≤0.65	AND	≥3 months
	>12 months	≤0.70	AND	≥5 months
	PFS- median (control)	HR*		AB gain
3	≤6 months	≤0.65	AND	≥1.5 months
	>6 months	≤0.65	AND	≥3 months

*Thresholds refer to the lower limit of the 95% CI.
ESMO, European Society for Medical Oncology; MCBS, Magnitude of Clinical Benefit Scale; OS, overall survival; PFS, progression-free survival.

AIMS

The aims of the current study are:

a. To provide the rationale behind the use of the ESMO-MCBS dual rule, with emphasis on the use of the LL95%CI rule for the assignment of the maximal RB and graphically illustrate its behaviour and corresponding field testing application
b. To evaluate the sensitivity and specificity of the ESMO-MCBS LL95%CI rule. To examine the robustness of its characteristics for reasonable power (80%–90%) and range of targeted and true HRs (range of HRs from 0.60 to 0.90), and how they compare with RB rules based on a range of PE threshold values
c. To evaluate whether the ESMO-MCBS LL95%CI rule behaves properly under reasonable design power, by not penalising trials that use high power for substantial benefit, when true, while penalising trials that use high power for small benefit when only small benefit exists
d. To compare sensitivity and specificity of the ESMO-MCBS v1.0 dual rule to other dual scores based on PE rules including those proposed by Sobrero *et al*[10]
e. To demonstrate the application of the ESMO MCBS v1.0 dual rule in field testing.

METHODS

The behaviour of the ESMO-MCBS v1.0 over the range of primary endpoints used in the ESMO-MCBS evaluation (either PFS or OS) for the appropriate median control range (≤12 or >12 months) and the corresponding design power (80%–90%) is illustrated across different design HRs. Behaviours were evaluated conceptually and with corresponding data from the field testing of ESMO-MCBS v1.0.

The ESMO-MCBS Working Group, considered a true relative decrease in risk of at least 35% or more (or 30% or more, for OS with median control >12 months), as deserving the maximal preliminary grade.[1] A decrease of risk by at least 20% is necessary to satisfy the required minimum observed AB chosen by the ESMO-MCBS Working Group for achieving the maximal preliminary grade when median control for OS or PFS is ≤12 months.[1] An HR of >0.80 is generally considered to correspond to a relatively small RB. This convention was endorsed by the Working Groups set by the ASCO Cancer Research Committee to propose new thresholds for clinically meaningful outcomes of cancer medications for metastatic colon cancer, metastatic breast cancer, non-small cell lung cancer and pancreatic cancer. For the cancer types examined, the maximum recommended design HR for a meaningful clinical trial goal was set at HR=0.80, for power 80%–90%.[11]

The benefit of an experimental treatment is evaluated based on the available information, which includes the targeted (or design) HR with a specific power according to the study design and the observed HR (with corresponding CI). The true HR is not known in real life and only through simulations assuming true HR in study scenarios one can elicit the behaviour of the different evaluation rules.

In the simulations, time-to-event data were produced from exponential distributions with parameters selected to satisfy the assumptions of the true values for the HR and median for the control group. Trial sizes and number of events to stop the trial were determined based on the design HR and the assumed power. In most of the examined scenarios, it was assumed that the trials were correctly designed by targeting the true HR (true HR=design HR). Random censoring was considered throughout the simulations. For each combination, 50 000 simulated trials were run. For each replication, several characteristics including the observed HR with the corresponding 95% Wald CI and the observed medians for the control and experimental treatment were produced. All calculations were performed using the R language for statistical computing V.3.2.2.[12]

Three approaches were used to evaluate the performance of the ESMO-MCBS LL95%CI threshold for RB. Comparison of the ESMO-MCBS LL95%CI rule with the PE rule for a range of threshold values was performed using receiver operating characteristic (ROC) curves, plotting true positive rate against false positive rate across a range of true RB levels and study power. Second, through simulations, the %acceptance of maximal RB score was used to compare the MCBS LL95%CI threshold to the PE thresholds exhibiting similar behaviour when the design is correctly targeting the true HR, and third, when the design and true HRs differ, over a range of true HRs, design HRs and study powers.

Further simulations were run to evaluate the performance of the ESMO-MCBS v1.0 dual rules on the per cent acceptance of maximal preliminary grade (%acceptance) over a range of plausible scenarios. The behaviour of dual rules based on the PEs exhibiting a behaviour more closely corresponding to the LL95%CI thresholds, as well as the ones described by Sobrero *et al*, was compared with the behaviour of the ESMO-MCBS v1.0 dual rules.[10]

RESULTS

Graphical illustration of the ESMO-MCBS v1.0 preliminary grading

Assuming trials are correctly designed, the behaviour of the ESMO-MCBS LL95%CI rule over the range of design HR from 0.60 to 0.90 (targeted HR used in the alternative hypothesis) and design power of 80% or 90% is presented graphically (figure 2A,B). Each figure reflects the MCBS LL95%CI rule for the corresponding median control range of the primary endpoint (figure 2A: LL95%CI ≤0.65, for PFS, and OS with median control ≤12 months; figure 2B: LL95%CI ≤ 0.70 for OS with median control >12 months; design HR=true HR).

There are several critical observations from this illustration:

1. The maximum observed HR leading to a statistically significant result (stars in figures) increases as the design HR value increases (x-axis). This is the direct result of targeting a progressively smaller benefit, resulting to a narrower 95% CI centred closer to 1.
2. *Trials targeting big benefit* (non-shaded area; figure 2A,B): for significant trials, the LL95%CI rule is always satisfied for true HR up to some value (identical circles and stars) (PFS and OS with median control ≤12 months, LL95%CI ≤0.65: up to HR ≤0.736 for 80% power, and HR ≤0.701 for 90% power; OS with median control >12 months, LL95%CI ≤0.70: up to HR ≤0.77 for 80% power, and HR ≤0.74 for 90% power).

   Thus, a penalty to the ESMO-MCBS v1.0 score for such a study will only depend on the magnitude of the observed AB, while the RB rule is inactive. It is important to emphasise that in this big benefit range for true HR, the LL95%CI rule awards maximal RB grade to 100% of the significant trials. This translates to a bigger number of trials assigned maximal grade when designed with 90% power than with 80% power (blue circles above red circles).
3. *Trials targeting smaller benefit* (shaded area: red for 80% power, blue for 90% power; figure 2A,B): the MCBS RB rule assigns the maximal score only for a subset of the significant trials. A statistically significant result in this case can occur even if the observed HR is

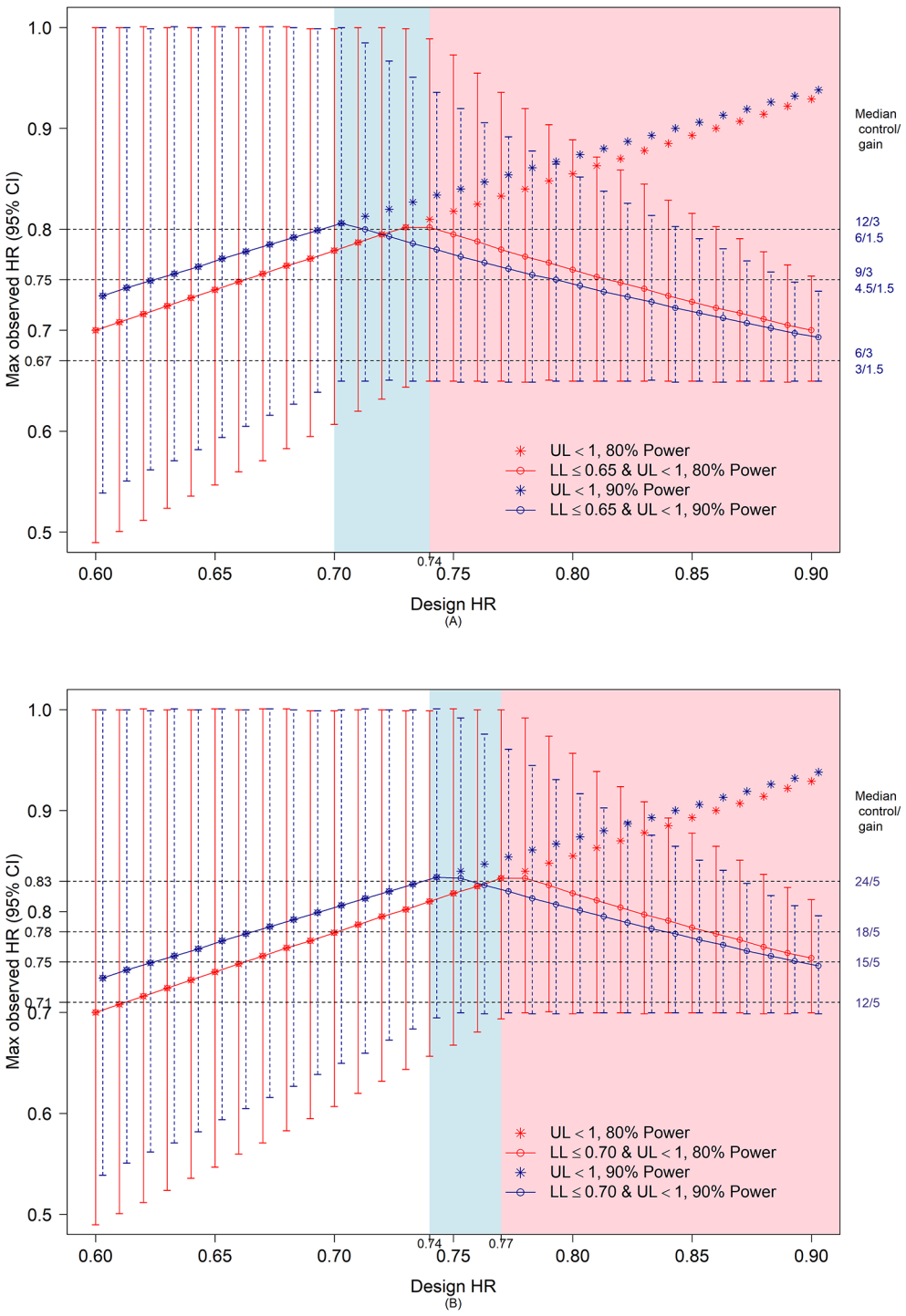


Figure 2 Maximum observed HR to achieve statistical significance (two-sided alpha=0.05) and maximum observed HR that satisfies the LL95%CI rule for maximal RB, for a range of design HRs . Note: design HR=true HR. (A) OS, median control ≤12 months; PFS, all medians: LL95%CI ≤0.65. (B) OS: median control >12 months: LL95%CI ≤0.70.

Annotation: *Symbols* and *shaded area*: red for power 80%, blue for power 90%. *Stars*: maximum observed HR to achieve statistical significance (two-sided significance level 0.05). *Circles*: maximum observed HR that would satisfy the LL95%CI rule for maximal RB score for significant trials. *Progressively smaller vertical linear segments*: corresponding estimated 95% CIs (for observed HR). *Dotted horizontal lines with median control/gain values listed on the right y-axis*: observed HR that would correspond to required AB gains and control medians, that is, HR=0.80 broadly corresponds to a 3-month gain for a 12-month median control. *Non-shaded area in the figures*: range for design HR for which stars and circles are identical, representing trials targeting relatively big benefit. *Shaded area in the figures*: range for design HR starting above a certain value, beyond which circles are progressively lower than stars, representing trials targeting smaller benefit.

LL95%CI, lower limit of the 95% CI; RB, relative benefit; OS, overall survival; PFS, progression-free survival; AB, absolute benefit; UL, upper limit.

very close to 1, thus in these large-sized trials, a small observed RB can still be statistically significant. When large studies target progressively smaller benefits, an increasing proportion of trials are penalised by the RB rule and consequently the maximum observed HR value allowed by the RB rule tends to move further away from the maximum observed HR achieving significance. This phenomenon is slightly amplified with higher power.

Simulation results 1: comparison of the LL95%CI and PE thresholds regarding sensitivity and specificity for maximal RB grading

Inherent to the choice of any threshold is a contrast between false negative and false positive results. The best rule would maximise both sensitivity and specificity. Assuming that a trial is correctly designed to detect with power 80%–90% a true benefit, the ideal rule should minimise false-negative results, by not misclassifying an experimental treatment as not providing big RB (true HR ≤0.65, ≤0.70) if in fact it does. This relates to sensitivity. In parallel, it should minimise false-positive results, by not misclassifying an experimental treatment as providing big RB, if in fact it does not (true HR >0.80). This relates to specificity.

By these criteria, superiority of the LL95%CI rule over a PE rule for the evaluation of RB is demonstrated in three tests:

1. ROC curves (figure 3): the optimal result on a ROC curve corresponds to the top left corner, that is, 100% true positive, 0% false positive. The relative characteristics of the MCBS rule for maximal RB classification, using the LL95%CI ≤0.65 threshold, are

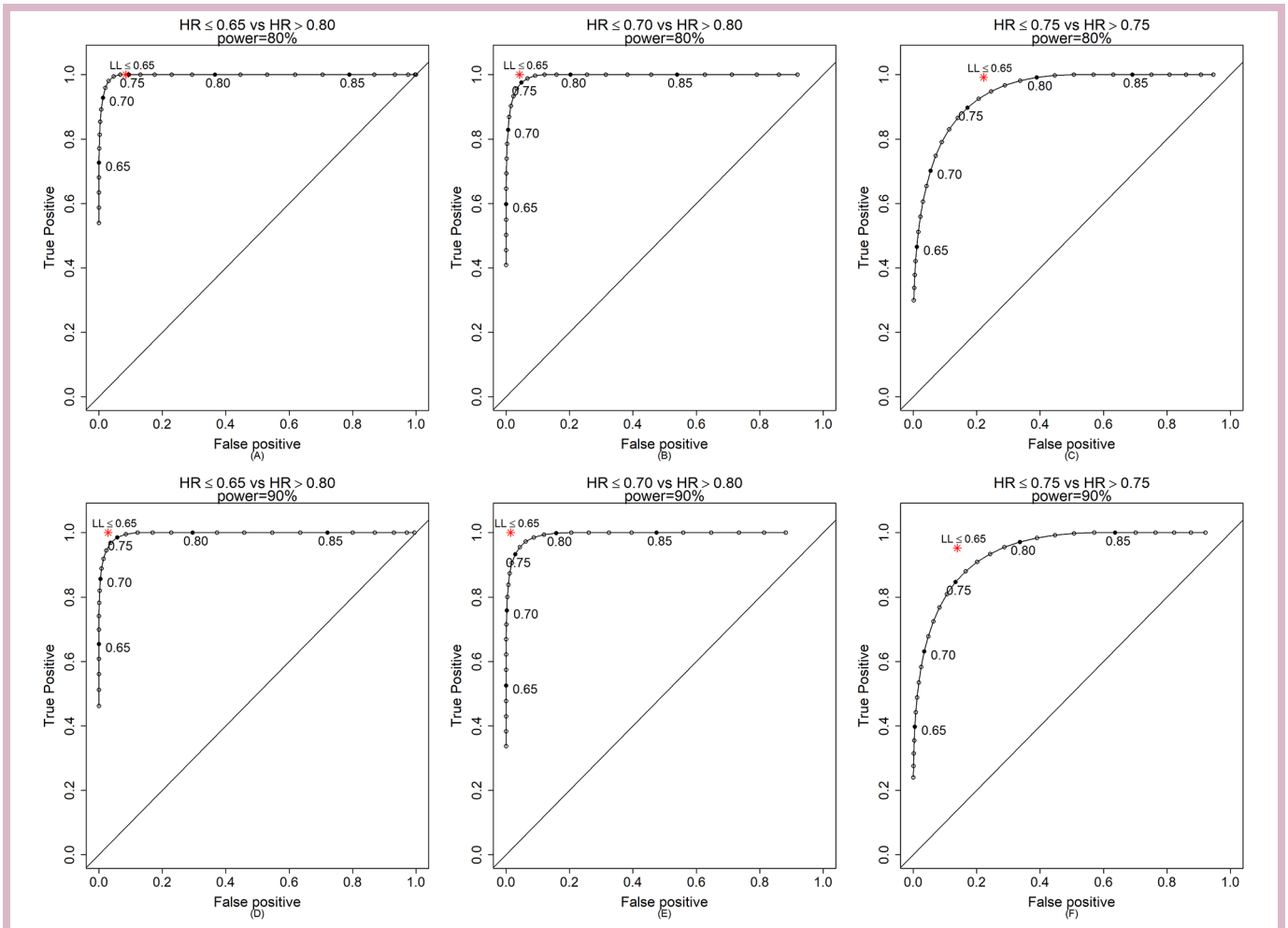


Figure 3 Comparison of the LL95%CI rule to a PE threshold for maximal RB classification: sensitivity and specificity for the MCBS LL95%CI ≤0.65 rule versus the ROC curve for PE rules ranging from 0.60 to 0.90 (time-to-event median control=6 months). (A–C) Power 80%. (D–F) Power 90%.

Annotation: *Left and centre panels:* true HR ≤0.65 and HR≤0.70, are respectively classified as big benefit, contrasted to a small benefit classification of true HR>0.80. *Right panels:* true HR ≤0.75 is classified as a relatively big benefit, contrasted to a small benefit classification of true HR>0.75. *Red star*: the MCBS LL95%CI≤0.65 rule for HR. *ROC curve* (*black*): PE rules ranging from 0.60 to 0.90.

LL95%CI, lower limit of 95% CI; PE, point estimate; RB, relative benefit; MCBS, Magnitude of Clinical Benefit Scale; ROC, receiver operating characteristic.

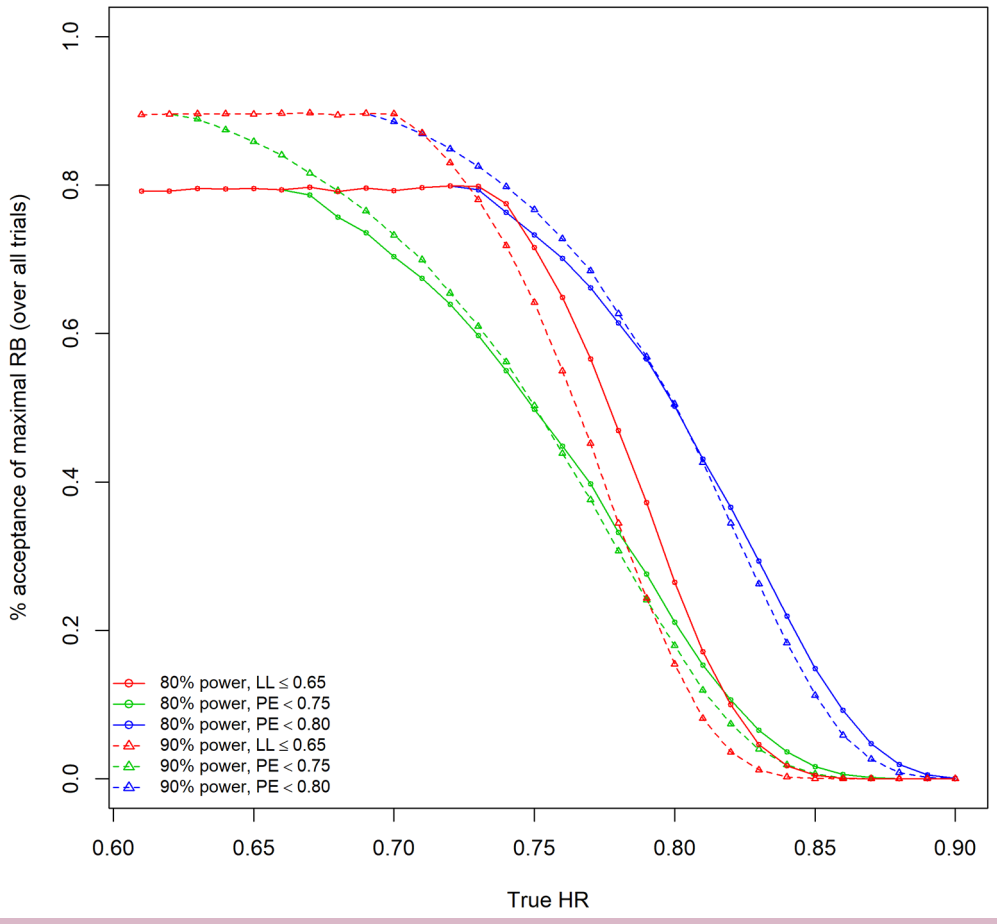


Figure 4 Comparison of the LL95%CI rule with PE rules with similar behaviour: %acceptance of maximal RB for power 80% and 90% over all trials, for LL95%CI ≤0.65 rule, PE <0.75 and PE <0.80.

Note: Initially, for big benefit studies, the %acceptance rate of maximal RB is higher when designed with 90% power versus with 80% power. Crossing of 80% and 90% power lines for: LL95%CI ≤0.65 (red lines) occurs at HR=0.73 with %acceptance 80%; PE <0.75 (green lines) occurs at HR=0.75 with %acceptance 50%; PE <0.80 (blue lines) occurs at HR=0.80 with %acceptance 50%.

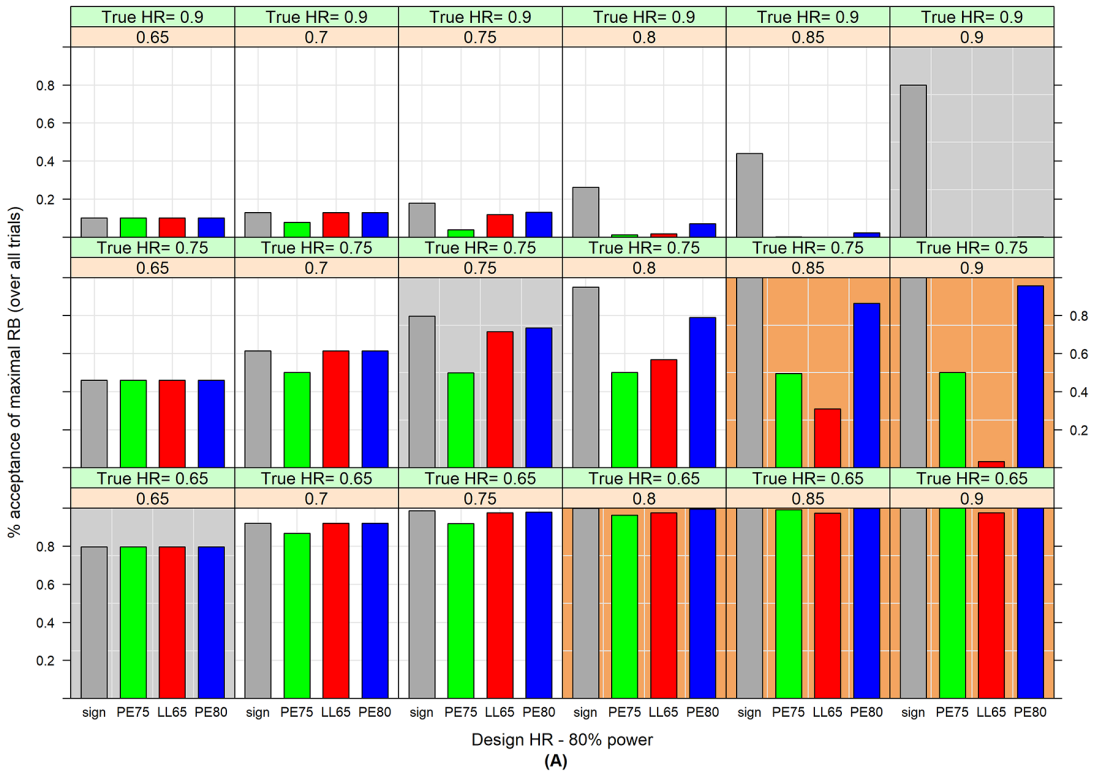
LL95%CI, lower limit of the 95% CI; PE, point estimate; RB, relative benefit.

compared with a range of PE thresholds and plotted in six ROC curves for true HR ranging from 0.60 to 0.90, for power 80% and 90% (figure 3). The LL95%CI rule (depicted as a red star) compared with a range of PE thresholds from 0.60 to 0.90, is almost always above and/or to the left of the PE ROC curve, indicating optimal balance between maximal true-positive and minimal false-positive results.

2. Comparison of the MCBS LL95%CI rule to PE rules with similar behaviour: translating the LL95%CI rule to similarly behaving PE thresholds, we find that LL95%CI ≤0.65 would more closely correspond to the behaviour of a PE threshold between 0.75 and 0.80, while the LL95%CI ≤0.70 to a PE threshold between 0.80 and 0.83. Based on %acceptance of maximal RB, it is again evident that the LL95%CI rule, for both powers 80% and 90%, exhibits consistently higher sensitivity for true big benefit and/or higher specificity for true small benefit (figure 4, online supplementary figure S1A–C).

In addition, comparing the effect of power on the LL95%CI rule, the %acceptance of maximal RB is evidently higher among all trials designed with 90% power than in studies designed with 80% power, up to a relatively big RB (true HR <0.73 for LL95%CI ≤0.65; true HR <0.77 for LL95%CI ≤0.70; figure 4, online supplementary figure S1B). The same is true for any PE rule, but in contrast to the LL95%CI threshold, it occurs only up to that specific PE value, at which a much lower %acceptance of only 50% is achieved (eg, if PE <0.75, then for true HR=0.75, the %acceptance among all trials cannot exceed 50%). This disadvantage is even more pronounced when using the lower PE threshold values of HR=0.60 or HR=0.65 (see online supplementary figure S2A,B).[10]

3. Sensitivity and specificity for true HR over a range of design HRs: the %acceptance of maximal RB grade over all trials for the LL95%CI ≤0.65 rule is presented for true HR values of 0.65, 0.75 and 0.90 in studies using 80% and 90% power, respectively (figure 5A,B).



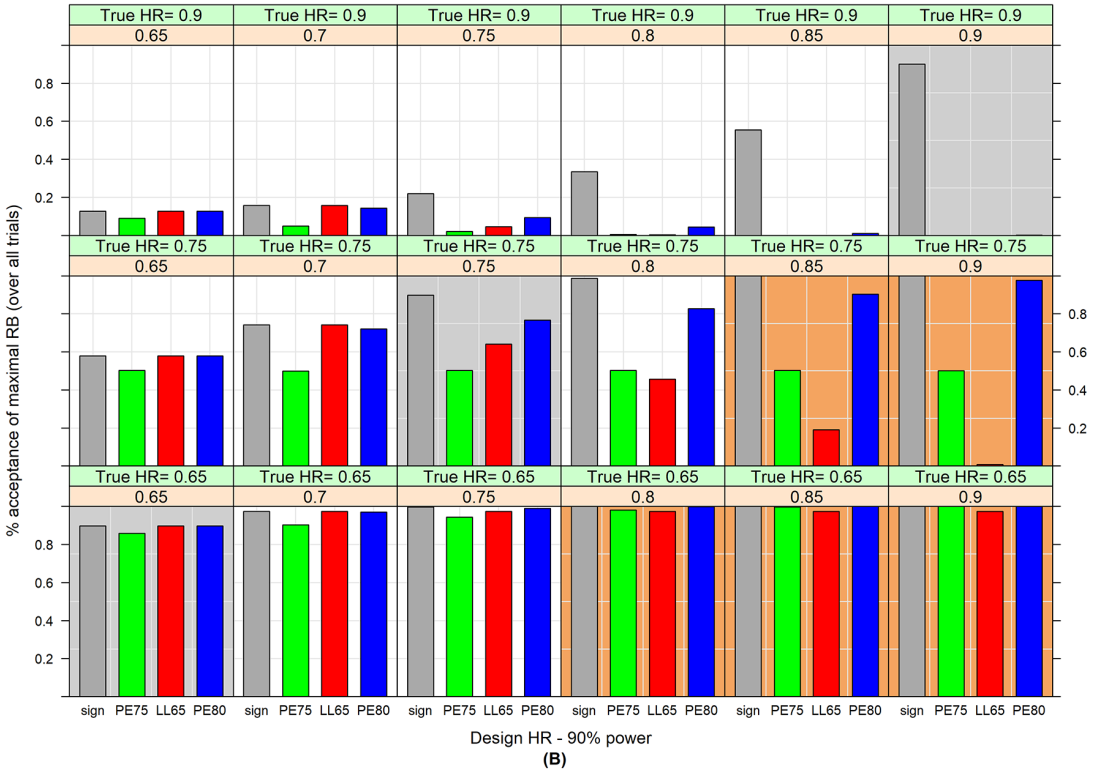


Figure 5 Comparison of the LL95%CI rule to PE rules with similar behaviour, when the true HR is different from the design HR: %acceptance of maximal relative benefit, over all trials, for true HR of 0.90, 0.75 and 0.65, over a range of design HRs.
Annotation: *Panel headers:* true HR (0.65, 0.75 and 0.90) is indicated above design HR (0.65 to 0.90). *Height of the bars:* proportion out of the 50 000 simulated trials satisfying each condition. *Grey bars:* proportion of simulated trials found statistically significant (at two-sided significance level 0.05). *Red, green and blue bars*: %acceptance using LL95%CI ≤0.65, PE of 0.75 and 0.80, respectively. *Grey-shaded areas:* scenarios with design HR=true HR. *Orange-shaded areas:* implausible scenarios (eg, when true HR=0.65 and design HR≥0.85).
LL95%CI, lower limit of 95% CI; PE, point estimate; RB, relative benefit.

The LL95%CI rule presents the desired robust behaviour across different targeted HRs for specified power.

For true HR=0.90 (top row in figure 5A,B), the LL95%CI rule (red bar) has a %acceptance close to zero for design HR ≥0.80, similar to PE <0.75 (green bar), discouraging the design of large size trials targeting small benefit (highest specificity: LL95%CI rule and PE <0.75). However, it can be seen that small-sized trials targeting relatively big RB (design HR ≤0.75) provide a minor but still existing risk to reach a false positive conclusion (LL95%CI rule similar to PE <0.80 (blue bar)).

On the other end of the spectrum, for true HR=0.65 (bottom row in figure 5A,B), the LL95%CI ≤0.65 rule has %acceptance equal to the corresponding power (80% or 90%) for a design HR=0.65, and it increases to almost 100% as targeted benefit magnitude decreases (highest sensitivity: LL95%CI rule and PE <0.80).

This advantage of the LL95%CI ≤0.65 rule is also apparent for true HR=0.75 (middle row in figure 5A,B), where the LL95%CI ≤0.65 rule provides high %acceptance when big benefit is targeted, diminishing %acceptance as targeted benefit gets smaller, while it achieves the smallest %acceptance for low targeted RB. The PE <0.75 provides %acceptance around 50% throughout, while the PE <0.80 gives higher and higher %acceptance the smaller the targeted benefit and the higher the power, thus indicating that the PE <0.75 is insensitive to the design HR and power, while the PE <0.80 promotes grossly overpowered trials.

In fairness, scenarios of evaluating results from grossly overpowered trials (eg, when true HR=0.65 and design HR ≥0.80) are not particularly plausible. First, it is not reasonable to run an unnecessarily large trial when expecting a true big benefit. Second, if such a trial was designed based on the initial assumption that the true effect is small, the study would probably not continue to completion and would stop at the interim analysis, which is ethically required in such a large trial.

Simulation results 2: comparison of the dual rule based on the LL95%CI ≤0.65 or ≤0.70 and on PE thresholds PE <0.75, PE <0.80 or PE <0.83 (thresholds exhibiting similar behaviour).

Assuming that a trial is correctly designed to detect with power 80%–90% a true benefit, the %acceptance over a range of plausible scenarios for significant trials is compared between the dual rule based on the PE thresholds behaving similarly to the proposed LL95%CI threshold incorporated in ESMO-MCBS v1.0 (figure 6A,B: PFS; figure 6C,D: OS; 80% and 90% power).

Big true benefit (illustrated for true HR=0.60, 0.65, 0.70)

In all of these scenarios, the LL95%CI ≤0.65 rule is consistently at least as lenient as the PE threshold of 0.80 (and more lenient than PE <0.75), accepting for highest grade almost 100% of the significant trials. The same is true for the LL95%CI ≤0.70 rule and the corresponding PE <0.80, PE <0.83, respectively, applied when median control OS >12 months.

The limiting effect of the strict AB rule, whereby maximal preliminary grade is only achieved when AB targets are met, is especially pronounced when the OS median control is 6 months or less. However, when the median OS is long (eg, 24 months), this constraint is lost since the AB rule (OS: 5 months gain) is always satisfied.

Smaller true benefit (illustrated for true HR=0.75, 0.80, 0.85)

The PE <0.75 threshold imposes initially a higher penalty, which becomes almost identical to the LL95%CI rule for true HR >0.80. For true HR=0.75, it overpenalises studies compared with the other criteria. When the true HR is 0.80 or 0.85, indicating a lower level of benefit, the LL95%CI ≤0.65 threshold behaves similarly to the PE >0.75 threshold and both are substantially stricter than the PE >0.80 threshold.

Thus, over the full range of true HRs, the ESMO-MCBS v1.0 dual rule is exhibiting the most desired discriminatory behaviour by being at least as lenient or more than the dual rule based on PE <0.80 when the true benefit is big, while it is similarly strict to the dual rule based on PE <0.75 when true benefit is small (true HR = 0.80, 0.85).

The PE thresholds for the assessment of RB for OS that were proposed by Sobrero *et al* are substantially more non-inclusive.[10] For example, for median control OS of 6 months, this rule leads to a false-negative rate of 61% for true HR=0.65, in a correctly designed study with power 80% and to 69%, for power 90% (see online supplementary figure S3A,B). In the field testing using these PE thresholds, two-thirds of the six studies (67%) achieving the maximal ESMO-MCBS preliminary grade with primary endpoint OS, and judged by the experts as deserving it, would have been unduly penalised (see online supplementary tables S3 and S4; studies highlighted with grey background).

Graphical illustration of the ESMO-MCBS v1.0 field testing results

The ESMO-MCBS v1.0 preliminary grades of trials evaluated in the field testing are shown along with the relevant RB cut-offs over the encountered full range of design HRs (figure 7; online supplementary figure S4A–E).[1] Each figure reflects the primary endpoint used in the ESMO-MCBS evaluation (either PFS or OS) for the appropriate median control range (≤12 or >12 months) and the corresponding design power (80%, or 85%, ≥90%). All symbols, lines and shaded areas are as defined before (in figure 2). Squares represent actual study information on observed HR with corresponding 95% CI (y-axis: observed HR vs x-axis: design HR), while their colour indicates whether the studies meet the criteria for maximal preliminary score (figure 7). The colour of the print in the tabulated data from the field testing that is presented in online supplementary tables S1–4 corresponds to the colour of the squares in figure 7.

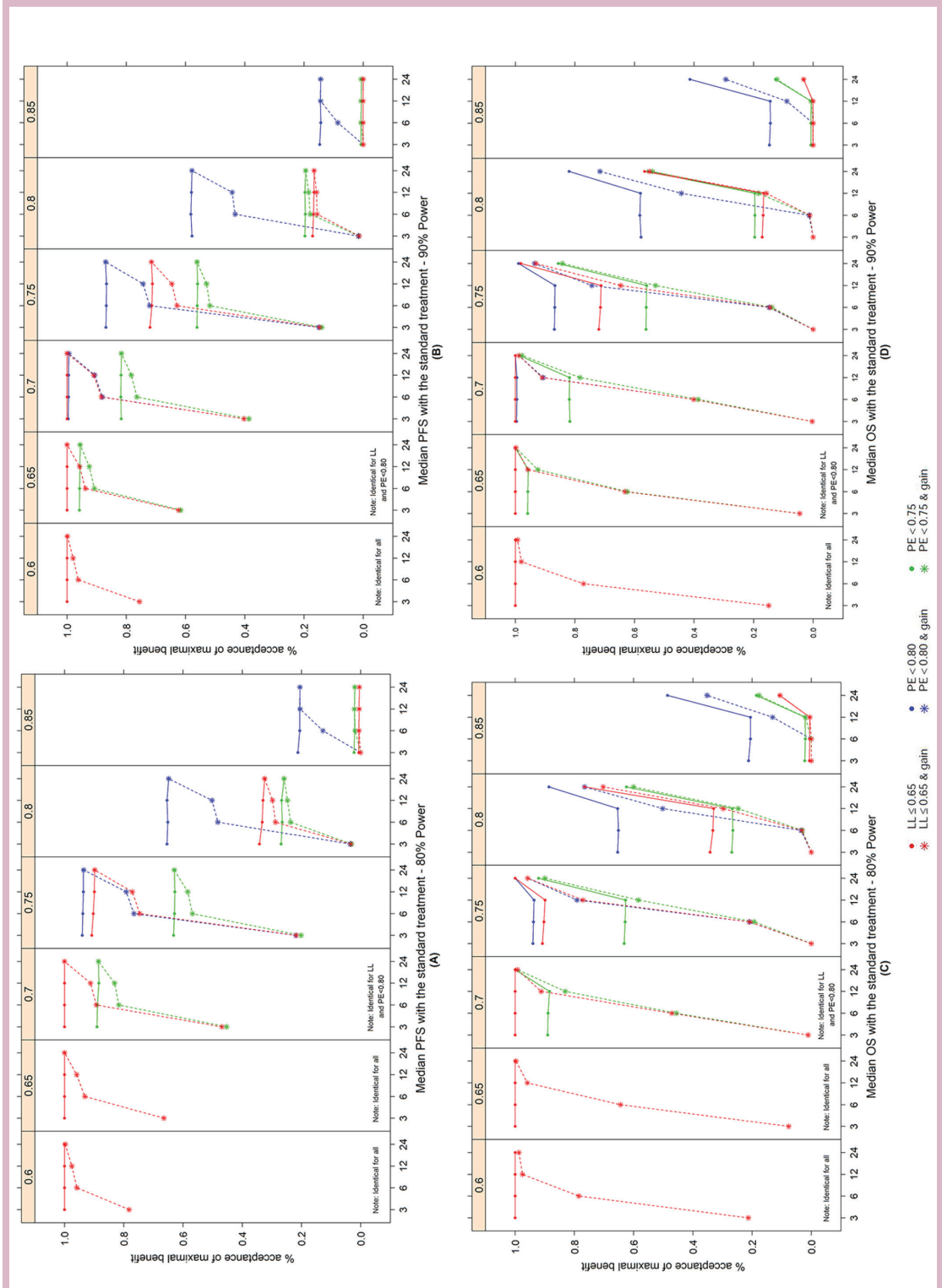


Figure 6 Comparison of the ESMO-MCBS dual rule and the dual rule with PE thresholds exhibiting similar behaviour: %acceptance of maximal RB (solid line: RB rule) and maximal preliminary grade (dotted line: dual rule) over significant trials for different true HRs.
(A-B) PFS, 80% power - 90% power. *Note:* Median control PFS ≤6 months: LL95%CI ≤0.65 rule vs PE<0.75 and PE<0.80; and gain ≥1.5 months. Median control PFS >6 months: LL95%CI ≤0.65 rule vs PE<0.75 and PE<0.80; and gain ≥3 months.
(C-D) OS, 80% power - 90% power. *Note:* Median control OS ≤12 months: LL95%CI ≤0.65 rule vs PE<0.75 and PE<0.80; and gain ≥3 months. Median control OS >12 months (24 months): LL95%CI ≤0.70 rule vs PE<0.80 and PE<0.83; and gain ≥5 months.
ESMO, European Society for Medical Oncology; MCBS, Magnitude of Clinical Benefit Scale; PE, point estimate; RB, relative benefit; PFS, progression-free survival; OS, overall survival.

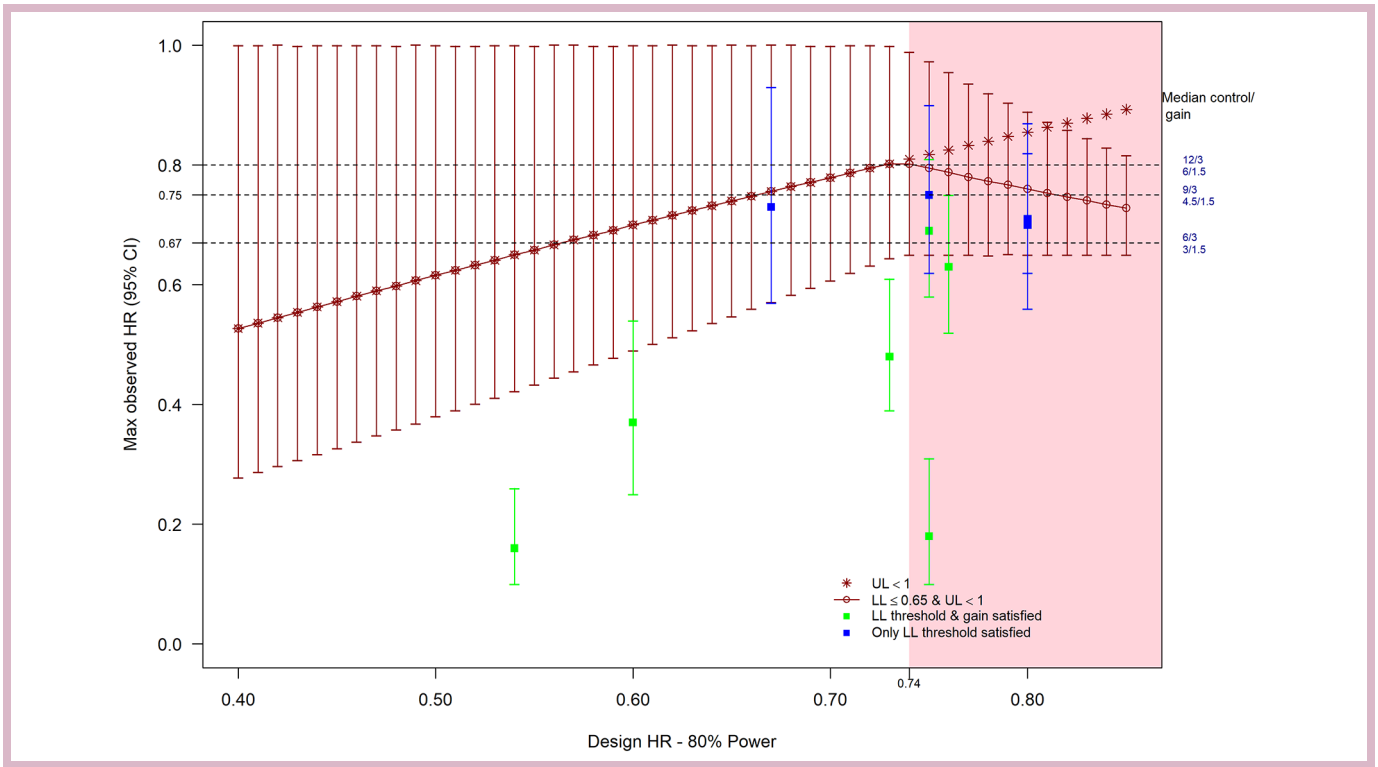


Figure 7 Field testing results relative to the maximal preliminary grade and maximum observed HR (95% CI) for primary outcome PFS—study power 80%.

Annotation: *Squares:* actual study information (y-axis: study observed HR and 95%CI; x-axis: study design HR); *Green squares:* study meeting both maximal RB & AB criteria; *Blue squares*: study meeting maximal RB but not AB criteria; *Red squares:* study meeting neither criterion; *Other* symbols, lines and shaded areas as in Fig. 2.

PFS, progression-free survival; RB, relative benefit; AB, absolute benefit; UL, upper limit; LL, lower limit.

As an example, 10 trials evaluated by the ESMO-MCBS v1.0, with PFS and median control ≤12 months targeting a design HR ranging from 0.54 to 0.80 with power 80%, are shown in figure 7. Among them, six satisfied both the RB and AB criteria for maximal preliminary grade, while four satisfied the RB criterion but not the AB (figure 7- online supplementary table S1).

Information is presented for 54 superiority trials involving nine tumour types in non-curative disease with adequate information on the study design (see online supplementary tables S1–4). Of note, all trials are targeting HR up to 0.80, with the exception of one trial (PREVAIL) that targets HR=0.83 with primary endpoint OS and median control survival above 12 months and, in fact, this study stopped at the interim analysis with HR=0.71 (95% CI 0.60 to 0.84) (see online supplementary figure S4C and table S3).

Overall, the balance between the RB and the AB rule, the first being more lenient and the latter generally stricter, appears to provide for a fair ESMO-MCBS grade assignment to the experimental treatments, as shown both through simulations and in the field testing results.

DISCUSSION

Credibility and validity are critically important for the development and value of the ESMO-MCBS, and ESMO has emphasised its compliance with standards for 'accountability for reasonableness' in the development process of the scale.[13 14] The validity of the ESMO-MCBS is derived from: (1) clinically relevant and reasonable criteria for prioritisation of different types of benefit, that is, that cure takes precedence over deferral of death, that direct endpoints such as OS and QoL take precedence over less reliable surrogates such as PFS or RR and that the interpretation of the evidence for benefit derived from indirect primary outcomes (such as PFS or RR) may be influenced by secondary outcome data; (2) coherence: procedural agreements regarding the evidence to be used/not used, how it will be analysed and evaluated and precautions to minimising bias (including conflict of interest issues) based on an understanding of the relative strengths and weaknesses of the usual measured outcomes OS and QoL, and their surrogates, as well as rigorous biostatistical review; (3) wide applicability over a range of solid cancers and a range of prognoses that have been rigorously tested; (4) statistical validity; and (5) a transparent process of development with scope for peer review, appeal and revision.[1]

Commitments to transparency and statistical validity underscore this unprecedented detailing of the statistical deliberations essential to the structure and workings of the ESMO-MCBS v1.0 and which will be continued in v1.1.[1]

The ESMO-MCBS aimed to have characteristics of inclusiveness and discernment. Inclusiveness refers to the quality that if a trial has adequate power and targets a clinically meaningful RB, then if found significant, the experimental treatment should not be penalised unfairly. Discernment, in contrast, refers to the ability to identify the situation where, if a trial is unjustifiably large and has adequate power for a small benefit gain that is less than the *clinically meaningful benefit,* then if found significant, the experimental treatment should be somehow penalised.

The choice of the LL95%CI as a threshold for assigning the MCBS score, first and foremost, serves the purpose of taking into account the variability of the estimate. The alternative of using the PE fails in that regard. The use of the LL95%CI for threshold evaluation has been the source of confusion, consternation and scepticism, with the often stated concern that it would increase the likelihood of small benefit studies being overcredited with greater benefit than is justifiable.[3–9] The data generated by the extensive simulations and field testing presented in this paper are reassuring, and they indicate that the use of the LL95%CI approach gives the best balance of minimising false-negative results (inclusivity) and minimising false-positive results (discernment) in the identification of studies reaching thresholds for maximal grade in the preliminary scoring of the ESMO-MCBS v1.0. In addition, the use of the LL95%CI demonstrates desired discriminatory behaviour for different powers, true HR values and different design HRs better than PE approaches.

When comparing the ESMO-MCBS scoring approach to an approach with PE thresholds exhibiting the closest behaviour to the chosen MCBS LL95%CI cut-off values (PE <0.75, 0.80 or 0.83), it was again demonstrated that for a range of true HRs, the MCBS rule was overall more sensitive and more specific. This advantage held whether the design was adequately applied for the true benefit (design HR=true HR), or not. In addition, this advantage was much more pronounced, when lower cut-off values, such as those proposed by Sobrero *et al* (HR=0.60, 0.65), were used.[10]

Finally, the concern that by taking into account the variability of the estimate, the LL95%CI threshold would penalise larger studies and promote smaller ones, was shown not to hold. In fact, the LL95%CI rule had a higher %acceptance for reasonable benefit trials, when designed with 90% power than when designed with 80% power. For smaller true benefit, this is reversed (ie, achieving a higher %acceptance for studies designed with 80% vs 90% power). Using the LL95%CI threshold, this reversal occurs at true HRs <0.73, or 0.77 for LL95%CI ≤0.65 and ≤0.70, respectively, and at an acceptance rate of approximately 80% for maximal benefit grade. When using a PE threshold, this reversal occurs at that PE threshold value and at an acceptance rate limited to 50% for maximal grading. Of note, this means that for trials evaluated based on the previously advocated PE cut-offs of 0.60 or 0.65, this reversal is present for much bigger RB values, which is a non-desirable property.[10]

Together these simulations demonstrate that the MCBS dual rule incorporating the LL95%CI threshold for HR addresses both goals of inclusiveness and discernment more effectively than a dual rule using PE thresholds. Inclusiveness: RB assessment using the LL95%CI for HR rather than a PE threshold avoids excluding a substantial proportion of big benefit positive studies from achieving due credit as would be the case if the PE were to be used. Discernment: the combined RB using the LL95%CI and AB assessment results in downgrading more trials with a statistically significant but clinically insignificant observed benefit. The stricter behaviour occurs earlier when power is 90% vs 80%, and this is in fact the desired behaviour. It results in not assigning maximal preliminary grade to more trials found significant based on a relatively small observed benefit (HR close to the null) due to a large sample size.

The dual rule incorporating both AB and RB is not unique to the ESMO-MCBS; it is also used in the approach advocated by Sobrero *et al*.[10] Among the magnitude of clinical benefit and value scales published thus far, only the ESMO-MCBS uses the LL95%CI rather than PE for evaluation of RB.[10 15–17]

A limitation of the exploration of the ESMO-MCBS v1.0 through both the field testing and the simulations reported here is that it is performed only for trials that make available the relevant design characteristics, that is, power and targeted RB, by including them in the published trial report.

Since the ESMO-MCBS aims to provide a structured, valid, reasonable and reproducible approach for data interpretation, the findings of this statistical exploration have implications for data interpretation in general while it promotes the importance of reporting HR CIs.[2 18]

Two important lessons were derived from these statistical explorations that will be salient for future revision of the ESMO-MCBS. The simulations identified a shortcoming in v1.0 such that when the control OS is greater than 24 months, there was no AB constraint on RB scoring. Second, it was shown that small-sized trials provide a relatively small but still existing risk to reach a false-positive conclusion. Both of these issues will be addressed in v1.1: the former with the introduction of a new prognostic subgroup for OS studies where the median OS of the control arm is >24 months; the latter with the specific identification of small randomized phase II studies and a disclaimer recommending that confirmation of the beneficial outcome based on larger trials is warranted.

Future research will evaluate the factors contributing to divergent grading outcomes when using the ESMO-MCBS and the ASCO Value Framework and the relative characteristics of threshold evaluation using LL95%CI as compared with the approach to RB evaluation of the German Institute for Quality and Efficiency in Health Care, which uses UL95%CI thresholds.[15 19 20] The outcome measures incorporated

in the grading system of the ESMO-MCBS are those prescribed in the CONSORT statements.[21 22] These standards are likely to evolve and, if so, ESMO-MCBS will be revised to remain constant with contemporaneous standards for reporting. Novel statistical approaches such as restrictive mean survival time have not yet gained CONSORT endorsement and are not widely reported.[23] Should they be incorporated into reporting standards, they will also be incorporated into future versions of the ESMO-MCBS.

CONCLUSION

This study illustrates the statistical rationale of the evaluation of RB thresholds using the LL95%CI in preference to a PE threshold. The simulations demonstrate that no set of rules can accommodate every possible situation and that the aims of scale development are to identify the best balance between incisiveness and discrimination while acknowledging that no approach will be perfect. The ESMO-MCBS is an evolving tool with underlying rules that will be regularly improved and adapted according to the results of repeated rigorous testing and feedback from users and stakeholders.

Author affiliations
[1]Laboratory of Biostatistics, School of Health Sciences, National and Kapodistrian, University of Athens, Athens, Greece
[2]Frontier Science Foundation-Hellas, Athens, Greece
[3]Department of Statistics, Athens University of Economics and Business, Athens, Greece
[4]Methodology Direction, EORTC Headquarters, Brussels, Belgium
[5]Department of Medical Oncology, Ioannina University Hospital, Ioannina, Greece
[6]Department of Medical Oncology, Vall d'Hebron University Hospital, Barcelona, Spain
[7]Division of Oncology, Medical University Vienna, Vienna, Austria
[8]Jules Bordet Institute, Universite Libre de Bruxelles, Bruxelles, Belgium
[9]Department of Medical Oncology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[10]ESMO Head Office, European Society for Medical Oncology, Lugano, Switzerland
[11]Department of Medical Oncology, Cancer Pain and Palliative Medicine Service, Shaare Zedek Medical Center, Jerusalem, Israel

Acknowledgements The authors wish to acknowledge the support and contribution of the ESMO Executive Board and extend their deepest gratitude in particular to Panagiota Zygoura, Katerina Vervita and Zoi Tsourti from Frontier Science Foundation-Hellas for their support.

Contributors NA.

Funding This project was funded by ESMO.

Competing interests The authors have declared the following: JB: director of EORTC. EORTC conducts many studies sponsored by, or otherwise supported by, a large number of companies. EORTC is an independent research organisation. MJP: board member: Radius. Consultant (honoraria): AstraZeneca, Lilly, MSD, Novartis, Pfizer, Roche-Genentech, Crescendo Biologics, Periphagen, Huya, Debiopharm and PharmaMar. Research grants to institute: AstraZeneca, Lilly, MDS, Novartis, Pfizer, Roche-Genentech, Synthon, Radius and Servier. Speakers bureau/stock ownership: none. GP: consulting and advisory services, and research support: Amgen, Merck, AstraZeneca, Roche, BMS, MSD and Lilly. J-YD: compensated participation to advisory boards, lecture and symposia: Amgen, Merck Serono, Bayer, Roche/Genentech, Sanofi, AstraZeneca, Boehringer-Ingelheim and Sirtex until April 2016. No further compensated participation in industry events from May 2016 onwards. JT: advisory boards for Amgen, Bayer, Boehringer Ingelheim, Celgene, Chugai, Genentech, Lilly, MSD, Merck Serono, Novartis, Pfizer, Roche, Sanofi, Symphogen, Taiho and Takeda. CCZ: honoraria: AstraZeneca, Celgene, Roche, Novartis, Bristol Myers Squibb, MSD, Ariad and Newgen. EGEdV: currently conducting research sponsored by the following companies: Amgen, Roche/Genentech, Chugai Pharma, Synthon, AstraZeneca, Radius Health, CytomX Therapeutics and Nordic Nanovector (all payments to the institution). Consulting or advisory role for the following companies: Synthon, Medication and Merck (all payments to the institution). Not a member of any speakers' bureau.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Cherny NI, Sullivan R, Dafni U, *et al*. A standardised, generic, validated approach to stratify the magnitude of clinical benefit that can be anticipated from anti-cancer therapies: the European Society for Medical Oncology Magnitude of Clinical Benefit Scale (ESMO-MCBS). *Ann Oncol* 2015;26:1547–73.
2. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986;292:746–50.
3. Hartmann M. The ESMO magnitude of clinical benefit scaling tool: from theory to practice. *Ann Oncol* 2015;26:2357–8.
4. Muhonen T, Joensuu H, Pfeiffer P. Comment on ESMO Magnitude of Clinical Benefit Scale. *Ann Oncol* 2015;26:mdv384.
5. Wild C, Grössmann N, Bonanno PV, *et al*. Utilisation of the ESMO-MCBS in practice of HTA. *Ann Oncol* 2016;27:2134–6.
6. Cheng S, McDonald EJ, Cheung MC, *et al*. Do the American Society of Clinical Oncology Value Framework and the European Society of Medical Oncology Magnitude of Clinical Benefit Scale Measure the Same Construct of Clinical Benefit? *Journal of Clinical Oncology* 2017;35:2764–71.
7. Lindgren P, Jönsson B, Wilking N. Assessing The ESMO Magnitude of Clinical Benefit Scale from a Health Economics Perspective. *Value Health* 2015;18:A569.
8. Del Paggio JC, Azariah B, Sullivan R, *et al*. Do Contemporary Randomized Controlled Trials Meet ESMO Thresholds for Meaningful Clinical Benefit? *Ann Oncol* 2017;28:157–62.
9. Sobrero A. The hard road to ranking the clinical benefit of antineoplastic agents: ESMO Award 2016 presentation. *ESMO Open* 2017;2:e000157.
10. Sobrero AF, Pastorino A, Sargent DJ, *et al*. Raising the bar for antineoplastic agents: how to choose threshold values for superiority trials in advanced solid tumors. *Clin Cancer Res* 2015;21:1036–43.
11. Ellis LM, Bernstein DS, Voest EE, *et al*. American Society of Clinical Oncology perspective: Raising the bar for clinical trials by defining clinically meaningful outcomes. *J Clin Oncol* 2014;32:1277–80.
12. Core Team R. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2014. http://www.R-project.org/
13. Daniels N. Decisions about access to health care and accountability for reasonableness. *J Urban Health* 1999;76:176–91.
14. Daniels N. Accountability for reasonableness. *BMJ* 2000;321:1300–1.
15. Schnipper LE, Davidson NE, Wollins DS, *et al*. Updating the American Society of Clinical Oncology Value Framework: Revisions and Reflections in Response to Comments Received. *J Clin Oncol* 2016;34:2925–34.
16. Carlson RW, Jonasch E. NCCN Evidence Blocks. *J Natl Compr Canc Netw* 2016;14:616–9.
17. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines) with NCCN Evidence Blocks™, 2015. http://www.nccn.org/evidenceblocks/ (accessed Jan 2016).
18. Kyriacou DN. The Enduring Evolution of the P Value. *JAMA* 2016;315:1113–5.
19. Schnipper LE, Davidson NE, Wollins DS, *et al*. American Society of Clinical Oncology Statement: A Conceptual Framework to Assess the Value of Cancer Treatment Options. *J Clin Oncol* 2015;33:2563–77.

20. Skipka G, Wieseler B, Kaiser T, *et al*. Methodological approach to determine minor, considerable, and major treatment effects in the early benefit assessment of new drugs. *Biom J* 2016;58:43–58.
21. Schulz KF, Altman DG, Consort MD. statement: updated guidelines for reporting parallel group randomized trials. *Ann Int Med* 2010;2010:726–32.
22. Piaggio G, Elbourne DR, Pocock SJ, *et al*. Reporting of noninferiority and equivalence randomized trials: extension of the CONSORT 2010 statement. *JAMA* 2012;308:2594–604.
23. Uno H, Claggett B, Tian L, *et al*. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol* 2014;32:2380–5.