ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ESMO OPEN CANCER HORIZONS

REVIEW

Biases in study design, implementation, and data analysis that distort the appraisal of clinical benefit and ESMO-Magnitude of Clinical Benefit Scale (ESMO-MCBS) scoring

B. Gyawali[1,2,3*], E. G. E. de Vries[4], U. Dafni[5,6], T. Amaral[7], J. Barriuso[8], J. Bogaerts[9], A. Calles[10], G. Curigliano[11,12], C. Gomez-Roca[13], B. Kiesewetter[14], S. Oosting[4], A. Passaro[15], G. Pentheroudakis[16], M. Piccart[17], F. Roitberg[18,19], J. Tabernero[20], N. Tarazona[21], D. Trapani[12], R. Wester[22], G. Zarkavelis[23], C. Zielinski[24], P. Zygoura[6] & N. I. Cherny[25]

Departments of [1]Oncology; [2]Public Health Sciences, Queen's University, Kingston, Ontario; [3]Division of Cancer Care and Epidemiology, Queen's University, Kingston, Ontario, Canada; [4]Department of Medical Oncology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; [5]Laboratory of Biostatistics, School of Health Sciences, National and Kapodistrian University of Athens, Athens; [6]Frontier Science Foundation-Hellas, Athens, Greece; [7]Skin Cancer Center, Department of Dermatology, Eberhard Karls University, Tuebingen, Germany; [8]The Christie NHS Foundation Trust and Division of Cancer Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK; [9]European Organisation for Research and Treatment of Cancer, Brussels, Belgium; [10]Medical Oncology Department, Hospital General Universitario Gregorio Marañon, Madrid, Spain; [11]Department of Oncology and Hemato-Oncology, University of Milan, Milan; [12]European Institute of Oncology, IRCCS, Milan, Italy; [13]Institut Universitaire du Cancer de Toulouse (IUCT), Toulouse, France; [14]Division of Oncology, Department of Medicine I, Medical University of Vienna, Vienna, Austria; [15]Division of Thoracic Oncology, European Institute of Oncology, IRCCS, Milan, Italy; [16]ESMO Head Office, Lugano, Switzerland; [17]Jules Bordet Institute, Université Libre de Bruxelles, Brussels, Belgium; [18]WHO Cancer Management Consultant, Geneva, Switzerland; [19]Instituto do Cancer do Estado de São Paulo (ICESP HCFMUSP), São Paulo, Brazil; [20]Vall d'Hebron Hospital Campus and Institute of Oncology (VHIO), UVic-UCC, IO-Quiron, Barcelona; [21]Department of Medical Oncology, Biomedical Research Institute INCLIVA, CIBERONC, University of Valencia, Valencia, Spain; [22]Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands; [23]University of Ioannina-Department of Medical Oncology, Ioannina, Greece; [24]Central European Cooperative Oncology Group and Central European Cancer Center, Wiener Privatklinik, Vienna, Austria; [25]Cancer Pain and Palliative Medicine Service, Department of Medical Oncology, Shaare Zedek Medical Center, Jerusalem, Israel

Available online 20 April 2021

Background: The European Society for Medical Oncology-Magnitude of Clinical Benefit Scale (ESMO-MCBS) is a validated, widely used tool developed to score the clinical benefit from cancer medicines reported in clinical trials. ESMO-MCBS scores assume valid research methodologies and quality trial implementation. Studies incorporating flawed design, implementation, or data analysis may generate outcomes that exaggerate true benefit and are not generalisable. Failure to either indicate or penalise studies with bias undermines the intention and diminishes the integrity of ESMO-MCBS scores. This review aimed to evaluate the adequacy of the ESMO-MCBS to address bias generated by flawed design, implementation, or data analysis and identify shortcomings in need of amendment.

Methods: As part of a refinement of the ESMO-MCBS, we reviewed trial design, implementation, and data analysis issues that could bias the results. For each issue of concern, we reviewed the ESMO-MCBS v1.1 approach against standards derived from Helsinki guidelines for ethical human research and guidelines from the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, the Food and Drugs Administration, the European Medicines Agency, and European Network for Health Technology Assessment.

Results: Six design, two implementation, and two data analysis and interpretation issues were evaluated and in three, the ESMO-MCBS provided adequate protections. Seven shortcomings in the ability of the ESMO-MCBS to identify and address bias were identified. These related to (i) evaluation of the control arm, (ii) crossover issues, (iii) criteria for non-inferiority, (iv) substandard post-progression treatment, (v) *post hoc* subgroup findings based on biomarkers, (vi) informative censoring, and (vii) publication bias against quality-of-life data.

Conclusion: Interpretation of the ESMO-MCBS scores requires critical appraisal of trials to understand caveats in trial design, implementation, and data analysis that may have biased results and conclusions. These will be addressed in future iterations of the ESMO-MCBS.

Key words: ESMO-MCBS, bias, clinical trial design, clinical trial implementation, clinical trial reporting, clinical trial analysis

*Correspondence to: Dr Bishal Gyawali, Division of Cancer Care and Epidemiology, Queen's University Cancer Research Institute, 10 Stuart Street, Kingston, Ontario K7L 3N6, Canada. Tel: +1-613-533-6000x78509; Fax: +1-613-533-6794

E-mail: gyawali.bishal@queensu.ca (B. Gyawali).

2059-7029/© 2021 The Author(s). Published by Elsevier Ltd on behalf of European Society for Medical Oncology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

INTRODUCTION

The European Society for Medical Oncology-Magnitude of Clinical Benefit Scale (ESMO-MCBS) was first published in 2015 and revised in 2017.[1,2] With a growing recognition that many cancer medicines provided modest benefits disproportionate to their high costs, the oncology community needed a tool that could objectively assess the clinical benefit from cancer medicines, assist in comparison with other similar medicines, and guide regulatory and reimbursement decisions. The ESMO-MCBS was established to address these needs.[1,2] To reduce bias and error in grading, the scale has been developed in close adherence to the principles of 'accountability for reasonableness',[3] a standard for ethical public health decision-making processes.

The ESMO-MCBS aims to highlight treatments with a substantial level of clinical benefit for patients and distinguish those from studies demonstrating only moderate, minor, or marginal clinical benefit. Within ESMO, the ESMO-MCBS is used in clinical practice guidelines and provides a structured approach to evaluate clinical research data. On its website, ESMO has an open access searchable portal detailing >230 clinical studies (Scorecards) assessed using the ESMO-MCBS.[4] Internationally, a high ESMO-MCBS score is currently valued and adopted by the World Health Organization Essential Medicines List (WHO EML) and Health Technology Assessment bodies worldwide. These global health applications underscore the importance of the ESMO-MCBS commitments to 'accountability for reasonableness' and continual efforts to improve the scoring process's validity.

ESMO-MCBS scores assume valid research methodologies and high-quality trial implementation. Studies that incorporate flawed design, implementation, and/or data analysis may generate biased outcomes and conclusions that exaggerate real benefit and are not generalisable. This subverts the intention of the ESMO-MCBS to give representative grading to the benefit observed in generalisable data and compromises its integrity.

Therefore, as part of the ongoing commitment to improving the validity of the scoring process, we undertook a review of trial design, implementation, and analysis issues that could bias the results and reviewed the adequacy of the ESMO-MCBS v1.1 to address these issues and identify shortcomings to redress in future revisions.

METHODOLOGY

Based on experience in evaluating the magnitude of benefit in clinical studies, ESMO-MCBS Working Group and Extended Working Group members (all listed in authorship) identified issues in study design, implementation, and data analysis that may influence study outcomes and compromise the veracity of the ESMO-MCBS scores. We conducted a review for each of these issues, including definitions, relevant policy documents derived from regulatory authorities, relevant literature, and illustrative studies. The policy documents included the World Medical Association Helsinki Declaration for Ethical Principles for Human Research,[5] and guidelines from the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH),[6-8] the Food and Drugs Administration (FDA),[9-11] the European Medicines Agency (EMA),[12-14] and the European Network for Health Technology Assessment.[15-19] For each issue we reviewed the ESMO-MCBS v1.1 approach to identify shortcomings of the scale to adequately address and document the corresponding sources of bias.

RESULTS

Design issues

Six issues in study design that could bias benefit evaluation were considered (Figure 1).

Substandard control arm

Rationale: Data derived from studies with a comparator (control) arm inferior to the standard of care (SOC), may bias the outcome by generating a larger benefit than if SOC had been used.[8,10,12,16]

Regulations: According to the Helsinki Declaration,[5] the comparator arm of a randomised, clinical trial (RCT) must be 'the best-proven intervention(s)'. The ICH guidelines emphasise the importance of using appropriate dosing and scheduling of the control.[8]

The Helsinki Declaration allows two exceptions[5]: (i) when no proven intervention exists and (ii) when there are compelling and scientifically sound methodological reasons for using a less than best-proven control therapy. The Helsinki Declaration allows the use of placebo, no intervention, or a lesser SOC if deemed necessary to determine an intervention's efficacy or safety. However this is only permitted on the condition that subjects receiving the control arm will not be subject to additional risks of serious or irreversible harm. The guidelines add the admonition that 'extreme care must be taken to avoid abuse of this option.' For non-inferiority (NI), the ICH emphasises that the control arm should comprise 'a drug acceptable in the region to which the studies will be submitted (for licensing) for the same indication'.[6]

Therefore, it is incumbent upon researchers to demonstrate that the control arm is consistent with the SOC at study initiation or that any deviation is adequately justified. The justification must present compelling and scientifically sound methodological reasons for the deviation and that participants will not be subject to serious harm. Institutional Review Boards (IRBs) are responsible for ensuring compliance with these conditions.[5] For registration trials, this adjudication is often guided by the regulatory agencies themselves.

Illustrative case: The NEMO study in treatment-naive or pretreated patients with advanced NRAS-mutated melanoma randomised 402 participants in a 2 : 1 ratio, between August 2013 and April 2015, to receive binimetinib or dacarbazine.[20] Seventy-nine percent of the participants

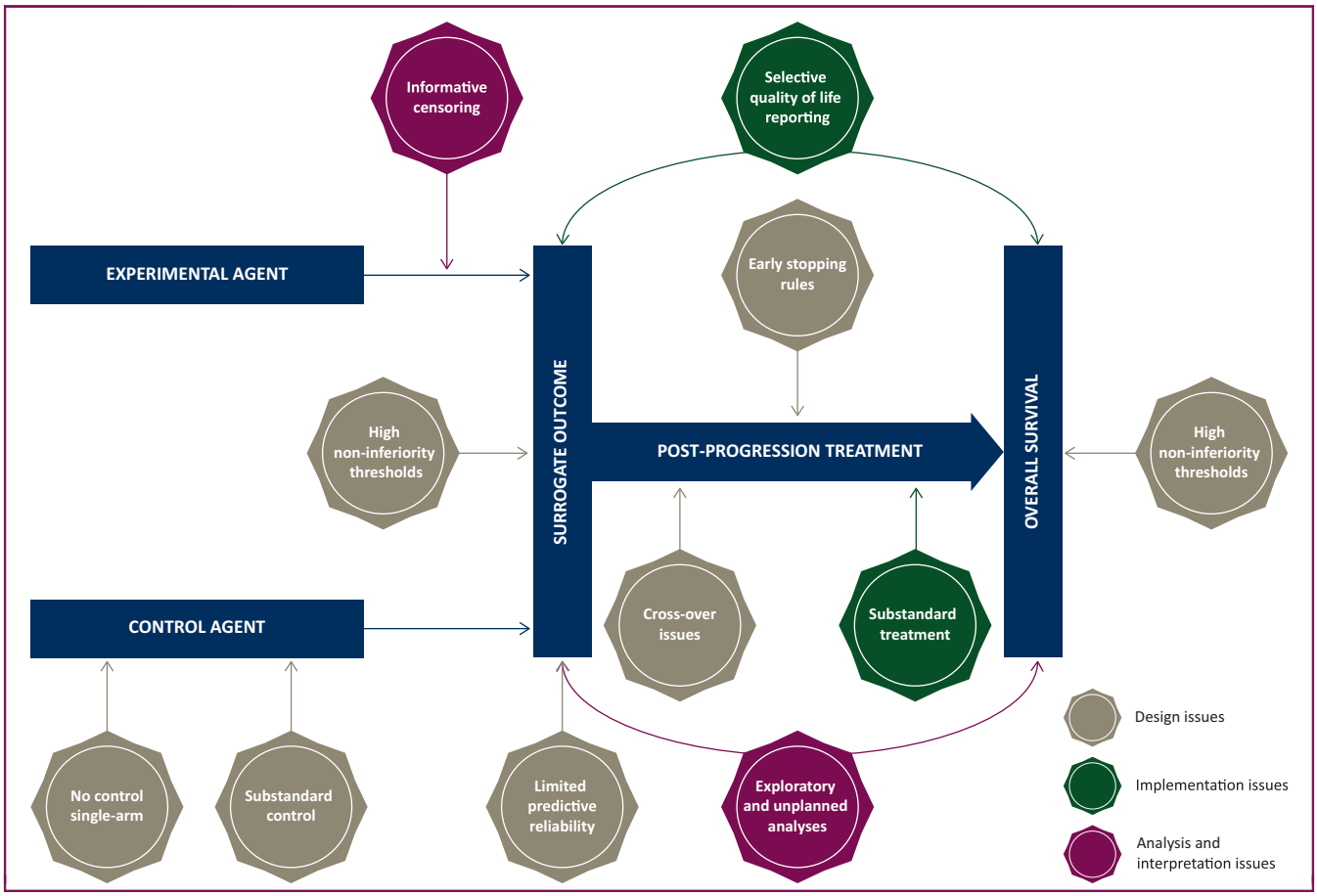

Figure 1. Issues in study design, implementation, and data analysis that may influence study outcomes and compromise the ESMO-MCBS scores.
HR, hazard ratio; NI, non-inferiority; QoL, quality of life.

were treatment-naive. Dacarbazine, the control arm for treatment-naive patients, was already proven to be inferior to ipilimumab immunotherapy plus dacarbazine.[21] Ipilimumab monotherapy was subsequently licensed as first-line treatment in 2011 by both the EMA[22] and FDA.[23] Consequently, patients in the control arm were deprived of the best, licensed upfront treatment, and in the first-line setting the marginal benefit of binimetinib was only demonstrated relative to a suboptimal comparator.

ESMO-MCBS v1.1: ESMO-MCBS relies on the integrity of the IRB and regulatory agencies to evaluate the control arm's adequacy.

Shortcoming: The ESMO-MCBS does not independently evaluate the control arm's appropriateness, nor does it have a mechanism to either indicate or penalise studies with a substandard control arm.

The predictive reliability of surrogate endpoints

Definitions: Surrogate outcome endpoints provide an indirect measurement when direct measurement of clinical effect is not feasible or practical.[8] While they aim to predict clinical benefits such as prolonged survival or improved quality of life (QoL), the reliability and strength of surrogates' predictive capacity vary.[24] The effect of an improved surrogate endpoint may not directly benefit the patient.[24] Commonly used surrogate outcomes in cancer trials include a decrease in tumour size response rate (RR) and delays in tumour progression [progression-free survival (PFS); disease-free survival (DFS)].[10,12,19]

Limitations of surrogate outcomes: The validity of a surrogate outcome depends on its reliability as a predictor of true clinical benefit, i.e. longer survival or improved QoL.[8,10,12,19] Hitherto, no outcome measure in oncology has been found to have absolute surrogacy for true clinical benefit across diseases and treatments.[25-29] As stated by the ICH, there is concern that they may not reliably predict clinical benefit.[7]

Evaluation of DFS as a surrogate for overall survival (OS) in adjuvant therapy studies, found that predictive reliability is variable across diseases and, overall, it is at best characterised as moderate.[25,27,30,31] Even within the same tumour type, there may be differences in predictive reliability of DFS based on tumour subtypes: for example, DFS is a better surrogate for OS in HER2-positive breast cancer than for other breast cancer subtypes.[30] In studies evaluating therapies in non-curative settings, PFS and time to progression provide information about the biological activity and may indicate the possibility of benefit to patients.[29,32] However, they are not reliable surrogates for improved OS[31-36] or QoL[36,37] in all

patients. RR and pathological complete response (pCR) rate are also weak predictors of improved OS.[25,30]

ESMO-MCBS v1.1: ESMO-MCBS v1.1 considers surrogacy in its weighting. Using ESMO-MCBS form 1, DFS scores are only creditable in the adjuvant setting if OS data are immature. If mature OS results do not demonstrate benefit, surrogacy is not confirmed, and the study is considered to not provide evaluable benefit (labelled 'No evaluable benefit'). Studies showing benefit based on pCR are credited at the lowest level, C, and only if a relatively high threshold marginal benefit is demonstrated.

In the non-curative setting, when the primary endpoint is PFS or RR, several stringencies are applied. The preliminary grades are capped: for studies using PFS as primary endpoint at 3 and for RR at 2, with penalties for adverse effects. Furthermore, when PFS is the primary endpoint a non-significant OS gain at mature follow-up and QoL evaluation indicating neither improvement nor delayed deterioration is considered as refutation of surrogacy, and the score is downgraded by one point.

Shortcoming: Hitherto, in v1.1, it was assumed that DFS did not confer patient benefit independent of OS. The approach of ESMO-MCBS v1.1 to the grading of DFS was recently reviewed and considered unreasonable.[38] Patients and other stakeholders appealed that the ESMO-MCBS approach to DFS does give credit to the benefit of added time without treatment or the burden of disease for a proportion of patients independent of any impact (or lack thereof) on mature OS.[39] This is illustrated by the meta-analysis of trastuzumab in HER2 overexpressed, hormone receptor-negative early breast cancer with less than two involved nodes. After a median of 8 years follow-up, there was a 5.9% gain in DFS, but the OS gain was not significant.[40] The ESMO-MCBS Working Group has concluded that DFS is an intermediate endpoint (i.e. a surrogate endpoint that may also directly have some patient benefits) that is worthy of a lower but persistent credit if OS benefit is not achieved. This consideration is incorporated in the draft revision of the ESMO-MCBS v2, and it is currently undergoing field testing and review.

Crossover

Definitions: In an RCT, crossover implies patients randomised to the control arm of the trial get the intervention allocated to the experimental arm upon disease progression. Crossover has methodological and ethical implications, depending on the medicine and line of therapy.[41,42]

When a medicine has already been approved, is the SOC for later lines, and is being evaluated for an earlier line, the trial design should incorporate crossover. This is called appropriate or desirable crossover.[41,43] In such situations, since the experimental therapy is part of subsequent standard care, the clinical question is whether using the same drug earlier improves OS versus using it later in the disease course. Failure to incorporate crossover in this setting harms participants on the control arm by not ensuring that they receive optimal post-progression therapy and may exaggerate the observed OS benefits.

If a medicine, never approved for a condition, is being tested in a trial, then crossover design is generally undesirable.[41-43] Since the new medicine's efficacy is unknown, there is no ethical mandate for the control arm patients to receive the medicine upon relapse.[42] Furthermore, crossover in this setting undermines the ability to determine the impact of the intervention on OS, and if crossover delays initiation of proven subsequent therapies, it may adversely impact patient well-being. For these reasons, crossover in this setting is discouraged by the EMA and FDA.[10,12]

Illustrative cases: *Failure to incorporate appropriate crossover.* Abiraterone acetate was approved for use in patients with chemotherapy-naive metastatic castration-resistant prostate cancer (CRPC) in 2012 and has become the SOC in that setting based on the COU-AA-302 trial showing prolonged OS.[44,45] Between 2013 and 2014, abiraterone was tested versus placebo in chemotherapy-naive patients with castration-sensitive prostate cancer in the LATITUDE trial.[46] In that study, only 11% of patients on the placebo arm received abiraterone upon progression to CRPC. A substantial OS benefit {hazard ratio (HR) 0.66 [95% confidence interval (CI): 0.56-0.78]} generated a high ESMO-MCBS score of 4. However, due to the lack of crossover, we do not know whether using abiraterone earlier while the tumour is castration-sensitive is better than using the same drug while castration-resistant. Furthermore, since abiraterone had improved OS for patients with CRPC, the control arm patients were potentially harmed by not receiving a proven post-progression therapy.

Incorporation of undesirable crossover: In the IMPACT trial, which randomised patients with low volume metastatic CRPC to the autologous dendritic cell therapeutic vaccine sipuleucel-T, or placebo,[47] patients who progressed on the control arm were allowed a frozen version of the vaccine, even though its efficacy had not been proven. Outside the trial, these patients would have immediately received docetaxel chemotherapy that had previously demonstrated survival advantage and improved QoL in this setting.[48] In the study, treatment with sipuleucel-T did not affect RR or PFS compared with placebo, but it was associated with improved OS. The crossover of 64% patients in the control arm to the frozen vaccine version confounded interpretation of the findings since it was uncertain whether prolonged survival was because of treatment efficacy in the experimental arm or delayed access to docetaxel in the control arm.[49]

ESMO-MCBS v1.1: ESMO-MCBS Scorecards indicate whether crossover is allowed or not allowed.

Shortcoming: The ESMO-MCBS does not have a mechanism to either indicate or penalise studies with inappropriate or inadequate crossover.

Early stopping of clinical trials

Definition: Early stopping rules allow for a study to terminate earlier than planned, with all patients crossing to the superior therapy, because of the result of an interim analysis showing larger than expected benefit or harm of the experimental intervention that adequately undermines equipoise.[8,12] These stopping boundaries are stringent and based on solid statistical methodology.[8,12] Cancer drug trials may be stopped early based on an interim analysis of time-to-event probability (DFS, PFS, or OS) when the HR crosses the stopping boundary.

Concern: Under the statistical rules applied, trials that are stopped early may overestimate the magnitude of benefit. The sooner the trial is stopped, the more impressive the HR will look since the stopping criteria are more stringent early in the trial course.[50] Hence, although the medicine is likely effective, the true benefit may be smaller in magnitude. Such overestimations of the treatment effect's magnitude are particularly important when the primary endpoint is not a definitive endpoint like OS but a surrogate endpoint such as PFS.[50]

ESMO-MCBS v1.1: In solid tumours, PFS is scorable only if the median PFS of the control arm has been reached. Consistent with EMA guidance,[12] there is no extra credit for early stopping based on PFS. If, however, early stopping is triggered by interim analysis of OS gain meeting pre-specified statistical criteria, the gain already credited for PFS in the preliminary score is upgraded by one point.

Shortcoming: None identified.

Inflated RRs and durations in single-arm trials

Definitions: In settings where there is no available therapy and where measurable reduction in tumour size meeting the RECIST criteria[51,52] can be attributed to the tested medicine, regulatory authorities often accept overall RR (ORR) and duration of response (DoR) derived from single-arm studies as adequate evidence supporting accelerated approval,[10,12,17] and occasionally full (regular) approval.

Limitations of single-arm studies: Studies have shown that ORR and DoR in single-arm trials are higher than the ORR and DoR when the same medicine for the same indication is tested in an RCT.[53,54] Furthermore, ORR is a poor surrogate for OS or QoL.[25,30]

ESMO-MCBS v1.1: The scoring of single-arm studies using the ESMO-MCBS form 3 applies two stringencies. The preliminary score for single-arm studies is capped at 3, and penalties are applied for adverse events. The score may be upgraded by one point if the findings are confirmed in a phase IV study or cancelled if accelerated approval is subsequently withdrawn.

Shortcoming: None identified.

NI design trials

Definition: In some cases, an investigational product is tested not to show superiority over the SOC but to demonstrate that for the primary outcome, the new agent is not worse than the active control by more than a pre-specified small amount, known as an NI margin.[8,10-12] Benefit from the novel agent is demonstrated if it is less burdensome, less expensive, if it has less adverse effects, or if associated with improved QoL.[53]

Defining the NI margin is critical. According to ICH standards, the NI margin, expressed by an upper limit of the 95% CI for the relevant endpoint, is the largest difference that can be judged as clinically acceptable. Moreover, it should be less than the gain observed in superiority trials of the active comparator.[8]

Non-adherence to the assigned treatment is particularly problematic in NI studies since it will bias the study toward concluding NI.[11] Consequently, monitoring treatment adherence by investigators and by the independent data-monitoring committee is crucial in these studies. Therefore, unlike superiority studies, both an intention-to-treat (ITT) analysis and a per-protocol analysis are required by the FDA and EMA for NI studies.[8,11,14,55,56]

Concerns regarding NI margin: If the defined NI margin is too lenient, there is a concern that treatments with true inferiority may seem non-inferior. Regretfully, the biostatistical rules for defining NI have not been standardised.[57] A recent analysis showed that cancer medicine trials used an NI threshold as high as 1.33 for the upper limit of the 95% CI for the HR of OS.[53] Consequently, it is plausible that if NI definitions are too lenient, NI may be credited even when substantial differences in the treatment arms exist. If a previous superiority trial has demonstrated gains, a substantial percentage of these gains must be preserved.

ESMO-MCBS v1.1: ESMO-MCBS relies on IRB processes' integrity to evaluate the validity of the NI thresholds. NI studies can be scored using the ESMO-MCBS form 1 in the adjuvant setting (grade B) and form 2c in the advanced setting (grade 4). The ESMO-MCBS v1.1 only credits NI design trials if NI is confirmed according to pre-specified statistical criteria and if the study demonstrates benefits of reduced costs, adverse effects, or benefits in global QoL. NI alone is not the basis for any credit of benefit.

Shortcoming: ESMO-MCBS does not have rules to determine the validity of the pre-specified NI margin.

Study implementation issues

Two issues of study implementation and reporting were considered: (1) the impact of post-progression subsequent treatments on OS and (2) the publication bias in the reporting of QoL data (Figure 1).

Post-progression subsequent therapies

Definition: Most RCTs involve evaluating a single period of randomisation between a novel treatment and an active control. In studies of first- or second-line therapies in solid tumours, most patients will subsequently receive one or more lines of post-progression treatment, which influences OS.[58] In some settings, such as hormone-responsive breast cancer, it is not uncommon for patients to receive more than five subsequent therapy lines.[59]

When patients receive optimal post-progression therapy, any advantage gained by the experimental treatment may be impacted by subsequent therapies.[58] When the PFS gain is maintained or even improved after optimal post-progression therapies and reflected in an OS gain, the benefit is recognised as being important. However, when PFS gains are diluted after optimal post-progression therapies and reflected in no significant OS gain, the benefits may be relatively trivial. This, however, is not the case when patients also derived qualitative benefits such as delayed deterioration or improvement in QoL.[33]

Regulations: The ICH guidelines state that efforts should be made to collect all data pertinent to the relevant outcomes, including the occurrence and timing of intercurrent events.[7] They emphasise that clinical trials are less generalisable if the sponsor tries to avoid or minimise these issues. Post-progression treatments constitute an intercurrent event that is pertinent to OS.[58] While some degree of attrition may be expected post-progression, the acceptable thresholds should be judged based on previous experiences from real-world studies.

Concerns regarding post-progression treatments: Failure to provide optimal post-progression treatment can exaggerate the impact of a PFS gain on OS even when both arms receive the same suboptimal therapies.[41,58,60] This underscores the importance of documenting post-progression subsequent treatments until death as part of routine follow-up data.[58]

Illustrative case: The MONALEESA-7 study evaluated hormonal therapy with ribociclib or placebo in the first- or second-line treatment of premenopausal women with estrogen receptor-expressing breast cancer.[61] Patients receiving ribociclib had a PFS gain of 10.8 months. A planned interim analysis of OS at 76% of anticipated deaths showed a large OS gain that met pre-specified significance thresholds. Applying ESMO-MCBS v1.1, the MONALEESA-7 study achieved a preliminary score of 4, which was upgraded to 5 after QoL data demonstrated delayed deterioration in global QoL.[62]

The paper indicated that 26.8% of the patients in the control arm and 31.1% of patients in the ribociclib arm received no further subsequent treatments at disease progression after the first line of therapy.[61] Although some degree of attrition is expected with each subsequent line of therapy, nearly one-third of patients not getting any subsequent therapy post first-line is an astoundingly aberrant figure given that most women with estrogen receptor-positive HER2-negative breast cancer routinely survive for >2 years after first progression and generally receive four subsequent lines of therapy or more.[59] This major divergence from SOC for a substantial proportion of patients renders the OS data from this study non-generalisable. Indeed, it is plausible that the failure to provide subsequent standard therapy to more than a quarter of the patients who progressed on the study may have exaggerated the OS gain from ribociclib compared with placebo.

Shortcoming: The ESMO-MCBS does not indicate or penalise studies in which OS benefit may have been exaggerated by substandard post-progression treatment.

Publication bias in the reporting of QoL data

Definition: Publication bias occurs when the outcome of an experiment or research study influences the decision to publish or otherwise distribute it.[63]

Publication bias in QoL results: QoL data remain missing for many trials.[64] Most QoL data from trials go unpublished or are substantially delayed, even when the primary study results are positive.[65]

ESMO-MCBS v1.1: When QoL is evaluated as a secondary outcome in clinical studies, the generated results impact ESMO-MCBS scoring. When the QoL benefits are reported in studies applying a valid scale, with an adequately complete dataset and using valid statistical criteria, ESMO-MCBS scores are upgraded one point for evaluations in the non-curative setting. When the primary outcome is PFS with secondary outcomes of OS and QoL, and the subsequent mature OS does not demonstrate any survival advantage, the surrogacy of the PFS finding is dependent on the QoL results. In this scenario, a negative QoL finding without improvement or delayed deterioration in global QoL results in readjusting the score with a one point downgrade. Failure to publish negative QoL results or substantial publication delay subverts this important score adjustment.

Shortcoming: ESMO-MCBS does not address non-publication or delayed publication of QoL data.

Issues related to analysis of trial data

Two issues related to the analysis and interpretation of trial data were considered: (1) conjectural findings from exploratory and unplanned analyses and (2) informative censoring (Figure 1).

Conjectural findings from exploratory and unplanned analyses

Definition: A conjecture is an unproven proposition suspected to be true based on preliminary supporting evidence. 'Conjectural findings' relate to the evaluation of efficacy based upon incomplete or suboptimal data. These include findings from *post hoc* subgroup analyses or exploratory analyses outside of the statistical plan.

'Conjectural findings' contrast with 'confirmatory findings' derived from primary analysis in a study with a pre-specified and justified statistical plan and a significant positive outcome.[8] In many instances, subgroup analyses with appropriate adjustment for multiplicity of testing and alpha splitting are part of the planned confirmatory analysis and are incorporated into the statistical plan.[8]

The EMA guideline on the investigation of subgroups in confirmatory clinical trials[13] describes two types of conjectural analyses: (i) when the evidence of benefit in the primary analysis population is statistically significant but of small magnitude, it is of *post hoc* interest to identify and to distinguish between subgroups more or less likely to derive clinically meaningful benefit, and (ii) when a study fails to establish statistically significant evidence of benefit in the primary analysis population, and there is interest in identifying a subgroup where the treatment may be effective.

Concerns: Conjectural findings increase the probability of false-positive findings, i.e. the magnitude of clinical benefit is falsely concluded to be greater than in the primary analysis population.[9,13] False-negative conclusions, in which a subgroup is inaccurately identified as being unlikely to benefit, are equally important.

Regulations: The ICH guidelines,[8] endorsed by FDA and EMA, exhort that findings from *post hoc* subgroup analyses should be interpreted cautiously. The EMA guideline outlines a structured approach to conjectural evaluation based on (i) external evidence that the subgroup of interest is well defined and clinically relevant, (ii) plausible explanation for different efficacy (or risk—benefit) in a sub-population and its complement, (iii) substantially different results and, when possible (iv) replication of similar subgroup findings from other relevant trials.[13] In a draft guideline that is not yet ratified,[9] the FDA expresses the concern that investigators' or sponsors' incentives can influence the choice of analyses to identify one or more positive findings.[9]

ESMO-MCBS v1.1: The ESMO-MCBS v1.1 distinguishes confirmatory findings, based on the pre-specified endpoints and statistical plan, and conjectural findings, based on *post hoc* and exploratory analyses. Confirmatory findings of clinical benefit, including pre-specified subgroups, are scored. The ESMO—MCBS v1.1 constrains the number of pre-specified subgroups (no more than 3) and allows separate subgroups grading when adjusted for multiplicity.

Conjectural findings based on *post hoc* subgroup analyses and exploratory endpoints are not eligible for scoring by the ESMO-MCBS v1.1. An exception is made for studies that incorporate tissue samples collection to enable restratification based on plausible new genetic or other biomarkers. When conjectural findings form the basis for regulatory approval, the ESMO Clinical Practice Guidelines and E-Updates' approach is to present the ITT and planned subgroup data and scoring in the tables. The relevant conjectural data relating to the regulatory approval are discussed in the text and annotated below the ESMO-MCBS tabulations.

Illustrative cases: The APHINITY trial[66] tested adjuvant pertuzumab in patients with HER2-positive breast cancer and showed marginal gains in DFS for the ITT population. The publication, however, reported the findings of 12 *post hoc* subgroup analyses and highlighted better outcomes among patients who had node-positive disease. In this case, the ESMO-MCBS v1.1 scored only the ITT (score B) results and not the *post hoc* subgroup findings.

More recently, atezolizumab was tested combined with nab-paclitaxel in triple-negative breast cancer in the IMpassion130 trial.[67] The median PFS was improved by 1.7 months in the ITT population and by 2.5 months in patients with programmed death-ligand 1 (PD-L1)-positive tumours compared with nab-paclitaxel alone. There was no difference in OS in the ITT population. The statistical plan incorporated hierarchical testing, which allowed evaluation of OS in the PD-L1-positive subgroup only if there was OS benefit in the ITT population. An exploratory analysis of the PD-L1-positive subgroup found an OS improvement of 10 months. The ESMO-MCBS v1.1 only scored the PFS result of the PD-L1-positive subgroup, since the OS data were derived from an exploratory analysis outside of the statistical plan.

Two examples illustrate the importance of the ESMO-MCBS exception for *post hoc* subgroup findings based on enabling restratification based on plausible new genetic or other biomarkers. The IPASS trial identified the importance of the *EGFR* mutation status for treatment with gefitinib,[68] and the PRIME[69,70] and CRYSTAL[71] studies identified the importance of RAS/RAF status for anti-*EGFR* therapy in metastatic colorectal cancer.

Shortcoming: The ESMO-MCBS does not explicitly state that the exception for *post hoc* subgroup findings based on plausible new genetic or other biomarkers is restricted to findings resulting into a modification in licensed indication.

Informative censoring

Definition: In clinical trials, the term 'censoring' refers to patients who do not complete the study in full and drop out without further measurements.[72] When dropouts are balanced between the two arms of a comparative superiority study, it is assumed that this does not impact the results. This is called 'uninformative censoring'. When patients discontinue for reasons related to the study drug, including lack of effect or side-effects, this assumption does not hold, and this is referred to as 'informative censoring.'[72]

The problem of informative censoring: In studies using the surrogate outcomes of DFS and PFS, patients who stop treatment before documentation of disease progression for reasons other than death are at risk of no longer being evaluated. When censoring is greater in patients receiving the experimental therapy than in the control arm, censoring poorly performing patients may exaggerate the benefit seen in these outcome measures.[72-74]

Four approaches to mitigate this bias are described, including (i) encouraging OS rather than surrogates as the primary endpoint, (ii) comparing PFS/DFS gains with time-

to-treatment-failure (TTF) differences, which includes discontinuations as failures, (iii) listing the reasons for censoring, and (iv) providing best-case (assuming all censored patients do not have disease progression) and worst-case (assuming all censored patients have progressed) sensitivity analyses.[72-74]

Regulatory requirements: The ICH guidelines address this issue, stating that 'the frequency and type of protocol violations, missing values, and other problems should be documented in the clinical study report and their potential influence on the trial results should be described'.[8]

Illustrative cases: The BOLERO-2 study of exemestane combined with everolimus or placebo in hormone-positive advanced breast cancer[75] reported a 6.5 months benefit in median PFS with HR 0.36 (95% CI, 0.27-0.47) for patients receiving everolimus. This result was reasonably impacted by informative censoring since 19% patients in the everolimus arm discontinued treatment due to adverse effects versus 4% in the placebo arm (since treatment discontinuation due to adverse effects does not count as a PFS event). Reanalysing the study data using TTF which considers progression or discontinuation as well as death, the median gain in TTF was only 1.1 months[76] and the difference in OS, which is based on ITT analysis, was not significant.[77]

ESMO-MCBS v1.1: ESMO-MCBS v1.1 does not evaluate the causes and rates for censoring when evaluating trials with DFS or PFS primary endpoint. The draft revision of the ESMO-MCBS v2, currently undergoing field testing and review, incorporates a 1-point downgrade for PFS studies where there is a difference of $\geq$10% in prevalence of treatment discontinuations for adverse effects.

Shortcoming: The ESMO-MCBS does not account for the impact of informative censoring on scores based on DFS.

DISCUSSION

The ESMO-MCBS scores assume valid research methodologies and high-quality trial implementation, and freedom from publication bias. To promote the integrity of the ESMO-MCBS scoring, there is a need to discern valid and biased research. Consequently, new approaches are needed to indicate or penalise studies with deficiencies in their research methodologies, trial implementation, analysis or publication strategy that may contribute to biased outcomes and conclusions.

The necessary preconditions for a valid study are outlined in Table 1. The ESMO-MCBS already addresses some of these issues in version 1.1 and its upcoming revisions. The ESMO-MCBS only scores studies with a clinically relevant hypothesis and statistically significant findings consistent with a valid pre-specified statistical plan. When indirect surrogate outcomes are used, the scale incorporates additional precautions and caps to minimise the risk of exaggerated claims of benefit unless surrogacy is verified. Regarding the QoL data, the Working Group is collaborating with partners in the European Organization for Research and Treatment of Cancer (EORTC) to refine new strategies to restrict credits to findings based on robust methodology and adequately complete datasets.

Table 1. The necessary preconditions for a valid study

1. Clinically relevant and appropriate hypothesis (primary outcome, targeted magnitude of benefit, secondary outcomes, type I and II errors)
2. Appropriate study design
3. In comparative studies: an adequate control arm that is consistent with the contemporaneous standard of care at the time of trial initiation
4. Inclusion and exclusion criteria that optimise the balance between generalisability and participant safety
5. Completeness of data collection
6. Valid statistical plan and adherence to that plan
7. When overall survival is either a primary or secondary outcome, post-progression treatment demonstrably consistent with the contemporaneous standard of care
8. Analysis of data that clearly distinguishes between confirmatory findings and conjectural conclusions

This review has identified seven shortcomings in the ESMO-MCBS approach to potential sources of bias in clinical studies that will need to be addressed in the future development of the scale:

1. The ESMO-MCBS does not independently evaluate the control arm's validity, nor does it have a mechanism to identify to either indicate or penalise studies with a substandard control arm. This is relevant to all ESMO-MCBS forms evaluating comparative studies.
2. The ESMO-MCBS does not evaluate crossover, its appropriateness, and when appropriate, its adequacy. This is relevant to scores derived from OS data using form 2a.
3. The ESMO-MCBS does not have discriminatory rules to determine the pre-specified NI margin validity. This is relevant to form 2c.
4. The ESMO-MCBS does not indicate or penalise studies in which OS benefit may have been exaggerated by substandard post-progression treatment. This is relevant to scores derived from OS data using form 2a.
5. The ESMO-MCBS exception for *post hoc* subgroup findings based on enabling restratification based on plausible new genetic or other biomarkers is not explicitly restricted to biomarkers generating a modification in licensed indications. This is relevant to the instructions regarding the use of forms 1 and 2.
6. The ESMO-MCBS does not indicate or penalise trials with differential rates of informative censoring in studies graded based on DFS. This is relevant to form 1.
7. ESMO-MCBS does not address non-publication or delayed publication of QoL data. This is particularly relevant to form 2b.

These issues will be addressed in future iterations of the ESMO-MCBS. The ESMO-MCBS Working Group will consider all potential options and would appreciate stakeholder feedback in this process. Options include developing a checklist for evaluating these issues, using annotations to indicate flawed studies, or possibly applying a downgrade to ESMO-MCBS scores.

Nevertheless, the appropriate interpretation of the ESMO-MCBS scores requires the critical appraisal of trials to understand these issues in trial design, implementation, and data analysis that may have biased the results and conclusions. The ESMO-MCBS facilitates unbiased evaluation of the magnitude of clinical benefit from cancer medicines, however, like all tools, its utility lies in the hands of the user. The ESMO-MCBS does not obviate the need to think critically about cancer medicine trial designs, and users should consider all these issues when appraising and scoring any clinical trial.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the editorial and organisational support provided by Nicola Latino ESMO Head of Scientific Affairs and Martina Galotti ESMO Scientific Affairs Specialist.

FUNDING

None declared.

DISCLOSURE

TA personal fees and travel grants from Bristol-Myers Squibb (BMS), personal fees, grants and travel grants from Novartis, personal fees from Pierre Fabre, grants from Neracare, Sanofi, and SkylineDx, personal fees from CeCaVa outside the submitted work; JBa reports grants, personal fees, and nonfinancial support from Ipsen; personal fees and nonfinancial support from Pfizer, Novartis; nonfinancial support from AAA, Nanostring, Roche; grants and personal fees from Servier; and personal fees from Nutricia outside the submitted work; JBo is a statistician on the SAG-O (scientific advice committee oncology) for EMA, and scientific director of EORTC Headquarters. EORTC carries out clinical trials with many (most) pharma and some biotechs either with financial or material support, or with the company as the sponsor (intent to register new indication); he is co-responsible for the management of EORTC. AC has received honoraria/consulting fees from AstraZeneca, Boehringer-Ingelheim, Pfizer, Roche/Genentech, Eli Lilly and Company, Takeda, Novartis, Merck Sharp & Dohme (MSD), and BMS; GC scientific advisory board for BMS, Roche, Novartis, Lilly, Pfizer, Seagen, AstraZeneca, Daichii Sankyo, and Veracyte; UD Tumour Agnostic Evidence Generation Working Group Member, Roche; BG reports receiving consulting fees from Vivio Health. CG-R BMS (institutional research and travel grants, speaker's honoraria), Roche/Genentech (institutional research and travel grants, speaker's honoraria), Pierre Fabre (travel and educational grants, speaker's honoraria); MSD (travel grants); Eisai (speaker's honoraria); Foundation Medicine (research grant, speaker's honoraria); BK Honoraria for lectures from Ipsen, Novartis and MSD; SO research grants from Celldex and Novartis to the institution; AP received honoraria for consulting, advisory role or lectures from AstraZeneca, Agilent/Dako, Boehringer Ingelheim, BMS, Eli Lilly, Janssen, MSD, Pfizer and Roche Genentech; GP received institutional financial support for advisory board/consultancy from Roche, Amgen, Merck, MSD, BMS, and institutional support for clinical trials or contracted research from Amgen, Roche, AstraZeneca, Pfizer, Merck, BMS, MSD, Novartis, Lilly; MP consulting or advisory role: AstraZeneca, Lilly, MSD, Novartis, Pfizer, Roche, Genentech, Crescendo Biologics, Periphagen, Huya Bioscience, Debiopharm Group, Odante Therapeutics; Consulting or Advisory Role: G1 Therapeutics, Menarini, Seattle Genetics, Camel-IDS, Immunomedics, Oncolytics, Radius Health; Research Funding: AstraZeneca (Inst), Lilly (Inst), MSD (Inst), Novartis (Inst), Pfizer (Inst), Roche (Inst), Genentech (Inst), Radius Health (Inst), Synthon (Inst), Servier (Inst); Other Relationship: Radius Health; JT personal financial interest in the form of scientific consultancy role for Array Biopharma, AstraZeneca, Bayer, Boehringer Ingelheim, Chugai, Daiichi Sankyo, F. Hoffmann-La Roche Ltd, Genentech, Inc., HalioDX SAS, Ikena Oncology, IQVIA, Imedex, Lilly, MSD, Menarini, Merck Serono, Mirati, Novartis, Peptomyc, Pfizer, Pierre Fabre, Samsung Bioepis, Sanofi, Seattle Genetics, Servier, Taiho, Tessa Therapeutics and TheraMyc. Institutional financial interest in the form of financial support for clinical trials or contracted research for Amgen Inc., Array Biopharma Inc., AstraZeneca Pharmaceuticals LP, Debiopharm International SA, F. Hoffmann-La Roche Ltd, Genentech Inc., Janssen-Cilag SA, MSD, Novartis Farmacéutica SA, Taiho Pharma USA Inc., Pharma Mar, Spanish Association Against Cancer Scientific Foundation and Cancer Research UK; EGEdV declares: institutional financial support for advisory board/consultancy from Sanofi, Daiichi, Sankyo, NSABP, Pfizer and Merck, and institutional support for clinical trials or contracted research from Amgen, Genentech, Roche, AstraZeneca, Synthon, Nordic Nanovector, G1 Therapeutics, Bayer, Chugai Pharma, CytomX Therapeutics, Servier and Radius Health; CZ consultancies and speaker's honoraria: Roche, Novartis, BMS, MSD, Imugene, Ariad, Pfizer, Merrimack, Merck KGaA, Fibrogen, AstraZeneca, Tesaro, Gilead, Servier, Shire, Eli Lilly, Athenex. Institution (Central European Cooperative Oncology Group): BMS, MSD, Pfizer, AstraZeneca, Servier, Eli Lilly; GZ received speaker's honoraria from Amgen and Ipsen. All other authors have declared no conflicts of interest.

REFERENCES

1. Cherny NI, Dafni U, Bogaerts J, et al. ESMO-Magnitude of Clinical Benefit Scale version 1.1. *Ann Oncol* 2017;28:2340-2366.
2. 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;26:1547-1573.
3. Daniels N. Accountability for reasonableness. *Br Med J* 2000;321:1300-1301.
4. European Society for Medical Oncology. The ESMO-MCBS Scorecards. Available at: https://www.esmo.org/guidelines/esmo-mcbs/esmo-mcbs-scorecards.
5. World Medical Association. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. In Edition 64th WMA General Assembly, Fortaleza, Brazil, October 2013. 2013. Available at: https://www.wma.net/policies-post/

wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/.
6. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH Topic E 10: Choice of control group and related issues in clinical trials. In Edition 2000. 35. Available at: https://database.ich.org/sites/default/files/E10_Guideline.pdf.
7. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH topic E9(R1): Addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. In Edition 2019. 22. 2019. Available at: https://database.ich.org/sites/default/files/E29-R21_Step24_Guideline_2019_1203.pdf.
8. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH Topic E 9: Statistical Principles for Clinical Trials. In Edition 1998. 39. 1998. Available at: https://database.ich.org/sites/default/files/E39_Guideline.pdf.
9. Food and Drug Administration. Multiple Endpoints in Clinical Trials Guidance for Industry. Draft Guidance. In Edition 2017. 51. Available at: https://www.fda.gov/media/102657/download.
10. Food and Drug Administration. Clinical trial endpoints for the approval of cancer drugs and biologics guidance for industry. In Edition 2019. 19. Available at: https://www.fda.gov/media/71195/download.
11. Food and Drug Administration. Non-Inferiority Clinical Trials to Establish Effectiveness Guidance for Industry. In Edition 2016. 56. Available at: https://www.fda.gov/media/78504/download.
12. European Medicines Agency. Guideline on the evaluation of anticancer medicinal products in man (EMA/CHMP/205/95 Rev.5). In Edition London: European Medicines Agency. 43. 2017. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-evaluation-anticancer-medicinal-products-man-revision-5_en.pdf.
13. European Medicines Agency. Guideline on the investigation of subgroups in confirmatory clinical trials. In Edition London: European Medicines Agency. 20. 2019. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf.
14. European Medicines Agency. Points to consider on switching between superiority and non-inferiority. In Edition London: European Medicines Agency. 11. 2000. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-switching-between-superiority-non-inferiority_en.pdf.
15. European network for Health Technology Assessment. Guideline: Endpoints used for Relative Effectiveness Assessment: Clinical Endpoints. In Edition. 20. 2015. Available at: https://eunethta.eu/wp-content/uploads/2018/2002/WP2017-SG2013-GL-clin_endpoints_amend2015.pdf.
16. European Network for Health Technology Assessment. Comparators & Comparisons: Criteria for the choice of the most appropriate comparator(s). In Edition. 23. 2015. Available at: https://eunethta.eu/wp-content/uploads/2018/03/Criteria_WP7-SG3-GL-choice_of_comparator_amend2015.pdf.
17. European Network for Health Technology Assessment. Internal validity of non-randomised studies (NRS) on interventions. In Edition. 33. 2015. Available at: https://eunethta.eu/wp-content/uploads/2018/2003/Criteria_WP2017-SG2013-GL-choice_of_comparator_amend2015.pdf.
18. European Network for Health Technology Assessment. Guideline: Endpoints used for relative effectiveness assessment of pharmaceuticals: Clinical endpoints v2 In Edition. 20. 2015. Available at: https://eunethta.eu/wp-content/uploads/2018/02/WP7-SG3-GL-clin_endpoints_amend2015.pdf.
19. European Network for Health Technology Assessment. Guideline: Endpoints used in relative effectiveness assessment of pharmaceuticals: Surrogate Endpoints v2. 23. 2015. Available at: https://eunethta.eu/wp-content/uploads/2018/03/surrogate_endpoints.pdf.
20. Dummer R, Schadendorf D, Ascierto PA, et al. Binimetinib versus dacarbazine in patients with advanced NRAS-mutant melanoma (NEMO): a multicentre, open-label, randomised, phase 3 trial. *Lancet Oncol* 2017;18:435-445.
21. Robert C, Thomas L, Bondarenko I, et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. *N Engl J Med* 2011;364:2517-2526.
22. European Medicines Agency. Yervoy (ipilimumab): An overview of Yervoy and why it is authorised in the EU. In Edition London: European Medicines Agency. 4. 2011. Available at: https://www.ema.europa.eu/en/documents/overview/yervoy-epar-medicine-overview_en.pdf.
23. Food and Drug Administration. Prescribing information: Yervoy (ipilimumab) injection, for intravenous use 2011. 71. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/label/2020/125377s110lbl.pdf.
24. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001;69:89-95.
25. Savina M, Gourgou S, Italiano A, et al. Meta-analyses evaluating surrogate endpoints for overall survival in cancer randomized trials: a critical review. *Crit Rev Oncol Hematol* 2018;123:21-41.
26. Kim C, Prasad V. Strength of validation for surrogate end points used in the US food and drug administration's approval of oncology drugs. *Mayo Clin Proc* 2016;91:713-725.
27. Ciani O, Buyse M, Garside R, et al. Meta-analyses of randomized controlled trials show suboptimal validity of surrogate outcomes for overall survival in advanced colorectal cancer. *J Clin Epidemiol* 2015;68:833-842.
28. Ciani O, Davis S, Tappenden P, et al. Validation of surrogate endpoints in advanced solid tumors: systematic review of statistical methods, results, and implications for policy makers. *Int J Technol Assess Health Care* 2014;30:312-324.
29. Shi Q, Sargent DJ. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. *Int J Clin Oncol* 2009;14:102-111.
30. Gyawali B, Hey SP, Kesselheim AS. Evaluating the evidence behind the surrogate measures included in the FDA's table of surrogate endpoints as supporting approval of cancer drugs. *EClinicalMedicine* 2020;21:100332.
31. Haslam A, Hey SP, Gill J, Prasad V. A systematic review of trial-level meta-analyses measuring the strength of association between surrogate end-points and overall survival in oncology. *Eur J Cancer* 2019;106:196-211.
32. Saad ED, Katz A, Hoff PM, Buyse M. Progression-free survival as surrogate and as true end point: insights from the breast and colorectal cancer literature. *Ann Oncol* 2010;21:7-12.
33. Booth CM, Eisenhauer EA. Progression-free survival: meaningful or simply measurable? *J Clin Oncol* 2012;30:1030-1033.
34. Saad ED, Katz A, Buyse M. Overall survival and post-progression survival in advanced breast cancer: a review of recent randomized clinical trials. *J Clin Oncol* 2010;28:1958-1962.
35. Wilkerson J, Fojo T. Progression-free survival is simply a measure of a drug's effect while administered and is not a surrogate for overall survival. *Cancer J* 2009;15:379-385.
36. Amir E, Seruga B, Kwong R, et al. Poor correlation between progression-free and overall survival in modern clinical trials: are composite endpoints the answer? *Eur J Cancer* 2012;48:385-388.
37. Miller K, Wang M, Gralow J, et al. Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. *N Engl J Med* 2007;357:2666-2676.
38. Paluch-Shimon S, Cherny NI, de Vries EGE, et al. Application of the ESMO-Magnitude of clinical benefit scale (V.1.1) to the field of early breast cancer therapies. *ESMO Open* 2020;5:e000743.
39. Robinson AG, Booth CM, Eisenhauer EA. Disease-free survival as an end-point in the treatment of solid tumours—perspectives from clinical trials and clinical practice. *Eur J Cancer* 2014;50:2298-2302.
40. O'Sullivan CC, Bradbury I, Campbell C, et al. Efficacy of adjuvant trastuzumab for patients with human epidermal growth factor receptor 2-positive early breast cancer and rumors $\leq$ 2 cm: a meta-analysis of the randomized trastuzumab trials. *J Clin Oncol* 2015;33:2600-2608.
41. Haslam A, Prasad V. When is crossover desirable in cancer drug trials and when is it problematic? *Ann Oncol* 2018;29:1079-1081.
42. Prasad V, Grady C. The misguided ethics of crossover trials. *Contemp Clin Trials* 2014;37:167-169.

43. Hilal T, Gonzalez-Velez M, Prasad V. Limitations in clinical trials leading to anticancer drug approvals by the US food and drug administration. *JAMA Intern Med* 2020;180:1108-1115.
44. Ryan CJ, Smith MR, de Bono JS, et al. Abiraterone in metastatic prostate cancer without previous chemotherapy. *N Engl J Med* 2012;368:138-148.
45. Ryan CJ, Smith MR, Fizazi K, et al. Abiraterone acetate plus prednisone versus placebo plus prednisone in chemotherapy-naive men with metastatic castration-resistant prostate cancer (COU-AA-302): final overall survival analysis of a randomised, double-blind, placebo-controlled phase 3 study. *Lancet Oncol* 2015;16:152-160.
46. Fizazi K, Tran N, Fein L, et al. Abiraterone plus prednisone in metastatic, castration-sensitive prostate cancer. *N Engl J Med* 2017;377:352-360.
47. Kantoff PW, Higano CS, Shore ND, et al. Sipuleucel-T immunotherapy for castration-resistant prostate cancer. *N Engl J Med* 2010;363:411-422.
48. Tannock IF, de Wit R, Berry WR, et al. Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. *N Engl J Med* 2004;351:1502-1512.
49. Agency for Healthcare Research and Quality. Technology assessment: Outcomes of Sipuleucel —T therapy. In Edition. 60. 2011. Available at: https://www.cms.gov/Medicare/Coverage/DeterminationProcess/downloads/id77TA.pdf.
50. Zhang JJ, Blumenthal GM, He K, et al. Overestimation of the effect size in group sequential trials. *Clin Cancer Res* 2012;18:4872-4876.
51. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009;45:228-247.
52. Seymour L, Bogaerts J, Perrone A, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. *Lancet Oncol* 2017;18:e143-e152.
53. Gyawali B, Tessema FA, Jung EH, Kesselheim AS. Assessing the justification, funding, success, and survival outcomes of randomized non-inferiority trials of cancer drugs: a systematic review and pooled analysis. *JAMA Network Open* 2019;2:e199570.
54. Zia M, Siu L, Pond G. Comparison of outcomes of phase II studies and subsequent randomized control studies using identical chemotherapeutic regimens. *J Clin Oncol* 2005;23:6982-6991.
55. D'Agostino Sr RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Stat Med* 2003;22:169-186.
56. Piccart MJ, Hilbers FS, Bliss JM, et al. Road map to safe and well-designed de-escalation trials of systemic adjuvant therapy for solid tumors. *J Clin Oncol* 2020;38:4120-4129.
57. Tanaka S, Kinjo Y, Kataoka Y, et al. Statistical issues and recommendations for noninferiority trials in oncology: a systematic review. *Clin Cancer Res* 2012;18:1837-1847.
58. Raphael J, Verma S. Overall survival (OS) endpoint: an incomplete evaluation of metastatic breast cancer (MBC) treatment outcome. *Breast Cancer Res Treat* 2015;150:473-478.
59. Seidman AD, Maues J, Tomlin T, et al. The evolution of clinical trials in metastatic breast cancer: design features and endpoints that matter. *Am Soc Clin Oncol Educ Book* 2020;40:44-54.
60. Prasad V, Berger VW. Hard-wired bias: how even double-blind, randomized controlled trials can be skewed from the start. *Mayo Clin Proc* 2015;90:1171-1175.
61. Im S-A, Lu Y-S, Bardia A, et al. Overall survival with ribociclib plus endocrine therapy in breast cancer. *N Engl J Med* 2019;381:307-316.
62. Harbeck N, Franke F, Villanueva-Vazquez R, et al. Health-related quality of life in premenopausal women with hormone-receptor-positive, HER2-negative advanced breast cancer treated with ribociclib plus endocrine therapy: results from a phase III randomized clinical trial (MONALEESA-7). *Ther Adv Med Oncol* 2020;12:1-8.
63. Easterbrook PJ, Gopalan R, Berlin J, Matthews DR. Publication bias in clinical research. *Lancet* 1991;337:867-872.
64. Hwang TJ, Gyawali B. Association between progression-free survival and patients' quality of life in cancer clinical trials. *Int J Cancer* 2019;144:1746-1751.
65. Marandino L, La Salvia A, Sonetto C, et al. Deficiencies in health-related quality-of-life assessment and reporting: a systematic review of oncology randomized phase III trials published between 2012 and 2016. *Ann Oncol* 2018;29:2288-2295.
66. von Minckwitz G, Procter M, de Azambuja E, et al. Adjuvant pertuzumab and trastuzumab in early HER2-positive breast cancer. *N Engl J Med* 2017;377:122-131.
67. Schmid P, Adams S, Rugo HS, et al. Atezolizumab and nab-paclitaxel in advanced triple-negative breast cancer. *N Engl J Med* 2018;379:2108-2121.
68. Fukuoka M, Wu Y-L, Thongprasert S, et al. Biomarker analyses and final overall survival results from a phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non—small-cell lung cancer in Asia (IPASS). *J Clin Oncol* 2011;29:2866-2874.
69. Douillard J, Siena S, Cassidy J, et al. Final results from PRIME: randomized phase 3 study of panitumumab with FOLFOX4 for first—line treatment of metastatic colorectal cancer. *Ann Oncol* 2014;25:1346-1355.
70. Douillard J-Y, Oliner KS, Siena S, et al. Panitumumab—FOLFOX4 treatment and RAS mutations in colorectal cancer. *N Engl J Med* 2013;369:1023-1034.
71. Van Cutsem E, Lenz HJ, Kohne CH, et al. Fluorouracil, leucovorin, and irinotecan plus cetuximab treatment and RAS mutations in colorectal cancer. *J Clin Oncol* 2015;33:692-700.
72. Shih WJ. Problems in dealing with missing data and informative censoring in clinical trials. *Curr Control Trials Cardiovasc Med* 2002;3:1-7.
73. Carroll KJ. Analysis of progression-free survival in oncology trials: some common statistical issues. *Pharm Stat* 2007;6:99-113.
74. Fleming TR, Rothmann MD, Lu HL. Issues in using progression-free survival when evaluating oncology products. *J Clin Oncol* 2009;27:2874-2880.
75. Baselga J, Campone M, Piccart M, et al. Everolimus in postmenopausal hormone-receptor—positive advanced breast cancer. *N Engl J Med* 2012;366:520-529.
76. Templeton AJ, Ace O, Amir E, et al. Influence of censoring on conclusions of trials for women with metastatic breast cancer. *Eur J Cancer* 2015;51:721-724.
77. Piccart M, Hortobagyi GN, Campone M, et al. Everolimus plus exemestane for hormone-receptor-positive, human epidermal growth factor receptor-2-negative advanced breast cancer: overall survival results from BOLERO-2. *Ann Oncol* 2014;25:2357-2362.