Are Surrogate Endpoints Unbiased Metrics in Clinical Benefit Scores of the ASCO Value Framework?

Sierra Cheng, BSc^a; Matthew C. Cheung, MD, SM^{a,b,*}; Di Maria Jiang, MD, FRCPC^b; Erica McDonald, BSc^a; Vanessa S. Arciero, BSc^a; Doreen Anuli Ezeife, MD, MSc^b; Amanda Rahmadian, BSc^a; Alexandra Chambers, MA^c; Kelley-Anne Sabarre, MSc^c; Ambika Parmar, MD^a; and Kelvin K.W. Chan, MD, MSc, PhD^{a,b,d,*}

ABSTRACT

Background: Clinical benefit scores (CBS) are key elements of the ASCO Value Framework (ASCO-VF) and are weighted based on a hierarchy of efficacy endpoints: hazard ratio for death (HR OS), median overall survival (mOS), HR for disease progression (HR PFS), median progression-free survival (mPFS), and response rate (RR). When HR OS is unavailable, the other endpoints serve as “surrogates” to calculate CBS. CBS are computed from PFS or RR in 39.6% of randomized controlled trials. This study examined whether surrogate-derived CBS offer unbiased scoring compared with HR OS-derived CBS. **Methods:** Using the ASCO-VF, CBS for advanced disease settings were computed for randomized controlled trials of oncology drug approvals by the FDA, European Medicines Agency, and Health Canada in January 2006 through December 2017. Mean differences of surrogate-derived CBS minus HR OS-derived CBS assessed the tendency of surrogate-derived CBS to overestimate or underestimate clinical benefit. Spearman’s correlation evaluated the association between surrogate- and HR OS-derived CBS. Mean absolute error assessed the average difference between surrogate-derived CBS relative to HR OS-derived CBS. **Results:** CBS derived from mOS, HR PFS, mPFS, and RR overestimated HR OS-derived CBS in 58%, 68%, 77%, and 55% of pairs and overall by an average of 5.62 (n=90), 6.86 (n=110), 29.81 (n=101), and 3.58 (n=108), respectively. Correlation coefficients were 0.80 (95% CI, 0.70–0.86), 0.38 (0.20–0.53), 0.20 (0.00–0.38), and 0.01 (–0.18 to 0.19) for mOS-, HR PFS-, mPFS-, and RR-derived CBS, respectively, and mean absolute errors were 11.32, 12.34, 40.40, and 18.63, respectively. **Conclusions:** Based on the ASCO-VF algorithm, HR PFS-, mPFS-, and RR-derived CBS are suboptimal surrogates, because they were shown to be biased and poorly correlated to HR OS-derived CBS. Despite lower weighting than OS in the ASCO-VF algorithm, PFS still overestimated CBS. Simple rescaling of surrogate endpoints may improve their validity within the ASCO-VF given their poor correlations with HR OS-derived CBS.

J Natl Compr Canc Netw 2019;17(12):1489–1496
doi: 10.6004/jnccn.2019.7333

Background

In 2015, ASCO published an evaluative framework to assess the value of cancer treatments via a single summary score.^{1,2} In response to the increasing disparity between cost and clinical benefit associated with novel oncology treatments, the ASCO Value Framework (ASCO-VF) was developed to distinguish between therapies delivering modest to substantial clinical benefit.^{3–5} Intended for clinical settings, the framework aims to inform patient–physician discussion surrounding the relative efficacy of treatment options in relation to their cost.^{1,2}

The ASCO-VF, updated in 2016,¹ computes a net health benefit score for randomized controlled trials (RCTs) based on 3 key elements: clinical benefit, toxicity, and symptom palliation.^{1,2}

Clinical benefit scores (CBS), derived before adjustments for toxicity, symptom palliation, improvements in quality of life, or prolonged survival, are critical components of final net health benefit scores.^{1,2} CBS derived from the hazard ratio for death (HR OS) are considered the reference standard. In the absence of HR OS, other efficacy endpoints are used to compute CBS in the following order: median overall survival (mOS), hazard ratio for disease progression (HR PFS), median progression-free survival (mPFS), and overall response rate (RR). Based on this hierarchy, CBS are calculated using scaling factors reflective of an endpoint’s relative importance when evaluating the efficacy of treatment options.^{1,2} Trials reporting OS-based endpoints are weighted the greatest $[(1 - \text{HR}) \text{ or } [\text{percentage difference in median OS}] \times 100]$, followed by those reporting only PFS-based endpoints

^aOdette Cancer Centre, Sunnybrook Health Sciences Centre, and ^bDepartment of Medicine, University of Toronto; ^cCanadian Agency for Drugs and Technologies in Health; and ^dCanadian Centre for Applied Research in Cancer Control, Toronto, Ontario, Canada.

*Contributed equally.

See JNCCN.org for supplemental online content.

See page 1569 for related commentary.

$([1 - \text{HR}] \text{ or } [\text{percentage difference in median PFS}] \times 80)$, and finally those reporting only RR-based endpoints $([\text{complete} + \text{partial RR}] \times 70)$.¹ As such, when scoring with HR OS or HR PFS, the greatest possible CBS are 100 and 80, respectively (as the HR approaches 0). The minimum possible CBS are undefined, extending into negative values for an HR > 1.0. Scores derived from mOS or mPFS can range from negative scores, when the percentage difference of the medians between the experimental arm and the control arm of an RCT is < 0, to scores much greater than 100 when the percentage difference is > 100. Finally, RR-derived CBS can range from 0 to 70 for an overall RR ranging from 0% to 100%. Scoring only deviates from the hierarchy when survival data are obscured by crossover, preferentially scoring using PFS rather than OS outcomes.¹ At the time of regulatory approval, 39.6% of RCTs either lacked OS reporting or reported crossover, and therefore a substantial proportion of RCTs require scoring with surrogate endpoints (S. Cheng, BSc, personal communication, 2018).

Although the ASCO-VF for advanced disease acknowledges the surrogacy of endpoints by placing varying weights on endpoints according to the proposed hierarchy of significance, it is unclear whether this intrinsic rescaling is empirically supported in the algorithm such that surrogate-derived CBS accurately reflect the clinical benefit in place of HR OS in the framework, as per framework intentions. We aimed to empirically examine the CBS component of the net health benefit to determine whether, based on the current scaling factors in the ASCO-VF, surrogate endpoints—defined as endpoints used for scoring when the hierarchy-defined preferred outcome is not available—provide unbiased scoring of clinical benefit compared with the framework's "preferred endpoint" of scoring based on HR OS.

Methods

Inclusion of RCTs

We identified RCTs cited as clinical efficacy evidence in oncology drug regulatory approvals of new indications and molecules by the FDA, European Medicines Agency (EMA), and Health Canada in January 2006 through December 2017. Included RCTs were restricted to double- or triple-arm phase III trials in advanced/noncurative settings. RCTs were eligible for scoring if any of the following clinical endpoints were reported: OS, PFS, time-to-progression (TTP), or overall RR. RCTs reporting crossover at the time of analysis were excluded because of the potential for the survival endpoint to be obscured. Based on the retrieved primary publications of RCTs, a cited reference search was conducted on Web of Science for follow-up publications of efficacy data on the intention-to-treat population. If crossover was reported in the follow-up, then data from the primary publication were used.

Scoring of RCTs

Two independent reviewers computed ASCO-VF CBS for each RCT based on all reported endpoints. When both independent and investigator assessments were reported, independent assessment data were extracted. Discordances in scores between reviewers were resolved through consensus.

Summary of CBS

Descriptive statistics including means, standard deviations, medians, and interquartile ranges (IQRs) were used to summarize CBS derived from all reported endpoints.

Difference in Clinical Benefit Derived From Surrogate Outcomes Compared With Reference Standard Outcomes

The absolute differences in CBS between surrogate and reference standard outcomes were calculated (surrogate-CBS minus HR OS-CBS). An average of the differences between surrogate-derived CBS and HR OS-derived CBS was calculated to evaluate the tendency for surrogate-derived CBS to overestimate or underestimate clinical benefit in the ASCO-VF. Although there is no established standard for a clinically meaningful difference in CBS, we previously determined the distribution of ASCO-VF CBS scores to have a median of 23 (IQR, 14–33) (S. Cheng, BSc, personal communication, 2018).⁶ Thus, proportions of overestimation and underestimation by > 20 points (approximately the range of the IQR) were calculated to represent substantial differences. MAE, an average of absolute differences and a common statistical metric, was used to assess the potential degree of imprecision or "prediction error" determining the average absolute difference between surrogate-derived and HR OS-derived CBS. A lower imprecision (MAE) value indicated better surrogacy of the endpoint for HR OS-derived CBS within the ASCO-VF.⁷

Bland-Altman plots, commonly used in method comparison studies,^{8,9} were constructed in this analysis to evaluate the agreement between HR OS-derived CBS (the reference standard) and surrogate-derived CBS. The Bland-Altman index (%) was calculated as a percentage of surrogate-derived CBS that fell beyond the limits of agreement (LOA). The surrogate showing the narrowest LOA and the lowest Bland-Altman index represented the surrogate with the best agreement with HR OS-derived CBS.

Rank Correlation Between Reference Standard and Surrogate-Derived CBS

For RCTs reporting HR OS and at least one other endpoint (mOS, HR PFS, mPFS, or RR), Spearman's rank correlations measured the correlation between surrogate (mOS, HR PFS, mPFS, or RR)-derived CBS compared with reference standard (HR OS)-derived CBS. Spearman's rank correlations are computed on the ranks of data

values rather than the data values themselves, therefore assessing whether the ranking for the HR OS benefits among drugs is preserved when using surrogates instead.

Sensitivity analyses were completed, restricted to RCTs reporting all endpoints, RCTs examining first-line therapy, RCTs with OS as the primary endpoint, RCTs without OS as the primary endpoint, RCTs with PFS as the primary endpoint, RCTs on hematologic malignancies only, and RCTs on nonhematologic malignancies. Sensitivity analyses restricted to specific cancer types—lung, breast, colorectal, and melanoma—were performed, as were analyses restricted by class of therapy (targeted agent or chemotherapy). Further sensitivity analyses examining RCTs with a PFS primary endpoint were completed using HR PFS as the reference standard outcome against which to compare instead of HR OS.

All analyses were performed using RStudio version 1.1.453 (RStudio, Inc).

Results

Included RCTs

Excluding duplicate entries, 401 RCTs were identified from the FDA, EMA, and Health Canada drug approval databases^{10–12} (Figure 1), with 127 RCTs eligible for scoring. Of these 127 RCTs, 108 RCTs comprising 116 treatment comparisons reported HR OS and were therefore included in the primary analysis. Key characteristics of the included RCTs are described in Table 1.

Summary of ASCO-VF CBS

Evaluation of the 108 RCTs via the ASCO-VF allowed for CBS calculation with 116 HR OS. These HR OS–derived CBS were paired with 90 mOS–derived, 110 HR PFS–derived, 101 mPFS–derived, and 108 RR–derived CBS for analysis (see supplemental eFigure 1 and eTable 1, available with this article at JNCCN.org).

Clinical Benefit Derived From Surrogate Outcomes Versus Reference Standard

Compared with HR OS–derived CBS, CBS derived from mOS, HR PFS, mPFS, and RR overestimated clinical benefit by an average of 5.62, 6.86, 29.81, and 3.58 points, respectively (Table 2). mOS–, HR PFS–, mPFS–, and RR–based surrogates overestimated clinical benefit in 58%, 68%, 77%, and 55% of scores, respectively, compared with HR OS–based surrogates. Proportions of overestimation by >20 points were 10% (mOS), 15% (HR PFS), 48% (mPFS), and 26% (RR). Proportions of underestimation by more than 20 points were 2% (mOS), 4% (HR PFS), 9% (mPFS), and 15% (RR) (Figure 2).

Calculation of the degree of precision or prediction error revealed the greatest average difference with mPFS compared with HR OS (MAE, 40.40), followed by RR

Figure 1. Process of screening and scoring randomized controlled trials.

Abbreviations: CBS, clinical benefit scores; EMA, European Medicines Agency; HR, hazard ratio; OS, overall survival; PFS, progression-free survival; RCTs, randomized controlled trials; RR, response rate; TTP, time-to-progression.

*If crossover was reported to have occurred at time of analysis and confounded HR OS data, then the RCT was excluded.

(MAE, 18.63), HR PFS (MAE, 12.34), and mOS (MAE, 11.32) (Table 2).

In evaluating Bland-Altman plots, all surrogates showed poor agreement (very wide LOAs, all much greater than 20 points), with mPFS showing the worst agreement. For mOS–derived CBS, the LOA ranged from −39.72 to 50.96 with a Bland-Altman index of 3.3% (Figure 3A). For HR PFS–derived CBS, the LOA ranged from −23.10 to 36.81 with a Bland-Altman index of 3.6% (Figure 3B). For mPFS–derived CBS, the LOA ranged

Table 1. Characteristics of RCTs (N=108)

	Treatment Comparisons n
Drug comparisons	116
Indication	
Myeloma	12
Breast cancer	9
Lung cancer	13
Colorectal cancer	12
Melanoma	8
Renal cell carcinoma	7
Prostate cancer	7
Other	40
Line of treatment	
1st	38
≥2nd	70
Reported efficacy data	
HR OS	116
mOS	90
HR PFS	110
mPFS	101
RR	108
Primary endpoint ^a	
OS	59
PFS	48
RR	2
Other	9

Abbreviations: HR OS, hazard ratio for death; HR PFS, hazard ratio for progression-free survival; mOS, median overall survival; mPFS, median progression-free survival; RCT, randomized clinical trial; RR, response rate.
^a10 RCTs had >1 primary endpoint.

from −66.31 to 125.93 with a Bland-Altman index of 5.0% (Figure 3C). For RR-derived CBS, the LOA ranged from −40.23 to 47.38 with a Bland-Altman index of 3.7% (Figure 3D).

Rank Correlation Between Reference Standard and Surrogate CBS

A strong Spearman's rank correlation of 0.80 (95% CI, 0.70–0.86; $P<.001$) was observed between HR OS–derived CBS and mOS-derived CBS. The CBS derived from HR PFS, mPFS, and RR were weakly correlated with the reference standard, at 0.38 (0.20–0.53; $P<.001$), 0.20 (0.00–0.38; $P=.05$), and 0.01 (−0.18 to 0.19; $P=.95$), respectively (supplemental eFigure 2).

Sensitivity analyses presented similar results (supplemental eTable 2). From the 127 originally identified RCTs that were eligible for scoring, correlations for RCTs with a PFS primary endpoint compared with HR PFS as the reference standard were poor for HR OS (0.30), mOS (0.38), and RR (−0.07). RR correlation was especially low for all analyses except for breast cancer RCTs (0.80).

Discussion

Our study shows that trial outcomes that are considered traditional surrogates for OS, including RR and PFS, do not allow an accurate calculation of clinical benefit as defined by the suggested intrinsic rescaling of CBS in the ASCO-VF algorithm. Furthermore, despite rescaling by the ASCO-VF, these surrogates still exhibited bias with a tendency to overestimate clinical benefit in comparison with scores calculated with HR OS.

The discovered poor correlations may indicate that different constructs of clinical benefit are being measured. Given the poor correlations between CBS derived from surrogate endpoints compared with those derived from HR OS and the variation in overestimation and underestimation in HR OS–derived CBS versus surrogate-derived CBS, simple rescaling of the surrogate endpoints (changing the weights given to different surrogate endpoints), which the ASCO-VF presently involves, will not serve to improve their validity (supplemental eFigure 2).

The inherent inadequate correlation between endpoints such as PFS and OS renders them difficult to translate into a framework. The ESMO Magnitude of

Table 2. Summary Statistics of Difference Between Surrogate-Derived and HR OS–Derived CBS

	mOS	HR PFS	mPFS	RR
Total ^a	90	110	101	108
Minimum	−22.90	−23.00	−64.19	−52.49
Median (IQR)	1.49 (−4.69 to 7.35)	5.10 (−2.75 to 16.50)	18.55 (1.16 to 50.00)	2.65 (−13.65 to 20.19)
Average	5.62	6.86	29.81	3.58
Maximum	169.80	65.00	193.30	55.80
SD	23.13	15.28	49.04	22.35
MAE	11.32	12.34	40.40	18.63

Abbreviations: CBS, clinical benefit scores; HR OS, hazard ratio for death; HR PFS, hazard ratio for progression-free survival; IQR, interquartile range; MAE, mean absolute error; mOS, median overall survival; mPFS, median progression-free survival; RR, response rate.

^aTotal number of pairs with HR OS–derived CBS.

Figure 2. Differences in surrogate- and HR OS–derived CBS for (A) mOS, (B) HR PFS, (C) mPFS, and (D) RR. Each bar represents the difference in CBS (surrogate-derived minus HR OS–derived) for a single randomized clinical trial. The red dashed lines represent the 20-point difference thresholds. Each bar exceeding this threshold (over or under) represents a HR OS–surrogate pair with substantial differences.

Abbreviations: CBS, clinical benefit scores; HR OS, hazard ratio for death; HR PFS, hazard ratio for disease progression; mOS, median overall survival; mPFS, median progression-free survival; RR, response rate.

Clinical Benefit Scale (ESMO-MCBS) possibly circumvents the concern of using OS as the preferred endpoint by scoring based on primary endpoints rather than a fixed hierarchy of endpoints.^{13,14} However, this may not be seen as meaningful.¹⁵ Ultimately, the selected preferred endpoint for a particular decision may be purpose-specific. Our study highlights concerns about future objectives to distinguish and prioritize treatments based on isolated surrogate measure outcomes as they stand today. A true comparison of treatments based on their clinical benefits may not be meaningful if such clinical benefits are derived from different metrics. As the framework is further developed, stringencies placed on different clinical

outcomes should be reassessed and efforts should be made to incorporate related outcome measures into unified scores. Still, it is difficult to determine the true value of a treatment via surrogate endpoints when current evidence suggests that they are not always reliable substitutes for OS or quality of life.

In keeping with the ASCO-VF recommendation for mOS as the preferred surrogate endpoint, the current results revealed that mOS exhibited the least bias during the estimation of clinical benefit. However, the utility of this surrogate is questioned given the limited instances of clinical trials reporting mOS without an HR OS, a situation only encountered with one trial in our study.¹⁶

Figure 3. Bland-Altman plots of surrogate- vs HR OS–derived CBS for (A) mOS, (B) HR PFS, (C) mPFS, and (D) RR.

Abbreviations: CBS, clinical benefit scores; HR OS, hazard ratio for death; HR PFS, hazard ratio for disease progression; mOS, median overall survival; mPFS, median progression-free survival; RR, response rate.

However, there are instances when mOS in place of HR OS is required, such as when the proportional hazard assumption of an RCT is violated, as is often observed in immunotherapy trials.^{17,18} In such cases, mOS would be the preferred endpoint after HR OS following the ASCO-VF hierarchy; however, medians fail to capture the long-term survival benefits of immunotherapy trials. Accordingly, the use of mOS alone likely leads to an inaccurate estimation of clinical benefit. Furthermore, considering a favorable HR OS does not necessarily translate to large absolute gains in survival, combining mOS and HR OS to achieve a single OS CBS may better represent a meaningful clinical benefit, similar to the approach by the ESMO-MCBS.^{13,14}

Generally, OS is considered a preferred endpoint to assess the efficacy of cancer drugs, because it is an objective, easily interpretable, and clinically meaningful measure. Measuring OS, however, necessitates a large sample size and extended follow-up, and it can be confounded by crossover to alternative therapies after progression.^{19–21} Thus, we excluded such RCTs reporting

crossover. Compared with other endpoints such as PFS, the statistical disadvantages of OS include a smaller number of events and therefore lower power, a longer median time to an event, and a smaller relative effect (HR OS tends to be closer to 1.00 than HR PFS).¹⁵

PFS outcomes are commonly reported as surrogates for survival, with HR PFS most often used for deriving CBS. HR PFS–derived CBS showed a considerable difference from HR OS–derived CBS and had the greatest tendency to overestimate after mPFS despite the framework’s rescaling. mPFS not only had the greatest tendency to overestimate but also overestimated by the greatest magnitude, on average, with almost half of the mPFS-derived CBS overestimating by >20 points. Although a minimum clinically meaningful difference value in CBS has yet to be defined, based on the distribution of CBS, this number may represent a very large proportion of mPFS-derived CBS substantially differing from HR OS–derived CBS. The LOA was by far the widest for mPFS-derived CBS, which also had the greatest Bland-Altman index. Notably, however, between 2008

and 2013, the FDA most commonly granted drug approvals for advanced solid tumors based on improvements in PFS.²² Accordingly, caution needs to be exercised with the use of ASCO-VF scores computed from PFS-based surrogates in clinical and policy decision-making.

These findings build on the current literature, whereby evidence does not support the surrogacy of PFS in place of OS in studies of drugs for advanced prostate, non-small cell lung, and breast cancer.^{23–28} A review of PFS/TTP meta-analyses found that PFS and TTP were weakly associated with OS across tumor types, with only support for surrogacy in advanced ovarian and colorectal cancers. Even then, validity was only shown for cytotoxic agents, and it is unclear whether the same applies for novel targeted molecular therapies.²⁹ Neither HR PFS- nor mPFS-derived CBS correlated strongly with HR OS-derived CBS for both targeted and chemotherapeutic agents in this analysis. Clinically, even in settings where PFS may be believed to be an appropriate surrogate for OS, whether this remains true when the endpoint has been transformed to ASCO-VF scores is unknown.

Poor surrogacy of efficacy endpoints hinders the utility of the ASCO-VF to enable transparent and informed physician–patient discussion regarding treatment options. Confronted with alternative treatment regimens, patients and physicians must collectively assess the value of treatment options, comparing relative clinical benefit to cost. This poses a challenge to impartial decision-making if relative CBS are overestimated for some therapies, leading to false expectations among patients. This is especially true given the current findings of extremely weak correlation across all analyses of RR-derived CBS compared with HR OS-derived CBS, coupled with the fairly high tendency of 26% of RR-derived CBS to overestimate scores by at least 20 points.

Our study is not without limitations. We are unable to comment on CBS derived from surrogate endpoints for specific cancer indications. Even the settings that represented the largest number of trials included in this study all still had relatively small sample sizes, limiting statistical power. However, the same correlation trends between HR OS and surrogates were observed, with significant associations between CBS derived from HR OS and mOS yet weak associations between CBS derived from HR

OS and those from subsequent surrogate endpoints in the hierarchy. Only CBS from breast cancer RCTs deviated from these trends, but overinterpretation is cautioned because of multiplicity, findings caused by chance, and small sample size. Future work examining CBS involving much larger sample sizes of RCTs for a specific tumor site of interest may be warranted.

Conclusions

Moving forward, investigators, physicians, and policy-makers should rediscuss clinically meaningful outcomes and invest further efforts in achieving a consensus regarding how clinical benefit should be measured among various evaluative frameworks. Related discussions about clinical benefit have already begun. In 2014, the ASCO Cancer Research Committee formed tumor site-specific working groups to achieve a consensus on which thresholds for benefit can be interpreted as clinically meaningful in OS and PFS outcomes.²¹ Ultimately, the consistent evaluation of treatments will not only allow more informed decision-making between patients and physicians but also provide unifying measures of clinical benefit to better influence their value.³⁰

Acknowledgments

The Canadian Centre for Applied Research in Cancer Control is funded by Canadian Cancer Society Research Institute Grant 2015-703549. The authors wish to thank Mahin Qureshi for contributions in all aspects of this study.

Submitted October 27, 2018; accepted for publication June 18, 2019.

Previous presentation: An abstract was presented at the 2017 Canadian Centre for Applied Research in Cancer Control Conference, May 25–26, 2017, Toronto, Ontario, Canada, and at the 2017 ASCO Annual Meeting, June 2–6, 2017, Chicago, Illinois.

Author contributions: *Principal investigator:* Chan. *Search strategy:* Chan. *Study concept and design:* Chan. *Data acquisition:* Cheng, Jiang, McDonald, Arciero, Ezeife, Rahmadian. *Data analysis and interpretation:* Cheng, Chan. *Manuscript preparation:* Cheng, Cheung, Chan. *Critical revision:* All authors. *Guarantor of work and final approval:* All authors.

Disclosures: The authors have disclosed that they have not received any financial considerations from any person or organization to support the preparation, analysis, results, or discussion of this article.

Correspondence: Kelvin K.W. Chan, MD, MSc, PhD, Division of Medical Oncology, Odette Cancer Centre, Sunnybrook Health Sciences Centre, 2075 Bayview Avenue, T2-058, Toronto, Ontario M4N 3M5, Canada. Email: kelvin.chan@sunnybrook.ca

References

- 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; 34:2925–2934.
- 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;33:2563–2577.
- Booth CM, Cescon DW, Wang L, et al. Evolution of the randomized controlled trial in oncology over three decades. *J Clin Oncol* 2008;26:5458–5464.
- Cressman S, Browman GP, Hoch JS, et al. A time-trend economic analysis of cancer drug trials. *Oncologist* 2015;20:729–736.
- Saluja R, Arciero VS, Cheng S, et al. Examining trends in cost and clinical benefit of novel anticancer drugs over time. *J Oncol Pract* 2018;14:e280–294.
- 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? *J Clin Oncol* 2017; 35:2764–2771.

7. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice*, 2nd ed. Lexington, KY: OTexts; 2018.
8. Zaki R, Bulgiba A, Ismail R, et al. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. *PLoS One* 2012;7:e37908.
9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;327:307–310.
10. U.S. Food & Drug Administration. Hematology/Oncology (cancer) approvals and safety notifications. Available at: <https://www.fda.gov/drugs/informationondrugs/approveddrugs/ucm279174.htm>. Accessed May 30, 2018.
11. European Medicines Agency. European public assessment reports. Available at: http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/landing/epar_search.jsp&mid=WC0b01ac058001d125. Accessed May 30, 2018.
12. Health Canada. The drug and health product register. Available at: <https://hpr-rps.hres.ca/reg-content/summary-basis-decision-result.php?lang=en&term=>. Accessed May 30, 2018.
13. 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.
14. Cherny NI, Dafni U, Bogaerts J, et al. ESMO-Magnitude of Clinical Benefit Scale version 1.1. *Ann Oncol* 2017;28:2340–2366.
15. Saad ED, Buyse M. Statistical controversies in clinical research: end points other than overall survival are vital for regulatory approval of anticancer agents. *Ann Oncol* 2016;27:373–378.
16. Miller KD, Chap LI, Holmes FA, et al. Randomized phase III trial of capecitabine compared with bevacizumab plus capecitabine in patients with previously treated metastatic breast cancer. *J Clin Oncol* 2005;23:792–799.
17. Hellmann MD, Kris MG, Rudin CM. Medians and milestones in describing the path to cancer cures: telling “tails.” *JAMA Oncol* 2016;2:167–168.
18. Rahman R, Fell G, Trippa L, et al. Violations of the proportional hazards assumption in randomized phase III oncology clinical trials [abstract]. *J Clin Oncol* 2018;36(Suppl):Abstract 2543.
19. Saad ED, Katz A, Hoff PM, et al. Progression-free survival as surrogate and as true end point: insights from the breast and colorectal cancer literature. *Ann Oncol* 2010;21:7–12.
20. Zhao F. Surrogate end points and their validation in oncology clinical trials. *J Clin Oncol* 2016;34:1436–1437.
21. 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;32:1277–1280.
22. Robinson AG, Booth CM, Eisenhauer EA. Progression-free survival as an end-point in solid tumours—perspectives from clinical trials and clinical practice. *Eur J Cancer* 2014;50:2303–2308.
23. Burzykowski T, Molenberghs G, Buyse M, et al. Validation of surrogate end points in multiple randomized trials with failure time end points. *J R Stat Soc Ser C Appl Stat* 2001;50:405–422.
24. Booth CM, Eisenhauer EA. Progression-free survival: meaningful or simply measurable? *J Clin Oncol* 2012;30:1030–1033.
25. Sherrill B, Amonkar M, Wu Y, et al. Relationship between effects on time-to-disease progression and overall survival in studies of metastatic breast cancer. *Br J Cancer* 2008;99:1572–1578.
26. Hackshaw A, Knight A, Barrett-Lee P, et al. Surrogate markers and survival in women receiving first-line combination anthracycline chemotherapy for advanced breast cancer. *Br J Cancer* 2005;93:1215–1221.
27. Miksad RA, Zietemann V, Gothe R, et al. Progression-free survival as a surrogate endpoint in advanced breast cancer. *Int J Technol Assess Health Care* 2008;24:371–383.
28. Soria JC, Massard C, Le Chevalier T. Should progression-free survival be the primary measure of efficacy for advanced NSCLC therapy? *Ann Oncol* 2010;21:2324–2332.
29. 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.
30. Schnipper LE, Schilsky RL. Converging on the value of value frameworks. *J Clin Oncol* 2017;35:2732–2734.

See JNCCN.org for supplemental online content.

**CONGRESS
SERIES™**

**Live
Meeting**

Friday, February 7, 2020
Chicago, IL

REGISTER TODAY

NCCN.org/events

BREAST CANCER

with Updates from the 2019 San Antonio Breast Cancer Symposium

Hosted by Robert H. Lurie Comprehensive Cancer Center of Northwestern University

Agenda topics:

- Genetics/Familial Risk Assessment for Breast Cancer: Controversies and Evolving Principles
- Adjuvant and Neoadjuvant Therapies for Early Stage Breast Cancer, Including SABCS Updates
- Surgical Management of Breast Cancer, Including SABCS Updates
- Is There a Role for Proton Therapy in the Treatment of Breast Cancer?
- Principles of Breast Reconstruction after Surgery
- Biomarker Assays for Treatment Selection of Advanced Breast Cancer
- Management of Metastatic Breast Cancer, Including SABCS Updates
- Management of Toxicities with Targeted Therapies for Breast Cancer
- Management of Rare Breast Cancers

JNCCN-N-0277-1219

Supplemental online content for:

Are Surrogate Endpoints Unbiased Metrics in Clinical Benefit Scores of the ASCO Value Framework?

Sierra Cheng, BSc; Matthew C. Cheung, MD, SM; Di Maria Jiang, MD, FRCPC; Erica McDonald, BSc; Vanessa S. Arciero, BSc; Doreen Anuli Ezeife, MD, MSc; Amanda Rahmadian, BSc; Alexandra Chambers, MA; Kelley-Anne Sabarre, MSc; Ambika Parmar, MD; and Kelvin K.W. Chan, MD, MSc, PhD

J Natl Compr Canc Netw 2019;17(12):1489–1496

eFigure 1: Boxplot of CBS Derived From Endpoints

eFigure 2: Relationship Between CBS Derived From HR OS Compared With Surrogate Endpoints

eTable 1: Summary Statistics of Clinical Benefit Scores Derived From Endpoints

eTable 2: Spearman's Rank Correlations of Surrogate-Derived vs HR OS-Derived CBS

eFigure 1. Boxplot of CBS derived from endpoints: (A) interquartile ranges and (B) all scores.

Abbreviations: CBS, clinical benefit scores; HR OS, hazard ratio for death; HR PFS, hazard ratio for disease progression; mOS, median overall survival; mPFS, median progression-free survival; RR, response rate.

eFigure 2. Relationship between CBS derived from HR OS compared with surrogate endpoints: (A) mOS, (B) HR PFS, (C) mPFS, and (D) RR. Each line connecting HR OS-derived CBS to surrogate-derived CBS corresponds to an RCT. The lines should be as parallel to each other as possible for a good surrogate measure.

Abbreviations: CBS, clinical benefit scores; HR OS, hazard ratio for death; HR PFS, hazard ratio for disease progression; mOS, median overall survival; mPFS, median progression-free survival; RCT, randomized clinical trial; RR, response rate.

eTable 1. Summary Statistics of Clinical Benefit Scores Derived From Endpoints

	HR OS	mOS	HR PFS	mPFS	RR
Total	116	90	110	101	108
Minimum	−25.0	−13.2	−6.40	−36.2	1.12
Median (IQR)	23.0 (14.0–30.6)	22.3 (12.2–35.4)	29.6 (20.8–40.0)	40.0 (20.0–75.8)	23.5 (11.1–40.3)
Average	23.0	27.1	29.7	51.9	26.7
Maximum	72.0	191.8	68.0	216.3	70.00
SD	13.8	26.8	15.1	50.3	18.4

Abbreviations: HR OS, hazard ratio for death; HR PFS, hazard ratio for disease progression; IQR, interquartile range; mOS, median overall survival; mPFS, median progression-free survival; RR, response rate.

eTable 2. Spearman's Rank Correlations of Surrogate-Derived vs HR OS–Derived CBS

RCTs	mOS vs HR OS			HR PFS vs HR OS			mPFS vs HR OS			RR vs HR OS		
	n	r_s	P Value	n	r_s	P Value	n	r_s	P Value	n	r_s	P Value
All RCTs	90	0.80	<.001	110	0.38	<.001	101	0.20	.05	108	0.01	.95
Reporting on all endpoints ^a	77	0.80	<.001	77	0.39	<.001	77	0.13	.25	77	−0.03	.78
Evaluating first-line therapy	30	0.73	<.001	38	0.36	.03	35	0.35	.04	40	0.14	.40
OS as primary endpoint	55	0.78	<.001	54	0.52	<.001	53	0.25	.07	54	0.15	.29
Non-OS primary endpoints	35	0.76	<.001	56	0.29	.05	48	0.11	.49	54	0.12	.44
PFS as primary endpoint	29	0.77	<.001	49	0.30	.03	41	0.17	.28	46	0.16	.28
Evaluating hematologic malignancies only	12	0.51	.09	25	0.39	.06	19	0.07	.77	24	0.12	.57
Excluding hematologic malignancies	78	0.81	<.001	85	0.39	<.001	82	0.21	.06	84	−0.08	.49
Evaluating lung cancer	12	0.97	<.001	13	0.45	.12	12	0.01	.97	13	0.21	.50
Evaluating breast cancer	8	−0.12	.77	9	−0.46	.21	8	−0.63	.10	9	0.80	.01
Evaluating colorectal cancer	15	0.78	<.001	15	0.50	.06	14	0.04	.89	14	−0.50	.07
Evaluating melanoma	6	0.89	.02	9	0.67	.05	8	0.49	.22	10	0.07	.85
Evaluating targeted agent therapies	56	0.79	<.001	76	0.33	<.01	69	0.15	.22	76	0.04	.73
Evaluating chemotherapeutic agents	27	0.84	<.001	28	0.48	<.01	26	0.24	.23	27	−0.19	.33
	HR OS vs HR PFS			mOS vs HR PFS			mPFS vs HR PFS			RR vs HR PFS		
	n	r_s	P Value	n	r_s	P Value	n	r_s	P Value	n	r_s	P Value
PFS primary endpoint compared with HR PFS	52	0.30	.03	30	0.38	.04	52	0.86	<.001	58	−0.07	.61

Abbreviations: CBS, clinical benefit scores; HR OS, hazard ratio for death; HR PFS, hazard ratio for disease progression; IQR, interquartile range; mOS, median overall survival; mPFS, median progression-free survival; RCT, randomized clinical trial; RR, response rate.