

Measuring the Value of New Drugs: Validity and Reliability of 4 Value Assessment Frameworks in the Oncology Setting

Tanya G.K. Bentley, PhD; Joshua T. Cohen, PhD; Elena B. Elkin, PhD; Julie Huynh, MD; Arnab Mukherjee, DrPH, MPH; Thanh H. Neville, MD, MSHS; Matthew Mei, MD; Ronda Copher, PhD; Russell Knoth, PhD; Ioana Popescu, MD, MPH; Jackie Lee, BS; Jenelle M. Zambrano, DNP, CNS, RN; and Michael S. Broder, MD, MSHS

ABSTRACT

BACKGROUND: Several organizations have developed frameworks to systematically assess the value of new drugs.

OBJECTIVE: To evaluate the convergent validity and interrater reliability of 4 value frameworks to understand the extent to which these tools can facilitate value-based treatment decisions in oncology.

METHODS: Eight panelists used the American Society of Clinical Oncology (ASCO), European Society for Medical Oncology (ESMO), Institute for Clinical and Economic Review (ICER), and National Comprehensive Cancer Network (NCCN) frameworks to conduct value assessments of 15 drugs for advanced lung and breast cancers and castration-refractory prostate cancer. Panelists received instructions and published clinical data required to complete the assessments, assigning each drug a numeric or letter score. Kendall's Coefficient of Concordance for Ranks (Kendall's *W*) was used to measure convergent validity by cancer type among the 4 frameworks. Intraclass correlation coefficients (ICCs) were used to measure interrater reliability for each framework across cancers. Panelists were surveyed on their experiences.

RESULTS: Kendall's *W* across all 4 frameworks for breast, lung, and prostate cancer drugs was 0.560 ($P=0.010$), 0.562 ($P=0.010$), and 0.920 ($P<0.001$), respectively. Pairwise, Kendall's *W* for breast cancer drugs was highest for ESMO-ICER and ICER-NCCN ($W=0.950$, $P=0.019$ for both pairs) and lowest for ASCO-NCCN ($W=0.300$, $P=0.748$). For lung cancer drugs, *W* was highest pairwise for ESMO-ICER ($W=0.974$, $P=0.007$) and lowest for ASCO-NCCN ($W=0.218$, $P=0.839$); for prostate cancer drugs, pairwise *W* was highest for ICER-NCCN ($W=1.000$, $P<0.001$) and lowest for ESMO-ICER and ESMO-NCCN ($W=0.900$, $P=0.052$ for both pairs). When ranking drugs on distinct framework subdomains, Kendall's *W* among breast cancer drugs was highest for certainty (ICER, NCCN: $W=0.908$, $P=0.046$) and lowest for clinical benefit (ASCO, ESMO, NCCN: $W=0.345$, $P=0.436$). Among lung cancer drugs, *W* was highest for toxicity (ASCO, ESMO, NCCN: $W=0.944$, $P<0.001$) and lowest for certainty (ICER, NCCN: $W=0.230$, $P=0.827$); and among prostate cancer drugs, it was highest for quality of life (ASCO, ESMO: $W=0.986$, $P=0.003$) and lowest for toxicity (ASCO, ESMO, NCCN: $W=0.200$, $P=0.711$). ICC (95% CI) for ASCO, ESMO, ICER, and NCCN were 0.800 (0.660-0.913), 0.818 (0.686-0.921), 0.652 (0.466-0.834), and 0.153 (0.045-0.371), respectively. When scores were rescaled to 0-100, NCCN provided the narrowest band of scores. When asked about their experiences using the ASCO, ESMO, ICER, and NCCN frameworks, panelists generally agreed that the frameworks were logically organized and reasonably easy to use, with NCCN rated somewhat easier.

CONCLUSIONS: Convergent validity among the ASCO, ESMO, ICER, and NCCN frameworks was fair to excellent, increasing with clinical benefit subdomain concordance and simplicity of drug trial data. Interrater reliability, highest for ASCO and ESMO, improved with clarity of instructions and specificity of score definitions. Continued use, analyses, and refinements of

these frameworks will bring us closer to the ultimate goal of using value-based treatment decisions to improve patient care and outcomes.

J Manag Care Spec Pharm. 2017;23(6-a):S34-S48

Copyright © 2017, Academy of Managed Care Pharmacy. All rights reserved.

What is already known about this subject

- In response to rapidly increasing health care spending, several organizations have developed frameworks to systematically assess the value of new drugs.
- Value assessment frameworks vary in their definition of components of value (efficacy, toxicity, quality of life) and their approach to assessment (quantitative or qualitative).
- To date, most published assessments of value frameworks have been primarily conceptual, and it has been unclear whether the frameworks provide valid and reliable measurements of value when a broad array of drugs are considered.

What this study adds

- This study represents one of the first quantitative assessments of the convergent validity and interrater reliability of the ASCO, ESMO, ICER, and NCCN value assessment frameworks.
- Overall concordance was fair to excellent, strongly influenced by concordance among clinical efficacy scores. Interrater reliability improved with clarity of instructions and specificity of score definitions.
- These frameworks appear to measure a similar underlying concept. Their continued use, evaluation, and refinements will advance the ultimate goal of using value-based treatment decisions to improve patient care and outcomes.

The soaring cost of new cancer drugs has prompted calls for more attention to the value of these products.¹⁻⁵ In response, several organizations, including the American Society of Clinical Oncology (ASCO), European Society for Medical Oncology (ESMO), Institute for Clinical and Economic Review (ICER), and National Comprehensive Cancer Network (NCCN), have developed tools for value assessment. While these tools all purport to provide some overall measure of value, they vary considerably in their definitions of the

components of value (efficacy, toxicity, quality of life) and their approach to assessment (quantitative or qualitative).

Value assessment frameworks also vary with regard to their intended use and users. Some, including ASCO and NCCN, have been advocated to facilitate shared decision making between oncologists and their patients.⁶⁻¹⁰ Others, such as ESMO and ICER, are intended primarily for use by policymakers or payers.¹¹⁻¹⁴ Independent of their intended use and target audience, all of these tools are designed to be applicable across a spectrum of agents and cancers.

To date, most published assessments of value frameworks have been primarily conceptual or editorial.¹⁵⁻²⁴ Although 2 recent analyses evaluated value frameworks' validity and reliability, the generalizability of their findings to a broader array of frameworks and drugs is uncertain. It thus remains unclear whether existing oncology value assessment frameworks provide valid and reliable measurements of a drug's or regimen's value. Our objective was to evaluate the convergent validity and interrater reliability of 4 of the most widely discussed frameworks.

Methods

This study expanded on a pilot study, of which detailed methods are described elsewhere.²² In the current study, we convened a panel of 8 clinicians and health services researchers to assess the value of new cancer drugs using the ASCO, ESMO, ICER, and NCCN value frameworks. The panel included 4 oncologists, 2 non-oncologist physicians, and 2 doctorate-level health services researchers. All panelists had prior experience reviewing clinical literature in the oncology setting.

A total of 15 drugs were assessed in 3 cancers: advanced lung, advanced breast, and castration-refractory prostate cancers (see Appendix). To select the drugs for study inclusion, we first identified more prevalent and costly cancers,²⁵ and from these, we selected cancers with drugs for which published value scores were available. We considered drugs representing a range of indications (curative and palliative), malignancies (solid and hematologic), and mechanisms (cytotoxic, biologic, and immunologic). After expert review, we developed a list of 12 cancers and 90 drugs for consideration. To evaluate the study's feasibility, a pilot phase was conducted with 5 of these drugs for advanced lung cancer.²² For the full study presented here, we examined 15 drugs for which sufficient published evidence existed for calculating the value scores; these drugs included the 5 pilot drugs and 5 each for advanced breast cancer and prostate cancer. The study sponsor played no role in determining the included cancers or drugs.

We provided panelists with published efficacy and safety data from phase III randomized controlled trials for each included drug to conduct the assessments (Appendix). We evaluated convergent validity of the 4 frameworks using Kendall's Coefficient of Concordance for Ranks (Kendall's *W*), and interrater

reliability using intraclass correlation coefficients (ICCs). After completing their value assessments for each drug and framework (15 drugs × 4 frameworks = 60 assessments per panelist), panelists provided comments regarding their experiences and answered survey questions that asked them to rate the different frameworks. The panelists conducted a total of 480 assessments.

Each framework produces scores on different scales. The ASCO "Net Health Benefit" score ranges from -20 (worst) to +180 (best) and reflects drug clinical efficacy, toxicity, effects on long-term survival, palliation, quality of life, and treatment-free interval. The ESMO score ranges from 1 (worst) to 5 (best) and reflects efficacy, toxicity, and quality of life.

Unlike the other frameworks, ICER does not produce scores that can be ranked. Instead, it comprises multiple components, including comparative clinical effectiveness, cost-effectiveness, and budget impact, each of which requires a specific methodology and an in-depth analysis.

To produce ICER scores that could be ranked, we used ICER's comparative clinical effectiveness component, the Evidence Rating Matrix. This online tool reports final grades from I (worst) to A (best) based on the comparative net benefits and the level of certainty associated with these benefits. Our analysis converted these grades to a numerical scale from 0 (worst) to 4 (best).

NCCN framework scores range from 1 (worst) to 5 (best) for each of the 4 health benefit measures: efficacy, safety, quality of evidence, and consistency of evidence. We averaged these scores for each of these components. The impact of this approach was evaluated in sensitivity analyses by only considering NCCN scores from the clinical subdomain, the measure most commonly represented in the other frameworks.

Analysis

We estimated mean scores and standard deviations (SDs) for each drug and framework, overall and by subdomain. For descriptive comparisons, we also rescaled means to 0-100. Convergent validity and interrater reliability were the primary outcomes. Convergent validity measures the extent to which each framework produced similar evaluations for the same list of drugs. For this analysis, convergent validity was evaluated using Kendall's *W*. Kendall's *W* measures the agreement of ranked items and was calculated by comparing ranked mean drug scores (i.e., from 1 to 5 to represents best to worst drug scores per cancer type) among the 4 frameworks. Kendall's *W* ranges from 0 (no agreement) to 1 (complete agreement). *P* values were reported to test the alternative hypothesis of complete agreement ($W > 0$) against the null hypothesis (no agreement). *W* may be interpreted using a scale similar to that for ICC (described in the Analysis section).

For each of the 3 cancer types, we calculated Kendall's *W*: overall, across the 4 frameworks; within each pair of frameworks; for framework subdomains of clinical benefit,

Measuring the Value of New Drugs: Validity and Reliability of 4 Value Assessment Frameworks in the Oncology Setting

TABLE 1 Framework Mean Scores for 15 Drugs, Overall, and by Subdomain^a

	ASCO	ESMO	ICER	NCCN
Breast Cancer Drugs				
Overall, mean (SD)				
A	53.51 (9.17)	2.75 (0.71)	3.44 (0.42)	3.41 (0.46)
B	16.45 (2.38)	1.63 (0.52)	3.00 (0.53)	3.38 (0.42)
C	55.26 (14.27)	2.25 (0.71)	1.88 (1.55)	3.34 (0.30)
D	43.19 (5.42)	4.75 (0.46)	3.81 (0.26)	3.75 (0.50)
E	36.25 (5.50)	3.88 (0.35)	3.63 (0.52)	3.97 (0.53)
Clinical benefit				
A	42.53 (4.27)	3.00 (0.00)	- (-)	3.13 (0.99)
B	19.00 (0.00)	1.63 (0.52)	- (-)	3.25 (0.89)
C	54.50 (11.22)	3.00 (0.00)	- (-)	3.13 (0.35)
D	32.00 (0.00)	3.88 (0.35)	- (-)	3.50 (0.53)
E	32.00 (0.00)	3.75 (0.71)	- (-)	4.13 (0.64)
Toxicity				
A	2.99 (1.69)	0.00 (0.00)	- (-)	3.00 (0.76)
B	-2.56 (2.38)	0.00 (0.00)	- (-)	2.75 (0.71)
C	-9.74 (4.64)	-0.13 (0.35)	- (-)	2.88 (0.83)
D	1.19 (0.96)	0.00 (0.00)	- (-)	3.88 (0.64)
E	-0.65 (1.05)	0.00 (0.00)	- (-)	3.38 (0.74)
Quality of life				
A	0.00 (0.00)	0.13 (0.35)	- (-)	- (-)
B	0.00 (0.00)	0.00 (0.00)	- (-)	- (-)
C	1.25 (3.54)	0.00 (0.00)	- (-)	- (-)
D	7.50 (4.63)	0.63 (0.52)	- (-)	- (-)
E	1.25 (3.54)	0.00 (0.00)	- (-)	- (-)
Certainty ^b				
A	- (-)	- (-)	1.88 (0.35)	3.75 (0.46)
B	- (-)	- (-)	2.25 (0.46)	3.75 (0.53)
C	- (-)	- (-)	2.25 (0.89)	3.69 (0.46)
D	- (-)	- (-)	1.38 (0.52)	3.81 (0.59)
E	- (-)	- (-)	1.63 (0.74)	4.19 (0.70)
Lung Cancer Drugs				
Overall, mean (SD)				
F	73.08 (13.20)	4.00 (0.00)	3.63 (0.35)	3.66 (0.52)
G	73.85 (6.23)	3.38 (0.74)	3.63 (0.23)	3.66 (0.76)
H	63.81 (11.27)	4.88 (0.35)	3.75 (0.27)	3.94 (0.46)
I	40.95 (10.65)	2.75 (0.46)	3.19 (0.53)	3.72 (0.59)
J	11.49 (8.77)	2.00 (0.76)	3.19 (0.46)	3.69 (0.53)
Clinical benefit				
F	45.38 (3.89)	3.00 (0.00)	- (-)	3.38 (0.52)
G	50.35 (0.14)	3.00 (0.00)	- (-)	3.13 (0.99)
H	41.00 (0.00)	4.00 (0.00)	- (-)	3.75 (0.89)
I	29.00 (0.00)	2.00 (0.00)	- (-)	3.50 (0.93)
J	21.00 (0.00)	2.00 (0.76)	- (-)	3.25 (0.89)
Toxicity				
F	0.71 (2.56)	0.13 (0.35)	- (-)	3.38 (0.92)
G	3.00 (3.79)	0.50 (0.53)	- (-)	3.50 (0.93)
H	13.31 (4.20)	0.88 (0.35)	- (-)	4.13 (0.64)
Lung Cancer Drugs				
Overall, mean (SD)				
K	47.90 (7.05)	4.63 (0.74)	3.56 (0.32)	3.75 (0.53)
L	37.67 (12.30)	2.50 (0.76)	2.44 (0.90)	3.34 (0.42)
M	37.41 (10.16)	2.25 (0.46)	3.00 (0.53)	3.72 (0.49)
N	51.65 (13.54)	4.13 (0.35)	3.81 (0.26)	3.94 (0.53)
O	-5.76 (20.15)	1.13 (0.35)	0.63 (0.74)	3.13 (0.57)
Clinical benefit				
K	35.00 (0.00)	3.63 (0.74)	- (-)	3.50 (0.53)
L	30.00 (0.00)	2.38 (0.74)	- (-)	3.38 (0.92)
M	24.00 (0.00)	1.50 (0.53)	- (-)	3.25 (0.71)
N	35.67 (16.92)	3.13 (0.35)	- (-)	3.63 (0.52)
O	-9.77 (16.56)	1.00 (0.00)	- (-)	2.25 (1.28)
Toxicity				
K	-2.10 (2.17)	0.00 (0.00)	- (-)	3.63 (0.92)
L	-3.58 (3.64)	0.00 (0.00)	- (-)	2.63 (0.74)
M	-1.59 (4.59)	0.00 (0.00)	- (-)	3.50 (0.93)
N	-4.78 (2.08)	0.00 (0.00)	- (-)	4.00 (0.76)
O	1.51 (2.81)	0.00 (0.00)	- (-)	3.00 (0.76)
Quality of life				
K	10.00 (0.00)	1.00 (0.00)	- (-)	- (-)
L	0.00 (0.00)	0.00 (0.00)	- (-)	- (-)
M	7.50 (4.63)	0.75 (0.46)	- (-)	- (-)
N	10.00 (0.00)	1.00 (0.00)	- (-)	- (-)
O	1.25 (3.54)	0.00 (0.00)	- (-)	- (-)
Certainty ^b				
K	- (-)	- (-)	1.63 (0.52)	3.94 (0.62)
L	- (-)	- (-)	1.88 (0.83)	3.69 (0.46)
M	- (-)	- (-)	1.75 (0.89)	4.06 (0.56)
N	- (-)	- (-)	1.38 (0.52)	4.06 (0.73)
O	- (-)	- (-)	1.75 (1.04)	3.63 (0.35)

^aAssessments from 8 panelists for each drug and framework; N=480 total assessments.

^bIn the certainty subdomain for ICER, lower scores represent higher rankings.

ASCO=American Society of Clinical Oncology; ESMO=European Society for Medical Oncology; ICER=Institute for Clinical and Economic Review; NCCN=National Comprehensive Cancer Network; SD=standard deviation.

toxicity, quality of life, and certainty; by individual panelist characteristics (oncologists vs. non-oncologists; physicians vs. nonphysicians); and for each individual panelist.

We assessed interrater reliability of each framework across all cancers using ICC and 95% confidence intervals (CIs). ICC measures the extent to which independent panelists arrive at the same assessment. Panelist scores for each framework were used to calculate ICC. ICC ranges from 0 to 1, where values <0.40 are generally taken to represent poor reliability, 0.40-0.59 fair, 0.60-0.74 good, and ≥ 0.75 excellent.^{26,27} Kendall's *W* has been interpreted using the same categories.^{26,27}

ICC calculations assumed that the 8 panelists represented a random sample from a larger population of framework users. Each panelist evaluated the same drugs with the 4 frameworks. In sensitivity analyses, we calculated ICC with each panelist removed one at a time. We also calculated ICC between oncologist versus non-oncologist and physician versus nonphysician panelists, and for the following subdomains relevant in each framework: clinical benefit, toxicity, quality of life, and certainty.

Data was collected using electronic forms exported into Excel, and analyzed using SAS software version 9.4 (SAS Institute, Cary, NC). All tests were 2-sided with a significance level of 0.05.

Results

Overall and subdomain mean scores and SDs for each of the 15 drugs using the ASCO, ESMO, ICER, and NCCN frameworks are shown by cancer type in Table 1.

Convergent Validity, by Cancer Type

Figures 1-3 show, by cancer type, drug rankings and rescaled values for each framework, as well as Kendall's *W* overall, pairwise, and by subdomain.

For breast cancer drugs, Kendall's *W* was 0.560 ($P=0.010$) across the 4 frameworks (Figure 1, Panel 1). Pairwise across frameworks, Kendall's *W* ranged from 0.300 to 0.950, highest for ESMO-ICER and ICER-NCCN and lowest for ASCO-ICER and ASCO-NCCN. When breast cancer drugs were ranked on the basis of distinct framework subdomains (Figure 1, Panels 2-5), Kendall's *W* was highest for certainty (ICER and NCCN) and lowest for clinical benefit (ASCO, ESMO, and NCCN). When Kendall's *W* was assessed for 1 panelist at a time, it remained below the mean of 0.560 for all but 2 panelists, for whom *W* was 0.630 ($P=0.003$) and 0.657 ($P=0.002$; data not shown).

For lung cancer drugs, Kendall's *W* was 0.562 ($P=0.010$) across the 4 frameworks (Figure 2, Panel 1). Pairwise across frameworks, Kendall's *W* ranged from 0.218 to 0.974, highest for ESMO-ICER and lowest for ASCO-NCCN. When lung cancer drugs were ranked on the basis of distinct framework subdomains (Figure 2, Panels 2-5), Kendall's *W* was highest for toxicity (ASCO, ESMO, and NCCN) and lowest for certainty (ICER and NCCN). When Kendall's *W* was assessed for 1 panelist at a time, it remained

below the mean of 0.562 for all but 2 panelists, for whom *W* was 0.572 ($P=0.009$) and 0.777 ($P<0.001$; data not shown).

For prostate cancer drugs, Kendall's *W* was 0.920 ($P<0.001$) across the 4 frameworks (Figure 3, Panel 1). Pairwise across frameworks, Kendall's *W* was ≥ 0.900 for all pairs and highest for ICER-NCCN. When prostate cancer drugs were ranked on the basis of distinct framework subdomains (Figure 3, Panels 2-5), Kendall's *W* was highest for quality of life (ASCO and ESMO) and lowest for toxicity (ASCO, ESMO, and NCCN). When Kendall's *W* was assessed for 1 panelist at a time, it remained below the mean of 0.920 for all but 2 panelists, for whom *W* was 0.946 ($P<0.001$) and 0.930 ($P<0.001$; data not shown). In sensitivity analyses, *W* results did not change substantially for any cancers when only the NCCN clinical subdomain was used.

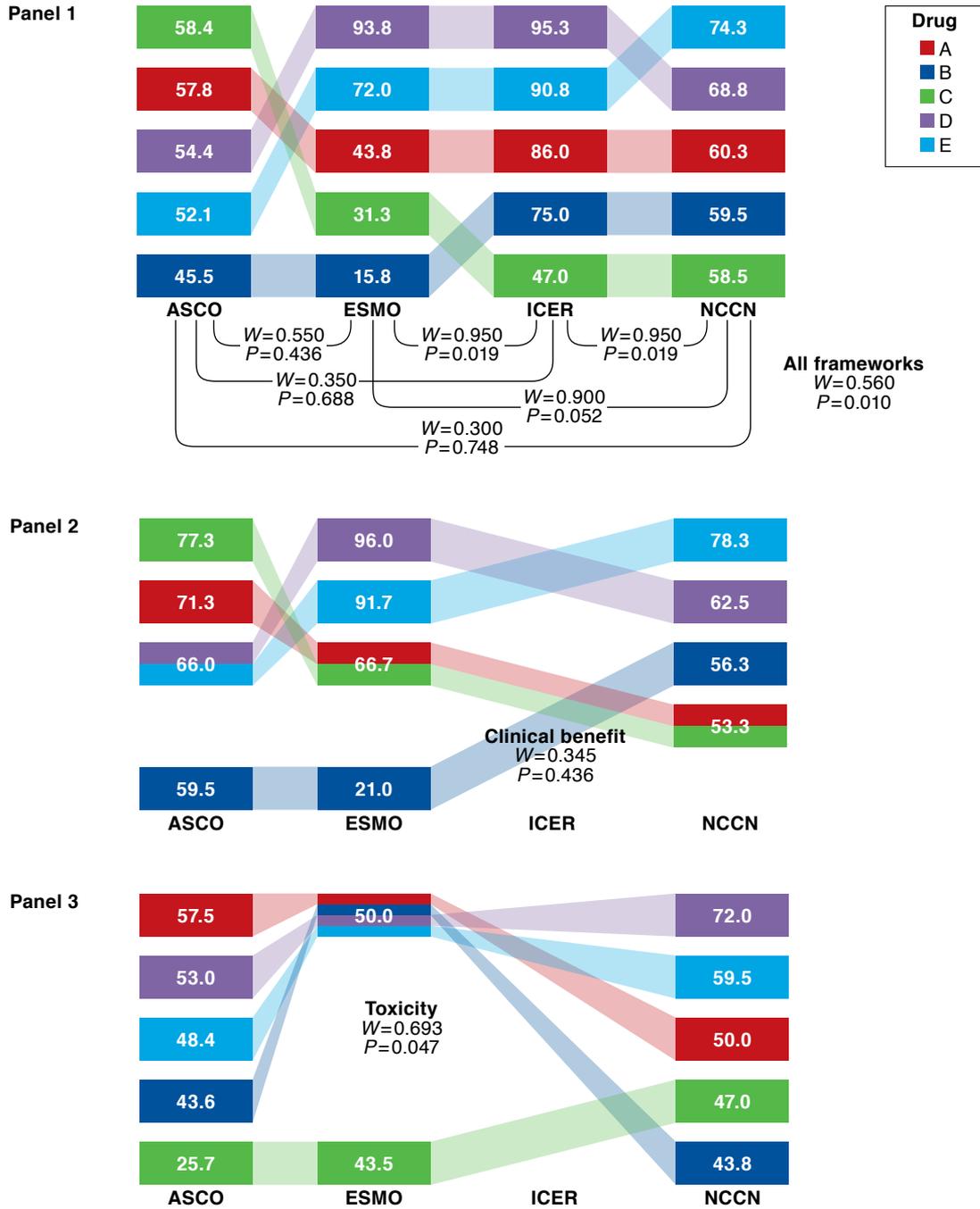
Interrater Reliability, Across Cancers

Table 2 shows ICC results by framework. ICCs for the panelists' assessments using the ASCO, ESMO, ICER, and NCCN frameworks were 0.800 (95% CI=0.660-0.913), 0.818 (95% CI=0.686-0.921), 0.652 (95% CI=0.466-0.834), and 0.153 (95% CI=0.045-0.371), respectively. For all frameworks, ICC results were similar although slightly higher among oncologists compared with non-oncologists. For ASCO, ESMO, and ICER, ICC results were higher among physicians compared with non-physicians; for NCCN, they were similar yet slightly lower for physicians compared with nonphysicians. When panelists were removed one at a time, ICC did not differ from the base case by more than 0.05 for any framework, with ranges of 0.046 for ASCO, 0.037 for ESMO, 0.070 for ICER, and 0.069 for NCCN. When framework subdomains were considered, ICC for clinical benefit was high for both ASCO and ESMO. For ASCO, ICC was also high in toxicity and lower for quality of life. For ESMO, ICC was high for quality of life and lower for toxicity. The ICC for all ICER and NCCN subdomains was <0.200.

Each value assessment took panelists approximately 25 minutes using the ASCO framework, 14 minutes using ESMO, 21 minutes using ICER, and 8 using NCCN. The mean (SD) times needed to review the literature (up to 2 manuscripts) for each drug assessed were 28 (20) minutes for ASCO, 22 (15) for ESMO, 25 (20) for ICER, and 11 (4) for NCCN, the framework for which all assessments were done last. Mean time to review literature was consistent among cancer types, considering all frameworks and excluding the panelists' first drug assessed.

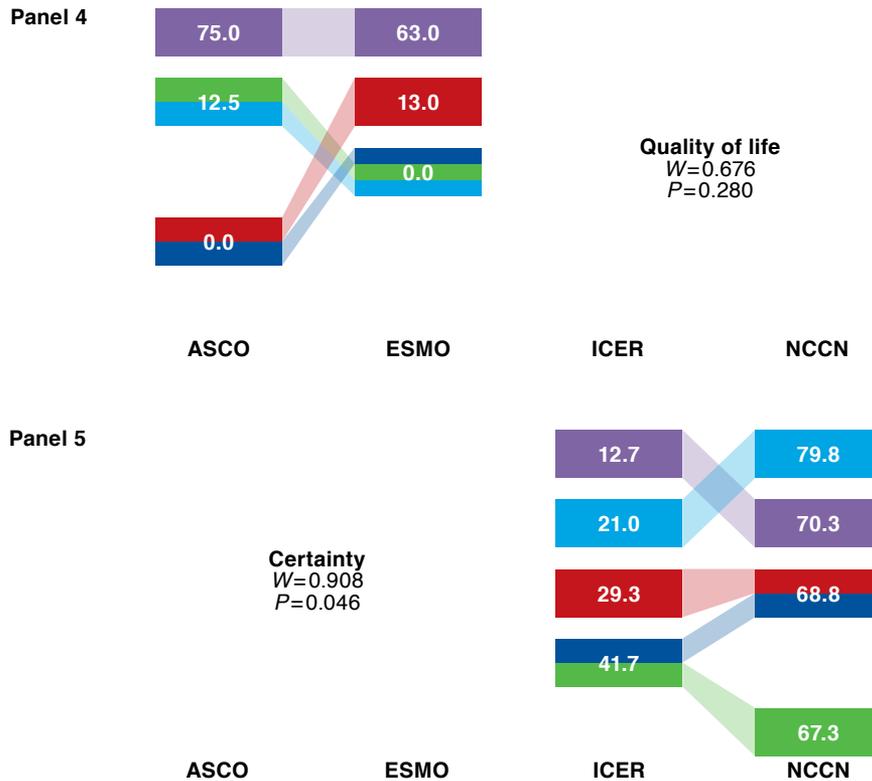
When asked about their experiences using the ASCO, ESMO, ICER, and NCCN frameworks, panelists somewhat agreed that the frameworks were logically organized and reasonably easy to use, with NCCN rated somewhat easier. Panelists neither agreed nor disagreed on whether they would be comfortable using the frameworks for assessing the value of cancer treatment for a loved one.

FIGURE 1 Kendall's *W* Among 4 Frameworks Overall, Pairwise, and by Subdomain: Rankings and Rescaled Scores of 5 Breast Cancer Drugs



continued on next page

FIGURE 1 Kendall's *W* Among 4 Frameworks Overall, Pairwise, and by Subdomain: Rankings and Rescaled Scores of 5 Breast Cancer Drugs (continued)



Note: Columns show drug rankings for each framework and rescaled mean scores (range 0-100). In Panel 1, Kendall's *W* is shown as a measure of concordance across all frameworks and each pairwise comparison. In Panels 2-5, it is shown for each subdomain. In the certainty subdomain for ICER (Panel 5), lower scores represent higher rankings.

Discussion

This analysis represents one of the first quantitative assessments of the convergent validity and interrater reliability of the ASCO, ESMO, ICER, and NCCN value assessment frameworks. The frameworks demonstrated fair-to-excellent overall concordance with one another, suggesting they measure a similar underlying concept, although, lacking a gold standard, we could not assess the extent to which that concept is one that physicians or patients would recognize as “value.” All frameworks except NCCN demonstrated good-to-excellent reliability.

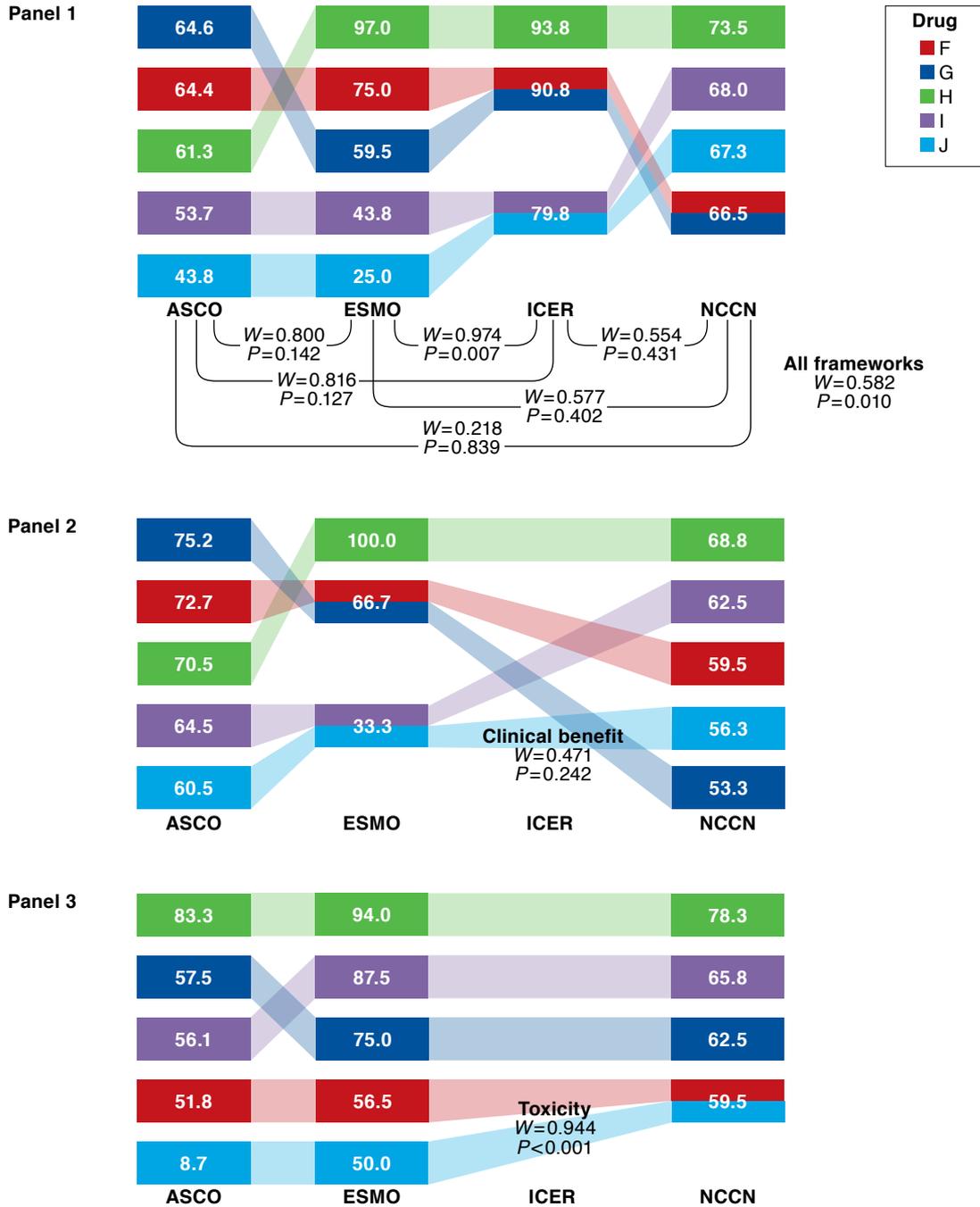
Clinical efficacy concordance had the greatest influence on overall concordance. For example, for breast cancer, clinical benefit concordance was poor (0.345) and, despite good or excellent concordance in toxicity and quality of life, overall concordance was only fair (0.560). In contrast, for prostate cancer, concordance was excellent for clinical benefit (0.956), and despite poor concordance for toxicity, overall concordance was excellent (0.920). That clinical benefit drives overall

concordance is consistent with efficacy being a primary driver of real-world clinical decisions and reflects the fact that frameworks place substantial weight on clinical benefit. These findings also provide evidence for the frameworks’ face validity, indicating that framework-driven decisions may reflect those made in clinical practice.²⁸

The complexity of the underlying data used to complete the assessments may have in turn influenced clinical concordance. Specifically, panelists appeared to more easily apply the correct numbers for the prostate cancer drugs than they did for breast and lung cancer drugs. That may be because in general, the prostate cancer papers each reported trial outcomes (i.e., hazard ratios) for fewer subgroups or endpoints than did the breast and lung cancer papers, and hence were easier to understand.

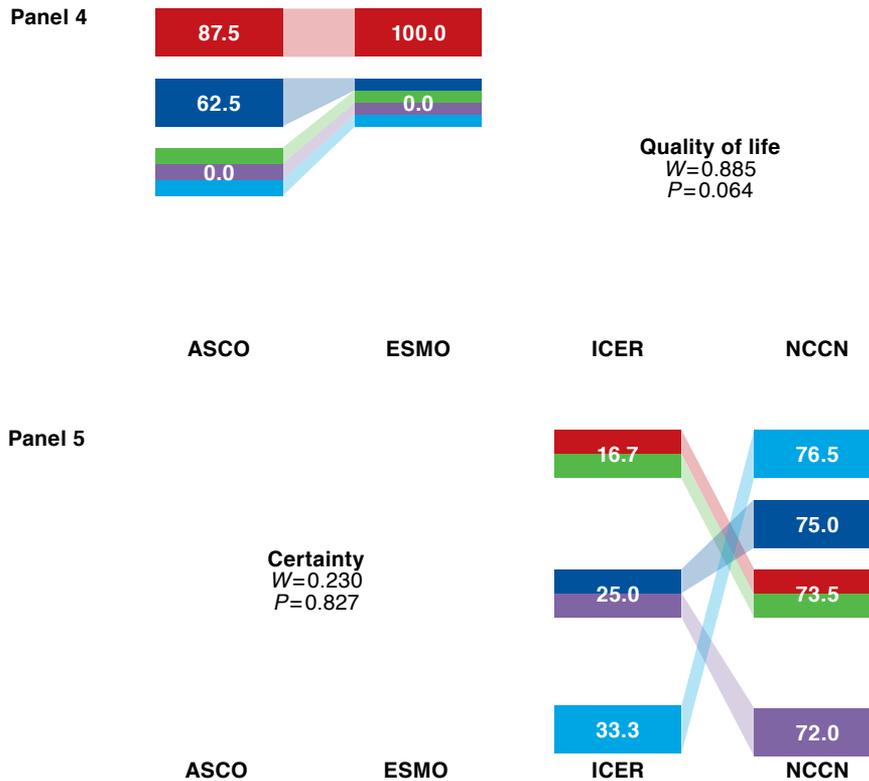
The ASCO, ESMO, and ICER frameworks demonstrated good-to-excellent reliability. Reliability of the NCCN framework was poor. In contrast with the other frameworks, which provide relatively detailed instructions and definitions, NCCN uses only short phrases in both instructions and category

FIGURE 2 Kendall's *W* Among 4 Frameworks Overall, Pairwise, and by Subdomain: Rankings and Rescaled Scores of 5 Lung Cancer Drugs



continued on next page

FIGURE 2 Kendall's *W* Among 4 Frameworks Overall, Pairwise, and by Subdomain: Rankings and Rescaled Scores of 5 Lung Cancer Drugs (continued)



Note: Columns show drug rankings for each framework and rescaled mean scores (range 0-100). In Panel 1, Kendall's *W* is shown as a measure of concordance across all frameworks and each pairwise comparison. In Panels 2-5, it is shown for each subdomain. In the certainty subdomain for ICER (Panel 5), lower scores represent higher rankings.

definitions. For example, efficacy, rated on a 5-point scale, can be described as “very effective,” defined as “sometimes provides long-term survival advantage or has curative potential,” or “highly effective,” defined as “often provides long-term survival advantage or has curative potential.” No definitions or anchors are given for the terms “sometimes” or “often,” which are the only differences between these definitions.

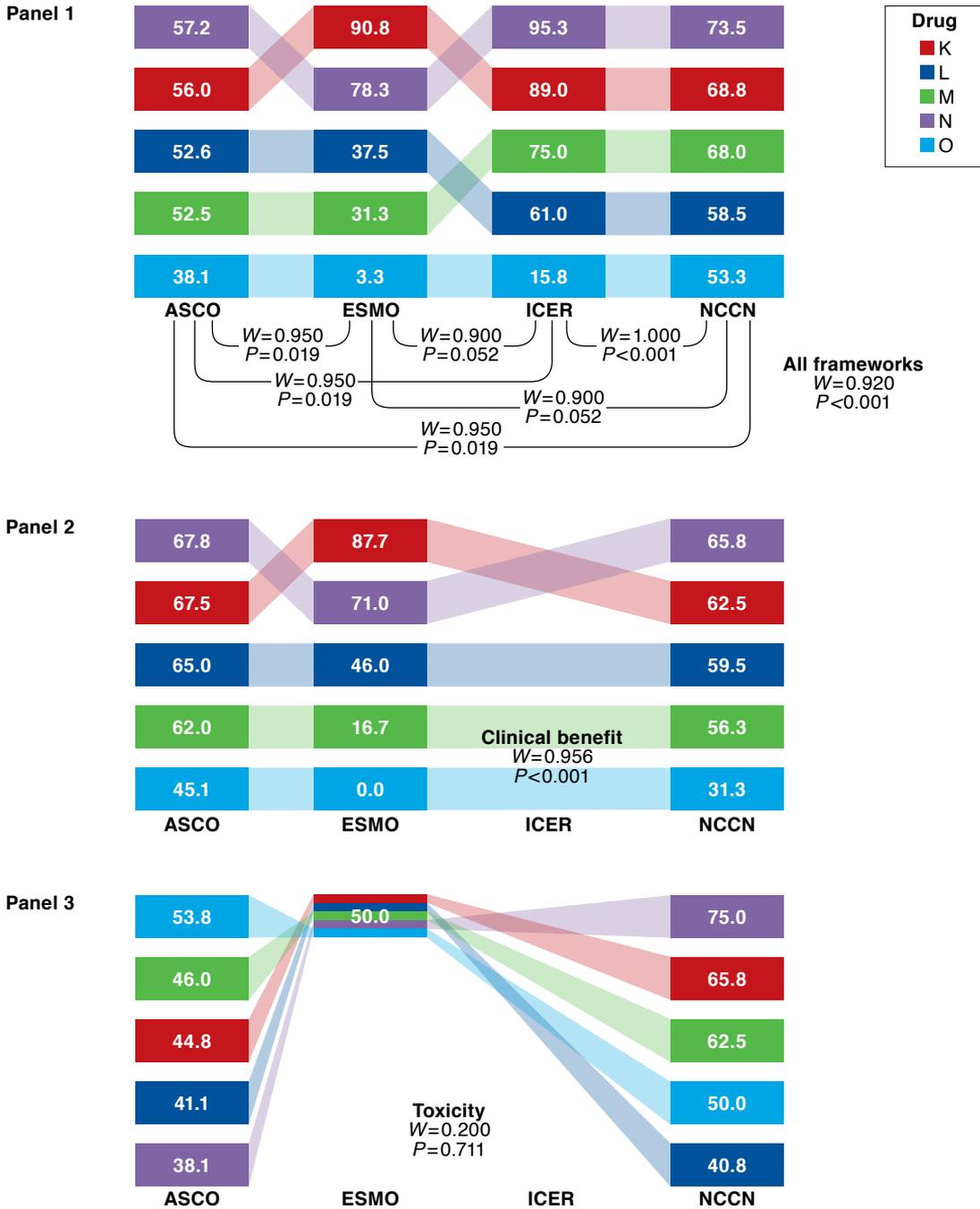
Conducting the NCCN assessments took less time on average than the others, but sparse instructions and broad categories may have limited the NCCN framework's ability to discriminate among drugs. The range of scores produced by the NCCN framework was narrow (with ratings for lung cancer drugs spanning only 7 points on a 100-point scale). In this analysis, ICC reflects the ratio of variation among drugs to the total variation, which includes variation among drugs and among raters. When the variation among drugs is small, most of the variation is explained by differences among raters.

Unlike NCCN, ASCO and ESMO frameworks incorporate detailed, mathematical approaches, with strict criteria or formulas for calculating benefit. ICER, while not formulaic,

provides detailed instructions and anchors terms such as “comparable” and “small/incremental” with half a dozen specific examples. The low interrater reliability suggests that the NCCN framework in its current form would perform best when based on large numbers of individual raters (as is now done for NCCN's published assessments).

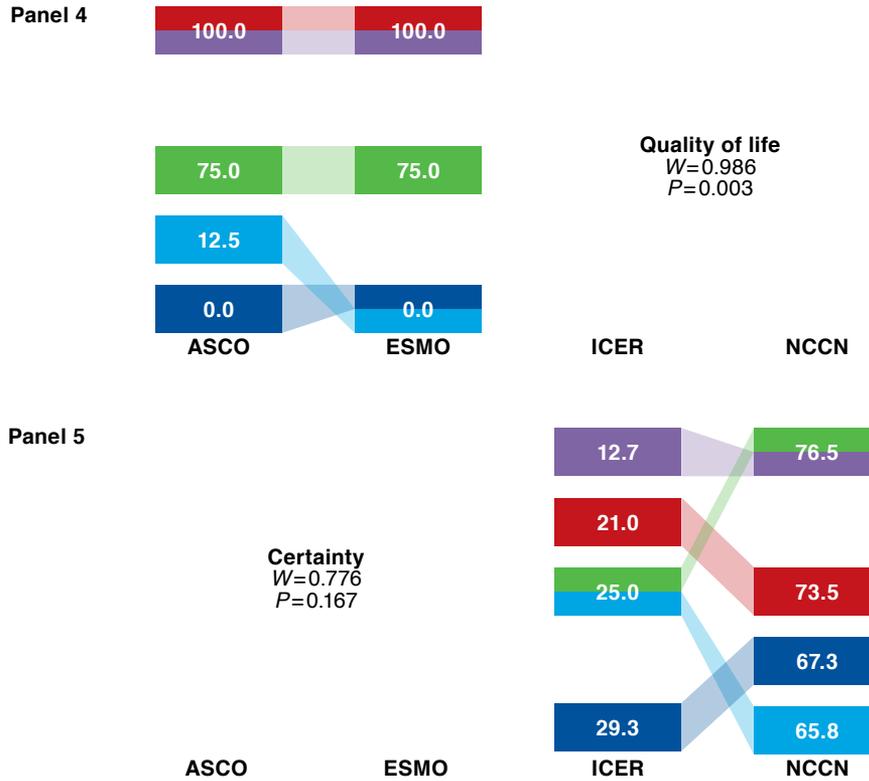
This study expands on a previously published pilot study²² by adding panelist assessments using the NCCN framework and including 10 more drugs (5 each in advanced breast cancer and prostate cancer). Including the 5 advanced lung cancer drugs assessed in the pilot using the ASCO, ESMO, and ICER frameworks, the current study evaluates a total of 15 drugs using 4 frameworks. In the pilot, drugs were ranked similarly by the 4 frameworks, with Kendall's *W* of 0.703 ($P=0.006$); interrater reliability was high for the ASCO and ESMO frameworks and poor for ICER. In the current, larger study, convergent validity was assessed by cancer type to accurately represent real-world drug-to-drug comparisons.

FIGURE 3 Kendall's *W* Among 4 Frameworks Overall, Pairwise, and by Subdomain: Rankings and Rescaled Scores of 5 Prostate Cancer Drugs



continued on next page

FIGURE 3 Kendall's *W* Among 4 Frameworks Overall, Pairwise, and by Subdomain: Rankings and Rescaled Scores of 5 Prostate Cancer Drugs (continued)



Note: Columns show drug rankings for each framework and rescaled mean scores (range 0-100). In Panel 1, Kendall's *W* is shown as a measure of concordance across all frameworks and each pairwise comparison. In Panels 2-5, it is shown for each subdomain. In the certainty subdomain for ICER (Panel 5), lower scores represent higher rankings.

With the inclusion of panelist assessments using the NCCN framework, convergence among the lung cancer drugs decreased from good in the pilot to fair ($W=0.562$, $P=0.010$). Pairwise analyses indicated that poor convergence with NCCN drove this finding, with all NCCN pairs having fair-to-poor convergence (Figure 2). In both studies, reliability was excellent ($ICC \geq 0.75$) for ASCO and ESMO, likely due to the clarity of instructions and methodologic approach taken by these frameworks. On the other hand, reliability for ICER was poor ($ICC=0.281$, $95\% \text{ CI}=0.055\text{-}0.799$) in the pilot and fair ($ICC=0.652$, $95\% \text{ CI}=0.466\text{-}0.834$) in the larger study reported here. These differences may be driven by the impact of sample size (5 vs. 15 drugs assessed in the pilot and full study, respectively) on statistical results and on panelists' increasing familiarity with the framework as more drugs were assessed.

Providers, payers, and patients have identified as a significant problem the lack of a clear approach to ensuring adequate "value" of new oncology treatments.^{17,29} Are these frameworks

a practical solution? Some authorities have posited that physicians will rise to the challenge of discussing value with their patients.³⁰ However, panelists in our study were only somewhat enthusiastic about the process of conducting individual drug assessments, and the hour spent reviewing literature and conducting just 1 assessment may be more time than most oncologists are able to devote. ASCO promises to develop a user-modifiable software tool with populated drug data and user-modifiable category weights, ICER intends to publish more reports, and ESMO proposes to assess every newly approved anticancer drug. Currently, however, only NCCN has published completed assessments for an extensive array of oncology regimens, and busy clinicians may choose ease-of-use and availability over reliability.

There is evidence that payers are considering framework ratings in their decisions.³¹ A recent survey found 27% of payers responding report to using value frameworks to guide preferred therapy decisions, alter coverage criteria through policies such as prior authorization to make "lower value" medications more

TABLE 2 Interclass Correlation Coefficients by Reviewer Type and Subdomain^a

	ASCO		ESMO		ICER		NCCN	
All reviewers (n=8), ICC (95% CI)	0.800	(0.660-0.913)	0.818	(0.686-0.921)	0.652	(0.466-0.834)	0.153	(0.045-0.371)
Oncologists vs. non-oncologists								
Oncologists (n = 4)	0.807	(0.638-0.920)	0.842	(0.699-0.936)	0.769	(0.582-0.903)	0.210	(0.020-0.501)
Other (n = 4)	0.786	(0.605-0.911)	0.816	(0.655-0.924)	0.603	(0.353-0.817)	0.156	(0 ^b -0.427)
Physicians vs. nonphysicians								
Physicians (n = 6)	0.825	(0.686-0.926)	0.831	(0.698-0.929)	0.641	(0.439-0.830)	0.156	(0.031-0.395)
Other (n = 2)	0.740	(0.375-0.905)	0.691	(0.302-0.884)	0.482	(0.023-0.784)	0.198	(0 ^b -0.597)
By subdomain								
Clinical benefit	0.829	(0.704-0.927)	0.809	(0.673-0.917)	-	(-)	0.149	(0.041-0.368)
Toxicity	0.755	(0.592-0.891)	0.597	(0.406-0.800)	-	(-)	0.194	(0.067-0.432)
Quality of life	0.671	(0.490-0.844)	0.818	(0.686-0.921)	-	(-)	-	(-)
Certainty	-	(-)	-	(-)	0.062	(0 ^b -0.247)	0.022	(0 ^b -0.129)

^aICC and CI are shown as measure of framework reliability.

^bNegative ICC estimate was observed, which suggested that the true ICC is very low; ICC of zero was therefore assumed.³⁴

ASCO=American Society of Clinical Oncology; CI=confidence interval; ESMO=European Society for Medical Oncology; ICC=interclass correlation coefficient; ICER=Institute for Clinical and Economic Review.

difficult to obtain, and educate providers on high- versus low-value therapies (e.g., through pathways programs). Another 41% of payers said they plan to use such tools in the future.

Can use of any of these frameworks help align the prices of cancer drugs with their value or improve patient access to high-value treatments? In their current form, the frameworks we evaluated focus much more on drug attributes other than cost, if they include cost at all. Cost is included in only some of the frameworks evaluated here, and it is considered in different ways even when included.¹⁶⁻¹⁷ The lack of transparent cost data and the variability in insurance coverage in the U.S. health care system means that identifying the correct cost inputs for a given patient is a colossal, if not an impossible, task. It may be that broader approaches to value-based reimbursement, such as the proposed Medicare Part B program changes, may have a greater impact on cancer drug pricing than value frameworks, such as those analyzed here.³²

Regardless of which framework is used, or who uses it, basing decisions on “value” will ultimately affect patients. Framework users should recognize that some highly valued patient outcomes—such as health-related quality of life, ease of use, subgroup differences, and long-term side effects—are incorporated in the frameworks unevenly. Even in frameworks that consider these outcomes, they can only be incorporated in each drug’s assessment to the extent that data are available, and such data are frequently lacking.³³

Although we did not evaluate the frameworks’ construct validity—the extent to which they actually measure the latent variable, “value”—our results indicate that the ASCO, ESMO, ICER, and NCCN frameworks provide value assessments that are fairly similar to one another. Defining and measuring value in oncology care is complex, and it is a remarkable testament to the efforts of the developers that these divergent approaches

are similar at all. Despite these broadly similar results, the differing approaches are most clearly visible in reproducibility of panelists’ assessments. The limited instructions and lower discriminatory ability of the NCCN framework meant its reliability was poor. Modifications of the instructions, provision of specific anchors, or other changes should be considered if the NCCN framework is to be used by individuals, rather than committees.

Limitations

Results of our analysis must be considered in light of its potential limitations. The use of rankings, rather than raw scores, was required for analyzing convergent validity with the Kendall’s W statistic. Because each framework produces scores on different scales, raw scores cannot be directly compared between frameworks. We addressed this in our analysis to allow for descriptive comparisons between frameworks by rescaling scores to 1-100. Further, since not all drugs assessed in this analysis are direct substitutes, this analysis may have been an easier test than a real-world comparison. Future iterations will incorporate larger subgroups of directly substitutable drugs to evaluate the impact that this incorporation may have on validity and reliability.

With panelists conducting multiple assessments among drugs and frameworks, contamination and training effects may occur. For example, as individual panelists conduct multiple assessments for the same drug, earlier assessments may impact subsequent scores. Indeed, it took panelists less time to complete the assessments for the framework they used last. With more assessments completed, convergence may improve as panelists get better at using the frameworks. Completing more assessments could falsely increase convergence for drugs assessed later in the study.

The set of drugs (prostate cancer drugs) that were evaluated last by all panelists did in fact have the best convergent validity. Although we posited that this result was due to the simplicity of trial data for these drugs, the impact of assessment sequence in our analysis is unknown. Future analyses should incorporate different sequencing of value assessments by drug and framework to evaluate the potential role of these factors.

Of the 4 frameworks considered here, only the ASCO and ESMO frameworks produce single clinical effectiveness scores. Published assessments with the full ICER value framework report multiple components, including clinical evidence reviews, cost-effectiveness models, and budget impact analyses. Since results from such multifaceted analyses could not be used to evaluate convergent validity and interrater reliability, the ICER comparative clinical effectiveness tool was used for our analysis. The NCCN Evidence Blocks also do not yield a single score for clinical effectiveness, instead producing separate scores for efficacy, safety, quality of evidence, and consistency of evidence. For our analysis, these scores were averaged to create 1 summary score. When this assumption was tested in sensitivity analyses, W results did not change substantially for any cancers.

Conclusions

Value assessment frameworks are an attempt to shift the discussion among providers, payers, and patients alike from price to value. Although the different approaches built into these frameworks suggest that stakeholders have yet to come to agreement on exactly how to define value, our findings indicate they may be closer than may have been expected. The ASCO, ESMO, ICER, and NCCN frameworks demonstrated fair-to-excellent convergent validity and appropriately focus on clinical efficacy. Interrater reliability was good or excellent for all frameworks except NCCN's, whose simpler approach made it user-friendly but more susceptible to small differences between users. Continued use, analyses, and refinements of these frameworks will bring us closer to the ultimate goal of using value-based treatment decisions to improve patient care and outcomes.

Authors

TANYA G.K. BENTLEY, PhD; JACKIE LEE, BS; JENELLE M. ZAMBRANO, DNP, CNS, RN; and MICHAEL S. BRODER, MD, MSHS, Partnership for Health Analytic Research, Beverly Hills, California. JOSHUA T. COHEN, PhD, Tufts Medical Center, Boston, Massachusetts; ELENA B. ELKIN, PhD, Memorial Sloan Kettering Cancer Center, New York, New York; JULIE HUYNH, MD, Hematology Oncology of San Fernando Valley, Encino, California; ARNAB MUKHERJEA, DrPH, MPH, Health Sciences Program, California State University, East Bay, Hayward, California; THANH H. NEVILLE, MD, MSHS, and IOANA POPESCU, MD, MPH, Department of Medicine, David Geffen School of Medicine, University of California, Los Angeles; and MATTHEW MEI, MD, City of Hope National Medical Center, Duarte, California. RONDA COPHER, PhD, and RUSSELL KNOTH, PhD, Eisai Inc., Woodcliff Lake, New Jersey.

AUTHOR CORRESPONDENCE: Tanya G.K. Bentley, PhD, Partnership for Health Analytic Research, 280 S. Beverly Dr., Ste. 404, Beverly Hills, CA 90212. Tel.: 310.858.9555; E-mail: tbentley@pharllc.com.

DISCLOSURES

This work was funded by Eisai Inc. Copher and Knoth are employees of Eisai Inc. Bentley, Lee, Zambrano, and Broder are employees of Partnership for Health Analytic Research, a health services research company paid by Eisai Inc. to conduct this research. For this study, Cohen, Huynh, and Neville report fees from Partnership for Health Analytic Research. Outside of this study, Cohen receives grants and direct consulting fees from various companies that manufacture and market pharmaceuticals. Mei reports a grant from Eisai Inc. during this study. The other authors have no disclosures to report.

Study concept and design were contributed by Bentley and Broder, with assistance from Elkin and Cohen. Bentley took the lead in data collection, along with Elkin, Huynh, Mukherjea, Neville, Mei, Popescu, Lee, and Zambrano. Data interpretation was performed by Bentley and Broder, along with Elkin, Cohen, Copher, and Knoth. The manuscript was written primarily by Bentley, along with Elkin and Broder, and revised by Bentley, Broder, Elkin, Cohen, Copher, and Knoth.

Select components of this work's methods were presented at ISPOR 19th Annual European Congress held in Vienna, Austria, October 29-November 2, 2016, and Society for Medical Decision Making 38th Annual North American Meeting held in Vancouver, Canada, October 23-26, 2016.

REFERENCES

1. Goldstein DA, Chen Q, Howard DH, Lipscomb J, Ramalingam SS, Flowers CR. Nectinumab in metastatic squamous cell lung cancer: establishing a value-based cost. *JAMA Oncol.* 2015;1(9):1293-300.
2. Kesselheim AS, Avorn J, Sarpatwari A. The high cost of prescription drugs in the United States: origins and prospects for reform. *JAMA.* 2016;316(8):858-71.
3. Ramsey SD, Lyman GH, Bangs R. Addressing skyrocketing cancer drug prices comes with tradeoffs: pick your poison. *JAMA Oncol.* 2016;2(4):425-26.
4. Salas-Vega S, Mossialos E. Cancer drugs provide positive value in nine countries, but the United States lags in health gains per dollar spent. *Health Aff (Millwood).* 2016;35(5):813-23.
5. Yu PP. Challenges in measuring cost and value in oncology: making it personal. *Value Health.* 2016;19:520-24.

6. 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(23):2563-77.
7. 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(24):2925-34.
8. National Comprehensive Cancer Network. NCCN Evidence Blocks frequently asked questions. 2016. Available at: <https://www.nccn.org/evidence-blocks/pdf/EvidenceBlocksFAQ.pdf>. Accessed April 5, 2017.
9. National Comprehensive Cancer Network. NCCN Evidence Blocks user guide. Available at: <https://www.nccn.org/evidenceblocks/pdf/EvidenceBlocksUserGuide.pdf>. Accessed April 5, 2017.
10. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology (NCCN Guidelines) with NCCN Evidence Blocks. Available at: <https://www.nccn.org/evidenceblocks/>. Accessed April 5, 2017.
11. 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(8):1547-73.
12. Ollendorf D, Chapman R, Khan S, et al. Treatment options for relapsed or refractory multiple myeloma: effectiveness, value, and value-based price benchmarks. Final evidence report. Institute for Clinical and Economic Review. May 5, 2016. Available at: http://icer-review.org/wp-content/uploads/2016/05/MWCEPAC_MM_Evidence_Report_050516-002.pdf. Accessed April 5, 2017.
13. Ollendorf D, Pearson SD. ICER evidence rating matrix: a user's guide. 2016. Available at: <http://icer-review.org/wp-content/uploads/2016/01/Rating-Matrix-User-Guide-FINAL-v10-22-13.pdf>. Accessed April 5, 2017.
14. Institute for Clinical and Economic Review. Addressing the myths about ICER and value assessment. 2016. Available at: <https://icer-review.org/blog/myths-icer-and-value-assessment/>. Accessed April 5, 2017.
15. Basch E. Toward a patient-centered value framework in oncology. *JAMA*. 2016;315(19):2073-74.
16. Chandra A, Shafrin J, Dhawan R. Utility of cancer value frameworks for patients, payers, and physicians. *JAMA*. 2016;315(19):2069-70.
17. Dalzell M. Considerations for designing "value calculators" for oncology therapies. *Am J Manag Care*. Published online May 12, 2016. Available at: <http://www.ajmc.com/journals/evidence-based-oncology/2016/peer-exchange-oncology-stakeholders-summit/considerations-for-designing-value-calculators-for-oncology-therapies>. Accessed April 5, 2017.
18. Dangi-Garimella S. Lessons to learn from the NICE cancer care model. *Am J Manag Care*. Published online July 15, 2016. Available at: <http://www.ajmc.com/journals/evidence-based-oncology/2016/july-2016/lessons-to-learn-from-the-nice-cancer-care-model/P-1>. Accessed April 5, 2017.
19. Feinberg B, Lal L, Swint M. Is there a mathematical resolution to the cost-versus-value debate? *Am J Manag Care*. 2015;21:SP542-44. Available at: <http://www.ajmc.com/journals/evidence-based-oncology/2015/december-2015/is-there-a-mathematical-resolution-to-the-cost-versus-value-debate>. Accessed April 5, 2017.
20. Goulart BH. Value: the next frontier in cancer care. *Oncologist*. 2016;21(6):651-63.
21. Neumann PJ, Cohen JT. Measuring the value of prescription drugs. *N Engl J Med*. 2015;373(27):2595-97.
22. Bentley TKG, Cohen JT, Elkin EB, et al. Validity and reliability of value assessment frameworks for new cancer drugs. *Value Health*. 2017;20(2):200-05.
23. Wilson L, Lin T, Wang L, et al. Evaluation of the ASCO Value Framework for anticancer drugs at an academic medical center. *J Manag Care Spec Pharm*. 2017;23(2):163-69. Available at: <http://www.jmcp.org/doi/10.18553/jmcp.2017.23.2.163>.
24. Young RC. Value-based cancer care. *N Engl J Med*. 2015;373(27):2593-95.
25. American Cancer Society. Cancer facts & figures 2017. Available at: <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html>. Accessed April 5, 2017.
26. Song G-Y, Zhao Z-H, Ding Y, et al. Reliability assessment and correlation analysis of evaluating orthodontic treatment outcome in Chinese patients. *Int J Oral Sci*. 2014;6(1):50-55.
27. Thomason ME, Dennis EL, Joshi AA, et al. Resting-state fMRI can reliably map neural networks in children. *NeuroImage*. 2011;55(1):165-75.
28. Kiesewetter B, Raderer M, Steger GG, et al. The European Society for Medical Oncology Magnitude of Clinical Benefit Scale in daily practice: a single institution, real-life experience at the Medical University of Vienna. *ESMO Open*. 2016;1(4):e000066.
29. National Health Council. The patient voice in value: the National Health Council patient-centered value model rubric. 2016. Available at: <http://www.nationalhealthcouncil.org/sites/default/files/Value-Rubric.pdf>. Accessed April 5, 2017.
30. Blayney DW. Where does dynamic value assessment fit into our role as agents advising our patients with cancer? *J Oncol Pract*. 2016;12(12):1211-13.
31. Shafer J, Rademacher K, Serluco J. The rising use of oncology value tools by payers. *Journal of Clinical Pathways*. 2016;2(6). Published online only. Available at: <http://www.journalofclinicalpathways.com/article/rising-use-oncology-value-tools-payers>. Accessed April 5, 2017.
32. Cohen JP. Frameworks for assessing the value of cancer drugs: purely an academic exercise? *Journal of Clinical Pathways*. 2016;2(6):29-33. Available at: <http://www.journalofclinicalpathways.com/article/frameworks-assessing-value-cancer-drugs-purely-academic-exercise>. Accessed April 5, 2017.
33. Johnson P, Greiner W, Al-Dakkak I, Wagner S. Which metrics are appropriate to describe the value of new cancer therapies? *Biomed Res Int*. 2015;2015:865101. Available at: <https://www.hindawi.com/journals/bmri/2015/865101/>. Accessed April 5, 2017.
34. Taylor PJ. An introduction to intraclass correlation that resolves some common confusions. Faculty paper. Programs in Science, Technology & Values, Critical & Creative Thinking, and Public Policy. University of Massachusetts, Boston, MA. 2010. Available at: <http://www.faculty.umb.edu/pjt/09b.pdf>. Accessed April 5, 2017.
35. Geyer CE, Forster J, Lindquist D, et al. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. *N Engl J Med*. 2006;355(26):2733-43. Erratum in: *N Engl J Med*. 2007;356(14):1487.
36. Zhou X, Cella D, Cameron D, et al. Lapatinib plus capecitabine versus capecitabine alone for HER2+ (ErbB2+) metastatic breast cancer: quality-of-life assessment. *Breast Cancer Res Treat*. 2009;117(3):577-89.
37. Cortés J, O'Shaughnessy J, Loesch D, et al. Eribulin monotherapy versus treatment of physician's choice in patients with metastatic breast cancer (EMBRACE): a phase 3 open-label randomized study. *Lancet*. 2011;377(9769):914-23.
38. Baselga J, Campone M, Piccart M, et al. Everolimus in postmenopausal hormone-receptor-positive advanced breast cancer. *N Engl J Med*. 2012;366(6):520-29.
39. Burris HA 3rd, Lebrun F, Rugo HS, et al. Health-related quality of life of patients with advanced breast cancer treated with everolimus plus exemestane versus placebo plus exemestane in the phase 3, randomized, controlled, BOLERO-2 trial. *Cancer*. 2013;119(10):1908-15.
40. Verma S, Miles D, Gianni L, et al. Trastuzumab emtansine for HER2-positive advanced breast cancer. *N Engl J Med*. 2012;367(19):1783-91.
41. Welslau M, Diéras V, Sohn JH, et al. Patient-reported outcomes from EMILIA, a randomized phase 3 study of trastuzumab emtansine (T-DM1) versus capecitabine and lapatinib in human epidermal growth factor receptor 2-positive locally advanced or metastatic breast cancer. *Cancer*. 2014;120(5):642-51.

42. Cortés J, Baselga J, Im YH, et al. Health-related quality-of-life assessment in CLEOPATRA, a phase III study combining pertuzumab with trastuzumab and docetaxel in metastatic breast cancer. *Ann Oncol*. 2013;24(10):2630-35.
43. Swain SM, Baselga J, Kim SB, et al. Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer. *N Engl J Med*. 2015;372(8):724-34.
44. Solomon BJ, Mok T, Kim DW, et al. First-line crizotinib versus chemotherapy in ALK-positive lung cancer. *N Engl J Med*. 2014;371(23):2167-77. Erratum in: *N Engl J Med*. 2015;373(16):1582.
45. Rosell R, Carcereny E, Gervais R, et al. Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *Lancet Oncol*. 2012;13(3):239-46.
46. Brahmer J, Reckamp KL, Baas P, et al. Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. *N Engl J Med*. 2015;373(2):123-35.
47. Herbst RS, Baas P, Kim DW, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet*. 2016;387(10027):1540-50.
48. Belani CP, Brodowicz T, Ciuleanu TE, et al. Quality of life in patients with advanced non-small-cell lung cancer given maintenance treatment with pemetrexed versus placebo (H3E-MC-JMEN): results from a randomised, double-blind, phase 3 study. *Lancet Oncol*. 2012;13(3):292-99.
49. Ciuleanu T, Brodowicz T, Zielinski C, et al. Maintenance pemetrexed plus best supportive care versus placebo plus best supportive care for non-small-cell lung cancer: a randomised, double-blind, phase 3 study. *Lancet*. 2009;374(9699):1432-40.
50. De Bono JS, Logothetis CJ, Molina A, et al. Abiraterone and increased survival in metastatic prostate cancer. *N Engl J Med*. 2011;364(21):1995-2005.
51. Harland S, Staffurth J, Molina A, et al. Effect of abiraterone acetate treatment on the quality of life of patients with metastatic castration-resistant prostate cancer after failure of docetaxel chemotherapy. *Eur J Cancer*. 2013;49(17):3648-57.
52. De Bono JS, Oudard S, Ozguroglu M, et al. Prednisone plus cabazitaxel or mitoxantrone for metastatic castration-resistant prostate cancer progressing after docetaxel treatment: a randomised open-label trial. *Lancet*. 2010;376(9747):1147-54.
53. Berthold DR, Pond GR, Roessner M, et al. Treatment of hormone-refractory prostate cancer with docetaxel or mitoxantrone: relationships between prostate-specific antigen, pain, and quality of life response and survival in the TAX-327 study. *Clin Cancer Res*. 2008;14(9):2763-67.
54. 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(15):1502-12.
55. Fizazi K, Scher HI, Miller K, et al. Effect of enzalutamide on time to first skeletal-related event, pain, and quality of life in men with castration-resistant prostate cancer: results from the randomised, phase 3 AFFIRM trial. *Lancet Oncol*. 2014;15(10):1147-56.
56. Scher HI, Fizazi K, Saad F, et al. Increased survival with enzalutamide in prostate cancer after chemotherapy. *N Engl J Med*. 2012;367(13):1187-97.

APPENDIX Drugs and Corresponding Data Used for Assessment

Study Drugs	Data
Breast	
Capecitabine + lapatinib	Geyer et al., 2006 ³⁵ Zhou et al., 2009 ³⁶
Eribulin	Cortés et al., 2011 ³⁷
Exemestane + everolimus	Baselga et al., 2012 ³⁸ Burris et al., 2013 ³⁹
T-DM1	Verma et al., 2012 ⁴⁰ Welslau et al., 2014 ⁴¹
Trastuzumab + chemotherapy + pertuzumab	Cortés et al., 2013 ⁴² Swain et al., 2015 ⁴³
Lung	
Crizotinib	Solomon et al., 2014 ⁴⁴
Erlotinib	Rosell et al., 2012 ⁴⁵
Nivolumab	Brahmer et al., 2015 ⁴⁶
Pembrolizumab	Herbst et al., 2016 ⁴⁷
Pemetrexed	Belani et al., 2012 ⁴⁸ Ciuleanu et al., 2009 ⁴⁹
Prostate	
Abiraterone + prednisone	De Bono et al., 2011 ⁵⁰ Harland et al., 2013 ⁵¹
Cabazitaxel + prednisone	De Bono et al., 2010 ⁵²
Docetaxel (Q7 or Q21) prednisone	Berthold et al., 2008 ⁵³ Tannock et al., 2004 ⁵⁴
Enzalutamide	Fizazi et al., 2014 ⁵⁵ Scher et al., 2012 ⁵⁶
Mitoxantrone + prednisone	De Bono et al., 2010 ⁵²