PRM 167

MEASURING THE VALIDITY AND RELIABILITY OF VALUE ASSESSMENT FRAMEWORKS FOR CANCER DRUGS: AN EVALUATION METHOD

Tanya G.K. Bentley, PhD[1], Joshua T. Cohen, PhD[2], Elena B. Elkin, PhD[3], Julie Huynh, MD[4], Arnab Mukherjea, DrPH, MPH[5], Thanh H. Neville, MD, MSHS[6], Matthew Mei, MD[7], Ronda Copher, PhD[8], Russell L. Knoth, PhD[8], Ioana Popescu, MD, MPH[9], Jackie Lee, BS[1], Jenelle M. Zambrano, DNP, CNS, RN[1], & Michael S. Broder, MD, MSHS[1]

1Partnership for Health Analytic Research, LLC, Beverly Hills, CA, USA, 2Center for the Evaluation of Value and Risk in Health, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA, 3Memorial Sloan-Kettering Cancer Center, New York, CA, USA, 4Hematology Oncology Medical Group of San Fernando Valley, Encino, CA, USA, 5California State University, East Bay, Hayward, CA, USA, 6David Geffen School of Medicine at UCLA, Department of Medicine, Los Angeles, CA, USA, 7City of Hope National Medical Center, Duarte, CA, USA, 8Eisai Inc., Woodcliff Lake, NJ, USA

BACKGROUND

- Several organizations, including 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 frameworks to assess the value of oncology drugs.
- We previously developed a methodology for evaluating the validity and reliability of value assessments and applied it in a pilot study.

OBJECTIVE

- This study aimed to evaluate our methodology's applicability for assessing a broader array of drugs and frameworks.

METHODS

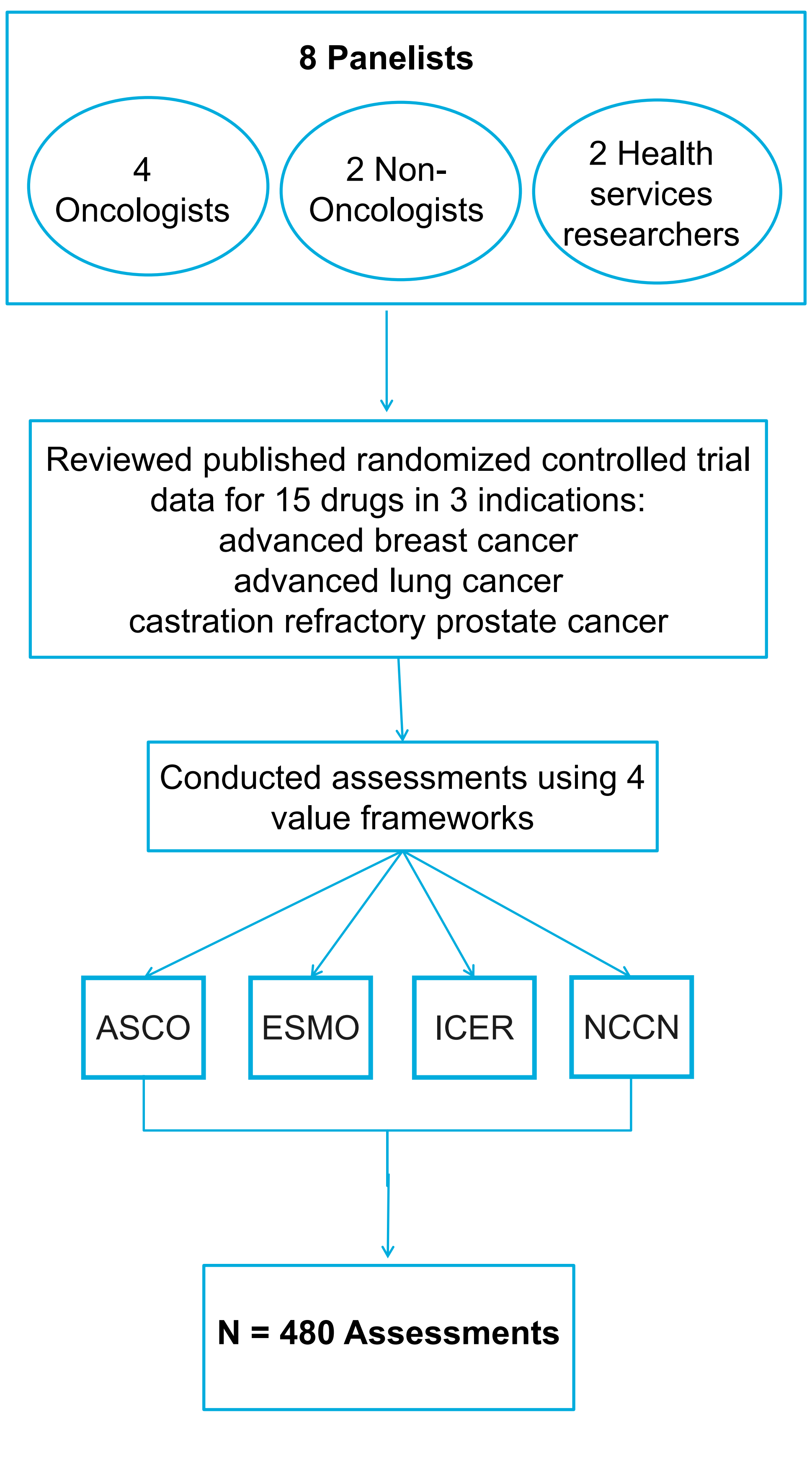
Overview

- Our method is based on two primary outcomes:
  1. Convergent validity: how correlated drug rankings are across frameworks.
     - We chose Kendall's coefficient of concordance for ranks (Kendall's *W*) as the correlation measure:
       a. We first calculated mean scores for each drug overall and by subdomain (clinical benefit, toxicity, quality of life, and certainty).
       b. We then ranked mean scores of each of the 15 drugs in 3 indications within each framework from highest to lowest.
       c. We calculated Kendall's *W* by comparing ranked mean drug scores among the frameworks.
     - Kendall's *W* ranges from 0 (no agreement) to 1 (complete agreement).
     - We used *p* values to test the alternative hypothesis of complete agreement ($W > 0$) against null hypothesis.
     - Means were re-scaled to 0-100 for easy comparisons.
  2. Inter-rater reliability: a measure of how stable framework value estimates are across users.
     - We chose intra-class correlations coefficients (ICC) with 95% confidence intervals (CI) as the statistical measure.
     - We calculated ICC separately for each framework, overall and by subdomain, assuming that the 8 reviewers represented a random sample from a larger population of reviewers.

Application

- We applied the method to 5 drugs for each of the 3 indications (total of 15 drugs):
  - Advanced breast cancer
  - Advanced non-small cell lung cancer
  - Castration refractory prostate cancer
- Eight panelists assessed the drugs: 4 oncologists, 2 non-oncologist clinicians, 2 health services researchers
- Each assessment produced a single numeric or ordinal outcome ("score")

Figure 1. Study Design



8 Panelists

4 Oncologists

2 Non-Oncologists

2 Health services researchers

Reviewed published randomized controlled trial data for 15 drugs in 3 indications:
advanced breast cancer
advanced lung cancer
castration refractory prostate cancer

Conducted assessments using 4 value frameworks

ASCO

ESMO

ICER

NCCN

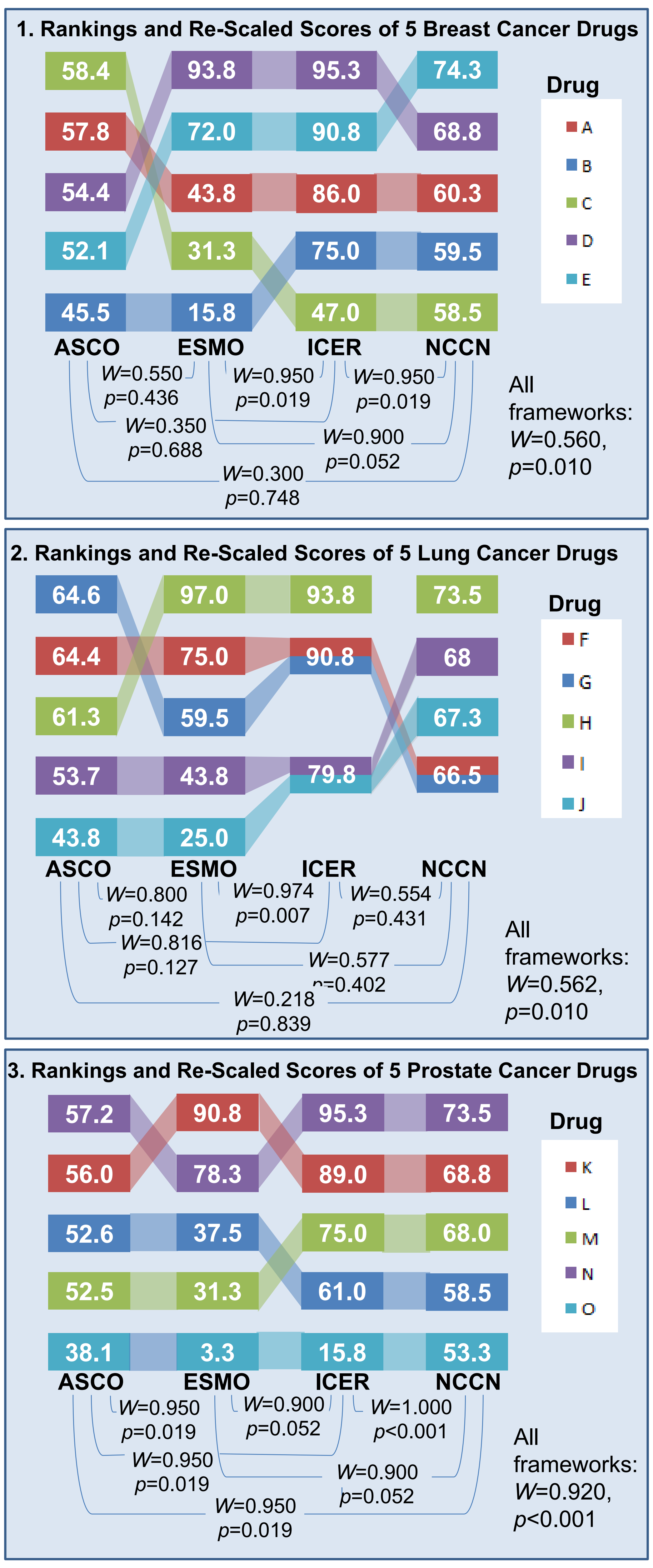
N = 480 Assessments

- Panelists responded to a survey after completing the value assessments to:
  - Rate the frameworks
  - Comment on their experiences

RESULTS

- The 8 panelists successfully completed a total of 480 assessments (4 frameworks * 8 panelists * 15 drugs)
- Results appear in **Figure 2** (convergent validity) and the **Table 1** (reliability).

Figure 2. Overall Ranking of Re-Scaled Scores of 15 Cancer Drugs in 3 Indications using 4 Frameworks



1. Rankings and Re-Scaled Scores of 5 Breast Cancer Drugs

2. Rankings and Re-Scaled Scores of 5 Lung Cancer Drugs

3. Rankings and Re-Scaled Scores of 5 Prostate Cancer Drugs

Columns show drug rankings for each framework and re-scaled mean scores (range: 0-100). In each panel, Kendall's *W* is shown as a measure of concordance across all frameworks and each pairwise comparison.

- Specifically:
  - Frameworks produced scores on different scales, so raw scores cannot be directly compared.
  - When re-scaled from 0 (worst) to 100 (best), score ranges varied across frameworks.
    - NCCN scores spanned the narrowest range.
    - ESMO scores spanned the broadest range.
  - Convergent validity among frameworks was fair to excellent, increasing with the clinical benefit subdomain concordance and simplicity of drug trial data.
- Overall convergent validity was excellent (≥0.75) only when also excellent among:
  - All pairwise comparisons
  - Clinical benefit subdomain scores (even when convergence among toxicity scores was poor, <0.40)
- For example:
  - Clinical benefit concordance was poor (<0.40) or fair (0.40-0.59) for the sets of drugs with fair overall concordance, despite good or excellent concordance among the toxicity and quality of life subdomains
  - Clinical benefit concordance was excellent (≥0.75) among the drugs with excellent overall concordance, despite poor toxicity concordance.

Table 1. ICC AND 95% CI by Panelist Type and Subdomain[a]

ICC (95% CI)				
	ASCO	ESMO	ICER	NCCN
All reviewers (n=8)	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)
Non-oncologists (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. Non-physicians				
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)
Non-physicians (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				
Certainty	n/a	n/a	0.053 (0[b] - 0.588)	0.022 (0[b] - 0.129)
Clinical Benefit	0.829 (0.704-0.927)	0.809 (0.673-0.917)	n/a	0.149 (0.041-0.368)
Quality of Life	0.671 (0.490-0.844)	0.818 (0.686-0.921)	n/a	n/a
Toxicity	0.755 (0.592-0.891)	0.597 (0.406-0.800)	n/a	0.194 (0.067-0.432)

n/a: subdomain is not a distinct component of the framework.

[a] ICC and CI shown as measures of framework reliability.

[b] Negative ICC estimate was observed, which suggested that the true ICC is very low; therefore, ICC of zero was assumed.

Panelists' survey results

- Assessment timing **(Table 2)**:

Table 2. Mean Panelists' Literature Review and Assessment Completion Times

	Mean time	
	Literature review for each drug assessed	**Completion of each assessment**
ASCO	28 minutes	25 minutes
ESMO	22 minutes	14 minutes
ICER	25 minutes	21 minutes
NCCN[a]	11 minutes	8 minutes

[a] Assessments were conducted last among all panelists

- No single framework emerged as:
  - easiest to use;
  - having highest global panelist rating (e.g., comfort with using framework to assess treatment for a loved one).

CONCLUSIONS

- This method allows quantitative analyses of value assessment frameworks' validity and reliability.
- When applied to 15 oncology drugs in 3 indications, this method successfully allowed us to draw conclusions about the convergent validity and inter-rater reliability of 4 value frameworks.
  - The frameworks demonstrated fair-to-excellent convergent validity, and appropriately focused on clinical efficacy.
  - Overall concordance was strongly influenced by concordance among clinical efficacy scores.
  - All frameworks except NCCN demonstrated good-to-excellent reliability.
- Mean scores produced by a committee will be more reliable than those produced by an individual.
- This method allowed us to identify key drivers of concordance and reliability between frameworks.
- When 2 frameworks produced similar clinical benefit scores, the overall scores were generally more concordant. Clinical benefit score primarily reflects efficacy, which is probably an important driver of clinical decision-making. Thus framework scores may reflect those made in clinical practice.
- Assessments were found to be time consuming, so their usefulness in practice may be enhanced with the release of more committee-based assessments from framework developers.
- Use of this method to determine how drugs may be valued by different frameworks will facilitate payer, provider, and patient decision-making.

Presented at ISPOR, May 20-24, 2017, Boston, Massachusetts

Research conducted by Partnership for Health Analytic Research, LLC.

Acknowledgements
This study was sponsored by Eisai, Inc.