A Novel Method for Evaluating Value Assessment Frameworks

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[6], Jenelle M. Zambrano, DNP, CNS, RN[1], Jackie Lee, BS[1], Eunice Chang, PhD[1], Michael S. Broder, MD, MSHS[1]

[1]Partnership for Health Analytic Research, LLC, Beverly Hills, CA; [2]Tufts Medical Center, Boston, MA; [3]Memorial Sloan-Kettering Cancer Center, NY; [4]Hematology Oncology of San Fernando Valley, Encino, CA; [5]California State University, East Bay, Hayward, CA; [6]David Geffen School of Medicine at UCLA, Department of Medicine, Los Angeles, CA; [7]City of Hope National Medical Center, Duarte, CA; [8]Eisai Inc., Woodcliff Lake, NJ

BACKGROUND

- Various frameworks have been developed to assess the value of oncology drugs.
- Organizations who have developed frameworks include:
  - American Society of Clinical Oncology (ASCO)
  - European Society for Medical Oncology (ESMO)
  - Institute for Clinical and Economic Review (ICER)
  - National Comprehensive Center Network (NCCN)
- Despite their common goals, it is unclear whether the frameworks actually provide valid and reliable measurements of value and how to assess such validity and reliability in practice.

OBJECTIVE

- We developed a methodology for evaluating the validity and reliability of value assessment frameworks.

METHODS

Overview

- We calculated convergent validity, defined as the correlation among drug rankings across frameworks.
  - Kendall's coefficient of concordance for ranks (Kendall's W) was chosen as the statistical measure.
    1. Calculated mean scores for each drug.
    2. Ranked mean scores of each of the 5 drugs within each framework from highest to lowest.
    3. Compared rankings among the frameworks.
    4. Kendall's W ranges from 0 (no agreement) to 1 (complete agreement). P values tested alternative hypothesis of complete agreement ($W > 0$) against null hypothesis.
    5. Means were re-scaled to 0-100 for easy comparisons.
- We used inter-rater reliability as a measure of how stable frameworks' estimates of value are across users.
  - Intraclass correlation coefficients (ICC) with 95% confidence intervals (CI) were chosen as the statistical measure.
  - ICC was calculated separately for each framework.
  - ICC calculations were done assuming the 8 reviewers represent a random sample from a larger population of reviewers.

Application

- We applied the method to 5 drugs for advanced non-small cell lung cancer.
- Each assessment produced a single numeric or ordinal outcome (in aggregate the "panelist scores").
  - Used along with NCCN's published assessments ("published scores") to evaluate convergent validity across 4 frameworks.

Figure 1. Study Design
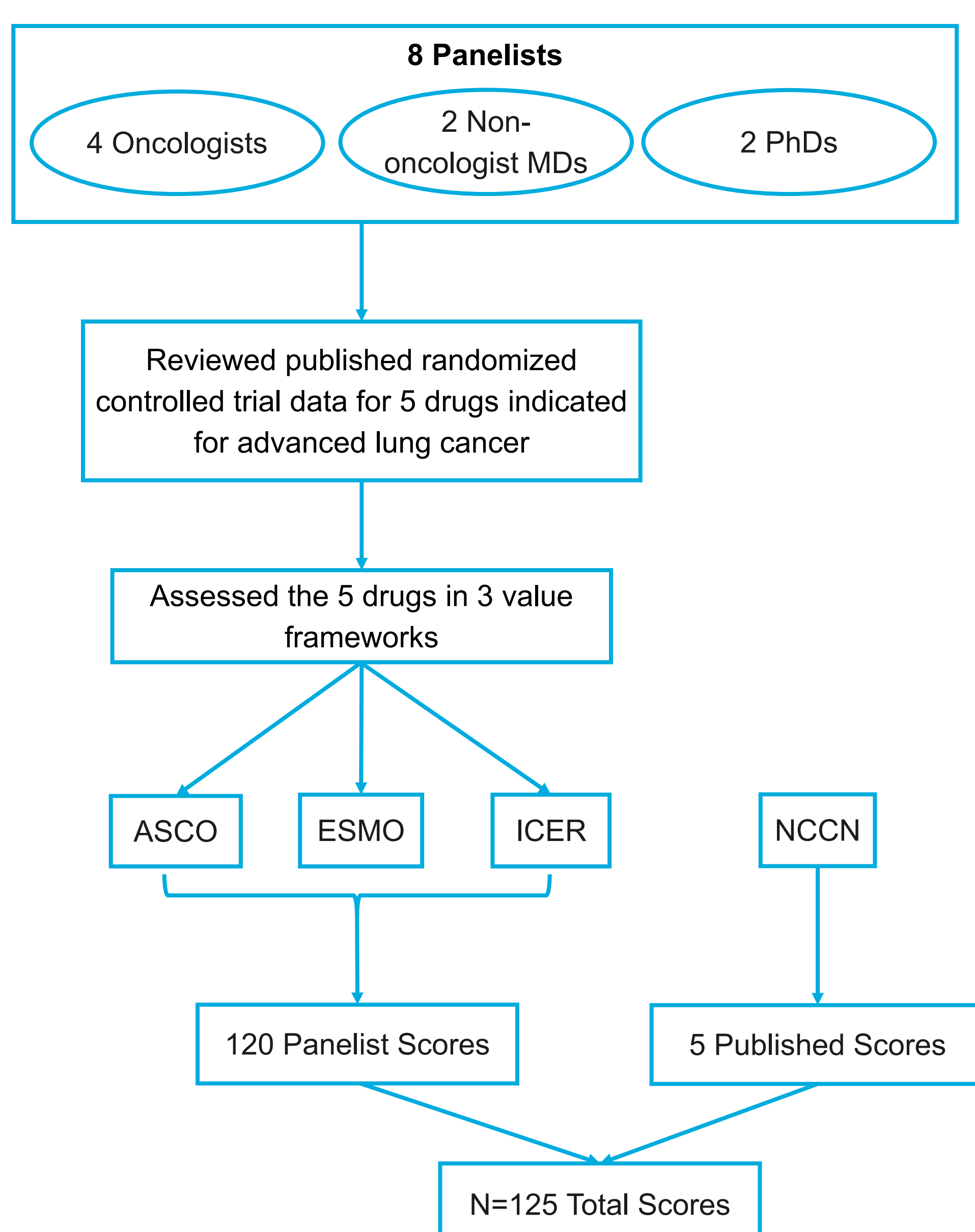
RESULTS

- Panelists successfully completed all value assessments for 5 selected drugs.
- Results of application are shown in Figure 2 (validity) and in the Table (reliability).
- Specifically:
  - Raw scores are on different scales and cannot be compared.
  - When re-scaled from 0 (worst) to 100 (best), score ranges varied among frameworks.
  - ASCO and ESMO had wider ranges: 31 and 72 points, respectively.
  - ICER and NCCN had much narrower ranges: 14 and 19 points, respectively.
    - ASCO: 16-47
    - ESMO: 25-97
    - ICER: 80-94
    - NCCN: 75-94
  - ASCO scores were the lowest, and NCCN scores were highest.
  - Kendall's W=0.703

Figure 2. Ranking of Re-Scaled Scores of 5 Lung Cancer Drugs using 4 Frameworks: Overall and by Subdomain
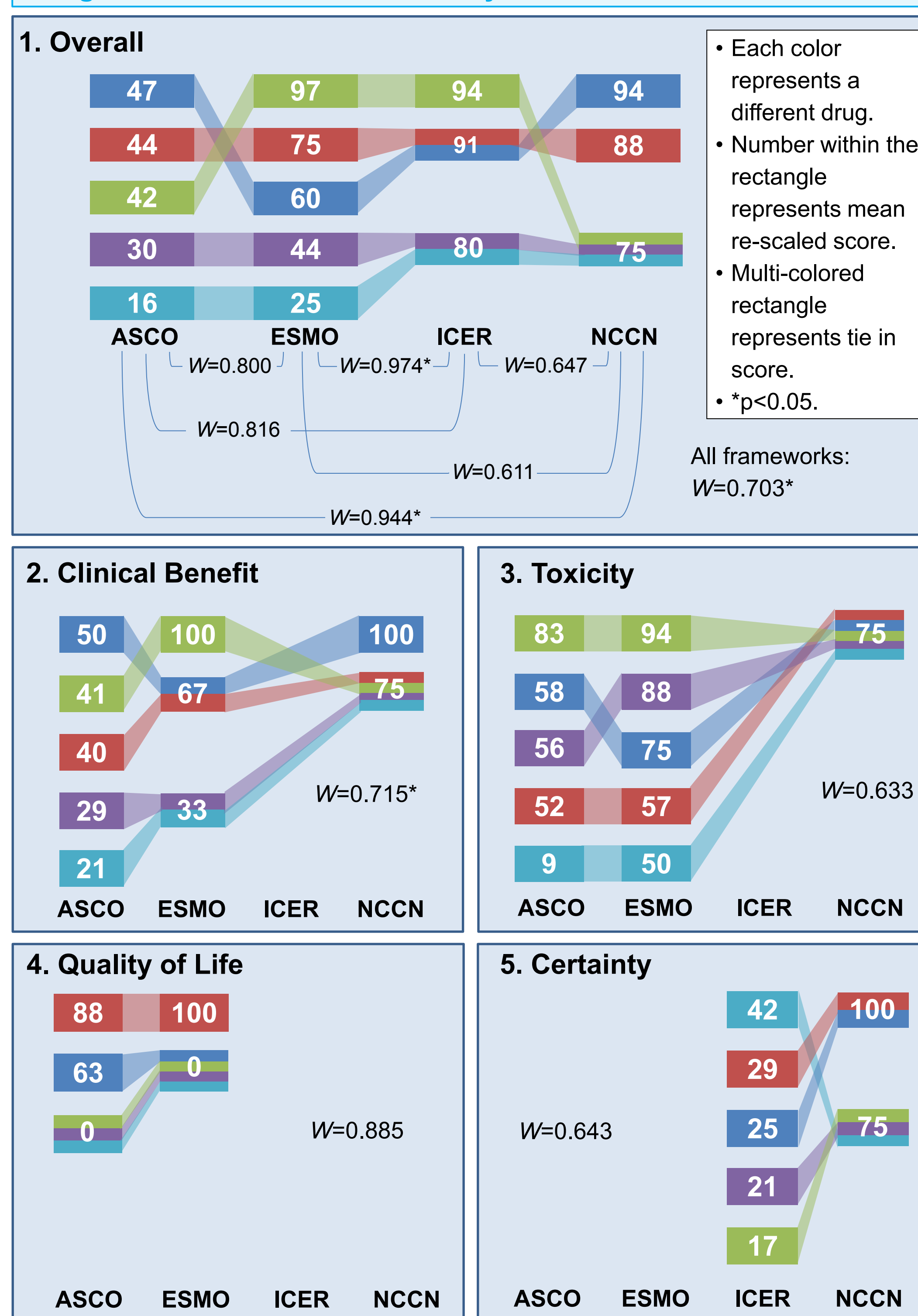
Columns represent each framework. Mean scores range from 0 to 100. In panel 1, Kendall's W is shown as a measure of concordance across all frameworks and each pairwise comparison. In panels 2-5 it is shown for each subdomain. Subdomain scores not shown are not distinct components of the framework.

Table. ICC (95% CI), Overall and by Panelist Type and Subdomain[a]

ICC (95% CI)	ASCO	ESMO	ICER
All reviewers (n=8)	0.786 (0.517 - 0.970)	0.804 (0.545 - 0.973)	0.281 (0.055 - 0.799)
Oncologists vs. Non-oncologists			
Oncologists (n=4)	0.835 (0.526 - 0.979)	0.843 (0.520 - 0.980)	0.120 (0[b] - 0.759)
Non-oncologists (n=4)	0.716 (0.331 - 0.959)	0.806 (0.477 - 0.974)	0.368 (0.029 - 0.861)
Physicians vs. Non-physicians			
Physicians (n=6)	0.855 (0.618 - 0.981)	0.793 (0.507 - 0.971)	0.228 (0[b] - 0.776)
Non-physicians (n=2)	0.562 (0[b] - 0.938)	0.769 (0[b] - 0.973)	0.222 (0[b] - 0.839)
By Subdomain			
Certainty	n/a	n/a	0.053 (0[b] - 0.588)
Clinical Benefit	0.692 (0.383 - 0.952)	0.857 (0.643 - 0.981)	n/a
Quality of Life	0.681 (0.372 - 0.950)	1.000 (n/a[c] - n/a[c])	n/a
Toxicity	0.825 (0.584 - 0.976)	0.468 (0.172 - 0.890)	n/a

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, suggesting that the true ICC is very low; therefore, ICC of zero was assumed.
[c] All reviewers had the same scores for each drug.

Panelists' Survey Results

- Panelists' mean time to complete each assessment:
  1. ASCO and ICER: ~30 minutes
  2. ESMO: 15 minutes
- Mean time to review literature for each drug for conducting assessments: 20-30 minutes.
- ESMO instructions were the clearest.
- ASCO was rated most logically organized.
- No single frameworks emerged as:
  - Easiest to use
  - Having highest global panelist rating (e.g., comfort with using framework to assess treatment for a loved one).

- Panelists were given a survey after completing the value assessments.
  - Rated different frameworks
  - Provided comments regarding their experiences.

CONCLUSIONS

- This method is the first to allow quantitative analyses of value assessment frameworks' validity and reliability.
- When applied to 5 oncology drugs, this method successfully allowed us to draw conclusions about the convergent validity and inter-rater reliability of 4 value frameworks.
  - Frameworks ranked similarly, indicating convergent validity.
  - Overall, reliability was quite good.
  - Reliability was better among oncologists and physicians for ASCO and ESMO, but not ICER.
  - Individuals who want to conduct their own value assessments in oncology (rather than use a published value) should choose either ASCO or ESMO, because these two frameworks demonstrated high validity and reliability.
  - Mean scores produced by a committee will be more reliable than those produced by an individual.
- Further exploration of differences among panelists will provide a better understanding of how to interpret value assessments produced by these frameworks in clinical practice.
- Although the approach can be used to determine the reproducibility of value assessments produced by these frameworks, reproducibility is only one component of an overall assessment of the frameworks' contribution to value-based decision-making. Importantly, the method presented here fails to measure frameworks' construct validity — the extent to which they actually measure the latent variable, "value." The true test of this will be how they influence decisions made by clinicians and patients when used in clinical practice settings.