# There but for GRACE? A Validated Screening Tool for Quality Observational Studies of Comparative Effectiveness

Nancy A. Dreyer, PhD, MPH[1]; Priscilla Velentgas, PhD[1]; Kimberly Westrich, MA[2]; and Robert Dubois, MD, PhD[2]
[1]Quintiles Outcome, Cambridge, MA, USA; [2]National Pharmaceutical Council, Washington, DC, USA

## Background

While there is growing demand for information about comparative effectiveness (CE), there is little understanding about whether non-interventional studies can be good enough for decision support.

## Objective

To develop and validate an item checklist that can be used to qualify non-interventional comparative effectiveness studies that are sufficiently rigorous in design and execution for decision support.

## Methods

An 11-item checklist was developed through literature review and consultation with experts from ISPOR, ISPE, payer groups, private sector, and academia. Since no gold standard exists for validation, three approaches were used to rate 88 articles about drugs, medical devices, and medical procedures for testing:

1) Quality assessments from published systematic reviews

2) Opinions of 9 recognized experts

3) Concordant assessments of articles from 5 experts

113 volunteers from five continents conducted a total of 280 assessments of 88 articles. Positive predictive values (PPV) and negative predictive values (NPV) of individual items compared testers' assessments to those of experts.

## Results

• Concordance of expert opinion was 52%. Most checklist questions performed better.

• The data questions showed better NPV than the methods questions, whereas the methods questions showed better PPV than data questions.

  - Ten of the 11 questions showed potential for NPV using a cut-point of 0.67.

  - Nine of the 11 questions showed some potential for their PPV.

• The single question that consistently performed best addressed the validity of the primary outcome(s) [D4].

## Conclusion

The single best performing criterion for distinguishing a study that is reasonably fit-for-purpose is to determine whether the primary outcome has sufficient validity for the decision at hand. Further enhancements to quality can be achieved by using epidemiologic methods to minimize the effects of bias. Study reports that rate relatively well on this screening require further examination to understand applicability, effect size, and bias. The absence of a broadly acceptable gold standard makes it difficult to validate which items cleanly distinguish high-quality non-interventional CE research.

### Predictive Values by Item for All Validation Test Sets

| Validation Test Sets By Sample | Adequate Intervention/ treatment | Adequate Outcomes | Objective Outcomes | Valid Outcomes | Similar Outcomes | Covariates recorded | New Initiators | Concurrent comparators | Covariates accounted for | Immortal Time Bias | Sensitivity Analysis |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | D1 | D2 | D3 | D4 | D5 | D6 | M1 | M2 | M3 | M4 | M5 |
| **POSITIVE PREDICTIVE VALUES** | | | | | | | | | | | |
| **Systematic Review 1** | | | | | | | | | | | |
| PPV | 0.40 | 0.49 | 0.54 | 0.59 | 0.47 | 0.60 | 0.38 | 0.46 | 0.69 | --† | 0.75 |
| N/D | 10/25 | 20/41 | 19/35 | 16/37 | 20/43 | 12/20 | 6/16 | 18/39 | 11/16 | --† | 9/12 |
| **Systematic Review 2** | | | | | | | | | | | |
| PPV | 0.48 | 0.48 | 0.58 | 0.52 | 0.47 | 0.75 | 0.33 | 0.33 | 0.91 | --† | 0.7 |
| N/D | 15/31 | 19/40 | 19/33 | 17/33 | 20/43 | 9/12 | 7/21 | 7/21 | 10/11 | --† | 7/10 |
| **Single Expert Review 1** | | | | | | | | | | | |
| PPV | 0.57 | 0.63 | 0.56 | 0.70 | 0.58 | 0.73 | 0.56 | 0.59 | 0.65 | 0.63 | 0.56 |
| N/D | 16/28 | 19/40 | 18/32 | 16/23 | 22/38 | 11/15 | 14/25 | 23/39 | 13/20 | 12/19 | 9/16 |
| **Single Expert Review 2** | | | | | | | | | | | |
| PPV | 0.68 | 0.62 | 0.59 | 0.74 | 0.58 | 0.64 | 0.65 | 0.58 | 0.68 | 0.55 | 0.69 |
| N/D | 5/8 | 16/26 | 18/29 | 14/19 | 19/33 | 9/14 | 13/20 | 0/1 | 13/19 | 11/20 | 11/16 |
| **Concordant Expert Review 1** | | | | | | | | | | | |
| PPV | 0.67 | 0.71 | 0.63 | 0.83 | 0.44 | 1.00 | 0.44 | 0.67 | 0.50 | 1.00 | 1.00 |
| N/D | 4/6 | 5/7 | 5/8 | 5/6 | 4/9 | 1/1 | 4/9 | 6/9 | 2/4 | 4/4 | 2/2 |
| **Concordant Expert Review 2** | | | | | | | | | | | |
| PPV | 0.63 | 0.60 | 0.55 | 0.67 | 0.55 | 0.50 | 0.67 | 0.55 | 0.50 | 0.50 | 0.50 |
| N/D | 5/8 | 6/10 | 6/11 | 4/6 | 6/11 | 2/4 | 6/9 | 6/11 | 2/4 | 1/2 | 1/2 |
| **No. rated ≥ 0.7**\*\* | 2 | 1 | 0 | 4 | 0 | 3 | 2 | 1 | 4 | 1 | 4 |
| **NEGATIVE PREDICTIVE VALUES** | | | | | | | | | | | |
| **Systematic Review 1** | | | | | | | | | | | |
| NPV | 0.60 | 1.00 | 1.00 | 0.73 | 1.00 | 0.67 | 0.46 | 0.80 | 0.65 | --† | 0.71 |
| N/D | 9/15 | 2/2 | 3/3 | 8/11 | 1/1 | 16/24 | 13/28 | 4/5 | 17/26 | --† | 24/34 |
| **Systematic Review 2** | | | | | | | | | | | |
| NPV | 0.67 | 0.83 | 0.83 | 0.78 | 1.00 | 0.66 | 0.45 | 0.45 | 0.70 | --† | 0.61 |
| N/D | 8/12 | 5/6 | 5/6 | 7/9 | 1/1 | 19/29 | 10/22 | 10/22 | 23/33 | --† | 22/36 |
| **Single Expert Review 1** | | | | | | | | | | | |
| NPV | 0.50 | -- | 0.50 | 1.00 | -- | 0.58 | 0.38 | -- | 0.46 | 0.00 | 0.38 |
| N/D | 2/4 | 0/0 | 1/2 | 6/6 | 0/0 | 11/19 | 3/8 | 0/0 | 6.13 | 0/1 | 8/21 |
| **Single Expert Review 2** | | | | | | | | | | | |
| NPV | 0.80 | 0.25 | 0.50 | 0.82 | 1.00 | 0.44 | 0.50 | 0.00 | 0.47 | 0.33 | 0.50 |
| N/D | 4/5 | 1/4 | 2/4 | 9/11 | 1/1 | 8/18 | 6/12 | 0/1 | 7/15 | 1/3 | 10/20 |
| **Concordant Expert Review 1** | | | | | | | | | | | |
| NPV | 0.50 | -- | 1.00 | 1.00 | -- | 0.56 | 0.33 | 1.00 | 0.43 | -- | 0.56 |
| N/D | 1/2 | 0/0 | 1/1 | 2/2 | 0/0 | 5/9 | 1/3 | 1/1 | 3/7 | 0/0 | 5/9 |
| **Concordant Expert Review 2** | | | | | | | | | | | |
| NPV | 1.00 | -- | -- | 0.67 | -- | 0.50 | 1.00 | 1.00 | 0.43 | 0.50 | 0.50 |
| N/D | 1/1 | 0/0 | 0/0 | 2/3 | 0/0 | 3/6 | 2/2 | 1/1 | 3/7 | 1/2 | 5/10 |
| **No. rated ≥ 0.7**\*\* | 3 | 2 | 3 | 6 | 3 | 2 | 1 | 3 | 2 | 0 | 1 |

Legend: N = numerator (number of articles that raters and experts agreed on quality); D = denominator (total number of articles rated on quality by raters); \* No raters responded "insufficient" to these questions; † Question not included in systematic review; \*\* values ending in 0.5 were rounded up

## GRACE Checklist Questions

### Data

**D1** Were treatment and/or important details of treatment exposure adequately recorded for the study purpose in the data source(s)? *Note: not all details of treatment are required for all research questions.*

**D2** Were the primary outcomes adequately recorded for the study purpose (e.g., available in sufficient detail through data source(s))?

**D3** Was the primary *clinical* outcome(s) measured objectively rather than subject to clinical judgment (e.g., opinion about whether the patient's condition has improved)?

**D4** Were primary outcomes validated, adjudicated, or otherwise known to be valid in a similar population?

**D5** Was the primary outcome(s) measured or identified in an equivalent manner between the treatment/intervention group and the comparison group(s)?

**D6** Were important covariates that may be known confounders or effect modifiers available and recorded? *Important covariates depend on the treatment and/or outcome of interest, (e.g., body mass index should be available and recorded for studies of diabetes; race should be available and recorded for studies of hypertension and glaucoma).*

### Methods

**M1** Was the study (or analysis) population restricted to new initiators of treatment or those starting a new course of treatment? *Efforts to include only new initiators may include restricting the cohort to those who had a washout period (specified period of medication nonuse) prior to the beginning of study follow-up.*

**M2** If one or more comparison groups were used, were they concurrent comparators? If not, did the authors justify the use of historical comparisons group(s)?

**M3** Were important covariates, confounding and effect modifying variables taken into account in the design and/or analysis? *Appropriate methods to take these variables into account may include: restriction, stratification, interaction terms, multivariate analysis, propensity score matching, instrumental variables or other approaches.*

**M4** Is the classification of exposed and unexposed person-time free of "immortal time bias"? *Immortal time in epidemiology refers to a period of cohort follow-up time during which death (or an outcome that determines end of follow-up) cannot occur.*

**M5** Were any meaningful analyses conducted to test key assumptions on which primary results are based? *E.g., were some analyses reported to evaluate the potential for a biased assessment of exposure or outcome, such as analyses where the impact of varying exposure and/or outcome definitions was tested to examine the impact on results.*

**See www.graceprinciples.org for full checklist with response options**