Optimal Data Analysis
Vol. 4 (April 21, 2015), 21-23

2155-0182/10/$3.00

Estimating Inter-Rater Reliability Using Pooled Data Induces Paradoxical Confounding: An Example Involving *Emergency Severity Index* Triage Ratings

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

The inter-rater reliability of patient triage ratings made using the Emergency Severity Index is computed and compared for pooled data within and across studies *versus* separately for pairs of independent raters. Findings reveal that results for the pooled findings exhibit confounding attributable to Simpson's Paradox. These findings raise the concern that all estimates of inter-rater (and parallel-forms) reliability based on pooled data reported for *all instruments discussed in the literature* are susceptible to paradoxical confounding, and likely result in overestimation of rating reliability.

A recent study conducted a meta-analysis of psychometric properties of patient triage scores assigned using the Emergency Severity Index (ESI).[1] Five studies included in the meta-analysis provided a pooled inter-rater reliability table indicating overall agreement between all unique pairs of raters in the study. In the meta-analysis these data were integrated into an inter-rater reliability table summarizing the agreement between all unique pairs of raters in all five studies. The pooled data indicated 35/44 (79.5%) ratings of triage code 1 were consistent; as were 730/868 (84.1%) triage code 2 ratings; 610/795 (76.7%) triage code 3 ratings; 587/767 (76.5%) triage code 4 ratings; and 489/646 (75.7%) triage code 5 ratings. For these data the *a priori* hypothesis that ratings between pairs of raters were consistent (i.e., that data fell along the major diagonal of the inter-rater table) was evaluated via maximum-accuracy[2] ("optimal") analysis performed using UniODA.[3,4] Analysis revealed relatively strong inter-rater agreement: ESS=73.2 (indicating 73.2% of the theoretical possible improvement *vs*. chance[3]), $p<0.0001$.

One of the five studies included in the pooled inter-rater reliability table involved all expert raters, and reported the highest weighted kappa for ratings on the ESI in the literature.[1,4-6] Optimal statistical analysis of the data in this latter study indicated that the *a priori* hypothesis was only tenable for 4 (40%) of the total of 10 unique inter-rater pairs: for the remaining 6 inter-rater pairs exploratory non-linear models reflecting various manifestations and magni-

2155-0182/10/$3.00

tudes of disagreement were identified.[4] For these ten pairs of raters the ESS values obtained ranged between 59.9 (relatively strong agreement) and 28.6 (moderate agreement). Findings obtained for the pooled ratings thus clearly indicate confounding attributable to Simpson's Paradox.[7-10]

These findings raise the concern that all inter-rater (and parallel-forms[11]) reliability estimates based on pooled data reported for *all instruments discussed in the literature* are susceptible to paradoxical confounding, and likely serve to over-estimate the empirical reliability.

References

[1]Mirhaghi A, Heydari A, Mazlom R, Hasanzadeh F (2015). Reliability of the Emergency Severity Index: Meta-Analysis. *Sultan Qaboos University Medical Journal*, *15*, e71-77.

[2]Yarnold PR (2014). "A statistical guide for the ethically perplexed" (Chapter 4, Panter & Sterba, *Handbook of Ethics in Quantitative Methodology*, Routledge, 2011): Clarifying disorientation regarding the etiology and meaning of the term *Optimal* as used in the Optimal Data Analysis (ODA) paradigm. *Optimal Data Analysis*, *3*, 30-31. URL: http://odajournal.com/2014/04/06/a-statistical-guide-for-the-ethically-perplexed-chapter-4-panter-sterba-handbook-of-ethics-in-quantitative-methodology-routledge-2011-clarifying-disorientation-regarding/

[3]Yarnold PR, Soltysik RC (2005). *Optimal data analysis: A guidebook with software for Windows*, Washington, DC, APA Books.

[4]Yarnold PR (2014). How to assess inter-observer reliability of ratings made on ordinal scales: Evaluating and comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis*, *3*, 42-49. URL: http://odajournal.com/2014/04/13/how-to-assess-inter-observer-reliability-of-ratings-made-on-ordinal-scales-evaluating-and-comparing-the-emergency-severity-index-version-3-and-canadian-triage-acuity-scale/

[5]Yarnold PR (2014). UniODA *vs*. kappa: Evaluating the long-term (27-year) test-retest reliability of the Type A Behavior Pattern. *Optimal Data Analysis*, *3*, 14-16. URL: http://odajournal.com/2014/03/29/unioda-vs-kappa-evaluating-the-long-term-27-year-test-retest-reliability-of-the-type-a-behavior-pattern/

[6]Yarnold PR (2014). UniODA *vs*. weighted kappa: Evaluating concordance of clinician and patient ratings of the patient's physical and mental health functioning. *Optimal Data Analysis*, *3*, 12-13. URL: http://odajournal.com/2014/03/28/unioda-vs-weighted-kappa-evaluating-concordance-of-clinician-and-patient-ratings-of-the-patients-physical-and-mental-health-functioning/

[7]Yarnold PR (1996). Characterizing and circumventing Simpson's paradox for ordered bivariate data. *Educational and Psychological Measurement*, *56*, 430-442. DOI: 10.1177/0013164496056003005

[8]Bryant FB, Siegel EKB (2010). Junk science, test validity, and the Uniform Guidelines for Personnel Selection Procedures: The case of *Melendez v*. Illinois Bell. *Optimal Data Analysis*, *1*, 176-198. URL: http://odajournal.com/2013/09/19/junk-science-test-validity-and-the-uniform-guidelines-for-personnel-selection-procedures-the-case-of-melendez-v-illinois-bell/

[9]Soltysik RC, Yarnold PR (2010). The use of unconfounded climatic data improves atmospheric prediction. *Optimal Data Analysis*, *1*, 67-100. URL: http://odajournal.com/2013/09/19/the-use-of-unconfounded-climatic-data-improves-atmospheric-prediction/

2155-0182/10/$3.00

[10]Yarnold PR (2013). Ascertaining an individual patient's *symptom dominance hierarchy*: Analysis of raw longitudinal data induces Simpson's Paradox. *Optimal Data Analysis*, *2*, 159-171. URL: http://odajournal.com/2013/11/07/ascertaining-an-individual-patients-symptom-dominance-hierarchy-analysis-of-raw-longitudinal-data-induces-simpsons-paradox/

[11]Yarnold, P.R. (2014). How to assess the inter-method (parallel-forms) reliability of ratings made on ordinal scales: Evaluating and comparing the Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale. *Optimal Data Analysis*, *3*, 50-54. URL: http://odajournal.com/2014/04/14/how-to-assess-the-inter-method-parallel-forms-reliability-of-ratings-made-on-ordinal-scales-emergency-severity-index-version-3-and-canadian-triage-acuity-scale/

Author Notes

This study involved secondary data analysis of published de-identified data and was exempt from Institutional Review Board review.

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646
USA