Optimal Data Analysis
Vol. 5 (May 31, 2016), 37-40

2155-0182/10/$3.00

Causality of Adverse Drug Reactions: The Upper-Bound of Arbitrated Expert Agreement for Ratings Obtained by WHO and Naranjo Algorithms

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

As a high-ranking cause of human mortality, adverse drug reactions (ADRs) are the focus of an enormous literature, and optimal statistical methods have proven undaunted by the analysis-challenging geometry of multi-site longitudinal medical data sets.[1-6] Two broadly-used causality assessment algorithms for identifying ADRs are the Naranjo and World Health Organization (WHO) ADR algorithms.[7] Ratings made using these algorithms haven't been validated, so the extent to which arbitrated ratings made by independent experts using these algorithms agree is important in assessing the expected upper-bound of inter-rater, inter-method reliability. Using data from India[7] on this issue, UniODA identified a strong to very strong relationship between ratings obtained using WHO and Naranjo algorithms for a sample of N = 200 randomly selected patients. Inter-algorithm disagreement occurred for 15.2% of cases indicated as "Probable" by the Naranjo algorithm, but as "Possible" by the WHO algorithm.

Separately for each observation in a randomly selected sample of N = 200 ADR proformas, two experts independently used the WHO and Naranjo algorithms to assess the likelihood that a causal relationship existed between taking a medication and then experiencing an adverse reaction.[7] After arbitrating all disagreements the final structure of inter-expert agreement is summarized in Table 1. All 134 cases classified as "Probable" by the WHO algorithm were also classified as "Probable" by the Naranjo algorithm. And, 42 (65.6%) of 64 cases classified as "Possible" by the WHO algorithm were likewise classified using Naranjo.

Table 1: Arbitrated Expert Agreement on WHO and Naranjo Ratings of Causality of ADR[7]

	Naranjo	
WHO	Probable	Possible
Probable	134	0
Possible	24	42

2155-0182/10/$3.00

The original study[7] used a weighted kappa coefficient to assess the inter-algorithm agreement for these data. Here, weighted kappa = 0.701, which the authors concluded indicates moderate to good agreement between the two expert raters using the two algorithms and arbitrating all of the inter-rater disagreements. Statistical significance of the kappa coefficient wasn't discussed. Problems associated with the use of kappa to assess inter-rater reliability are discussed elsewhere.[1,2,8,9]

UniODA[1] and MegaODA[10-12] command syntax given below[13-15] was used to evaluate the *a priori* hypothesis that the algorithms agreed on the causality ratings assigned to the cases:

```
OUTPUT adr.out;
OPEN adr.dat;
VARS naranjo who;
CLASS naranjo;
ATTR who;
DIR < 0 1;
MC ITER 25000;
GO;
```

The attribute wasn't indicated as being categorical because the underlying metric of the scale for both algorithms is ordinal (probable is more likely than possible). The DIR syntax sets the *a priori* hypothesis that "possible" scores (coded as 0 for both algorithms) on the WHO scale are associated with "possible" scores on the Naranjo scale, and that "probable" scores (coded as 1 for both algorithms) on the WHO scale are associated with "probable" scores on the Naranjo scale.[1,2] The UniODA model was: if WHO rating = 0 then predict that the Naranjo rating = 0; otherwise if WHO rating = 1 then predict that Naranjo rating = 1. The confusion matrix summarizing classification performance of this model is given in Table 2: ESS = 84.8 corresponding to a very strong effect, exact $p < 0.0001$. There is thus very strong, statistically significant agreement using WHO ratings to predict Naranjo ratings.

Table 2: Confusion Table for UniODA Model Predicting Naranjo Ratings using WHO Ratings

	Predicted Rating		
Actual Rating	Possible	Probable	Sensitivity
Possible	42	0	100%
Probable	24	134	84.8%

Switching the class and attribute assignments within the earlier UniODA/MegaODA syntax (CLASS who; ATTR naranjo;) produces a UniODA model (if Naranjo rating = 0 predict WHO rating = 0; if Naranjo rating = 1 predict WHO rating = 1) with ESS = 63.6, corresponding to a strong effect with exact $p < 0.0001$ (see Table 3). There is thus strong, statistically significant agreement using Naranjo ratings to predict WHO ratings.

Table 3: Confusion Table for UniODA Model Predicting WHO Ratings using Naranjo Ratings

	Predicted Rating		
Actual Rating	Possible	Probable	Sensitivity
Possible	42	24	63.6%
Probable	0	134	100%

Presently, treating the WHO rating as an attribute (independent variable), and the Naranjo rating as a class (dependent) variable, produced the most accurate model normed against chance. As seen in Table 1, Naranjo ratings had lower variability than WHO ratings, with 42 versus 64 "Possible" ratings, respectively. A design with a homogenous class variable and a heterogeneous attribute will generally yield superior accuracy compared to a design involving a heterogeneous class variable and a homogeneous attribute: the latter design, in fact, is prone to identification of degenerate models in which not all classes are included in the model.[1,2,16]

Further study of the factors inducing disagreement between these two algorithms is warranted. An initial analytic approach would entail recoding the data, setting the 24 cases of algorithm disagreement as class 1, and all the re-

2155-0182/10/$3.00

maining cases as class 0, and then conducting CTA to discriminate class 0 and 1 cases on the basis of various aspects of the algorithms, in particular including as potential attributes the aspects bearing directly on the decision-making regarding probable *vs*. possible classifications.

References

[1]Yarnold PR, Soltysik RC. *Optimal data analysis: Guidebook with software for Windows*. Washington, DC: APA Books, 2005.

[2]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books.

[3]Yarnold PR (2013). Initial use of hierarchically optimal classification tree analysis in medical research. *Optimal Data Analysis*, *2*, 7-18.

[4]Bennett CL, Raisch DW, Lyons EA, Nebeker JR, Samore MH, Feldman MD, McKoy JM, Carson KR, Kut V, Belknap SM, Trifilio SM, Schumock GT, Yarnold PR, Davidson CJ, Morse RE, Kuzel TM, Parada JP, Cournoyer D, West DP, Sartor O, Tallman MS (2005). Introducing RADAR: the Research on Adverse Drug events And Reports (RADAR) project. *JAMA*, *293*, 2131-2140.

[5]Nebeker JR, Yarnold PR, Soltysik RC, Sauer BC, Sims SA, Samore MH, Rupper RW, Swanson KM, Savitz LA, Shinogle J, Xu W (2007). Developing indicators of inpatient adverse drug events through non-linear analysis using administrative data. *Medical Care*, *45*, S81-S88.

[6]Lu K, Kessler SJ, Schultz R, Bian J, Chen B, Wu J, Noxon V, Rao GA, Leibnitz R, Restaino J, Maxwell W, Norris LB, Qureshi ZP, Martin L, Love BL, Bookstaver B, Sutton S, Fayad R, Jacob S, Georgantopoulos P, Sartor O, Yarnold PR, Huff D, Hrusheshky W, Raisch DW, Ablin R, Bennett CL (2014). Systematic approach to pharmacovigilance beyond the limits: The Southern Network on Adverse Reactions (SONAR) projects. *Advances in Pharmacoepidemiology & Drug Safety*. DOI: 10.4172/2167-1052.1000149

[7]Mittal N, Gupta MC (2015). Comparison of agreement and rational uses of the WHO and Naranjo adverse event causality assessment tools. *Journal of Pharmacology & Pharmacotherapeutics*, *6*, 91-93.

[8]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.

[9]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.

[10]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Separating the chaff. *Optimal Data Analysis*, *2*, 194-197.

[11]Soltysik RC, Yarnold PR (2013). MegaODA large sample and BIG DATA time trials: Harvesting the wheat. *Optimal Data Analysis*, *2*, 202-205.

[12]Yarnold PR, Soltysik RC (2013). MegaODA large sample and BIG DATA time trials: Maximum velocity analysis. *Optimal Data Analysis*, *2*, 220-221.

[13]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.

2155-0182/10/$3.00

[14]Yarnold PR (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.

[15]Yarnold PR (2015). Estimating inter-rater reliability using pooled data induces paradoxical confounding: An example involving Emergency Severity Index triage ratings. *Optimal Data Analysis, 4,* 21-23.

[16]Yarnold PR (2016). Maximizing overall percentage accuracy in classification: Discriminating study groups in the National Pressure Ulcer Long-Term Care Study (NPULS). *Optimal Data Analysis, 5,* 29-30.

Author Notes

The study analyzed de-individuated data and was exempt from Institutional Review Board review. No conflict of interest was reported.

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