Supplemental materials for: Diagnostic test score validation with a fallible criterion

Real data example

Because the Method of Bounds-Test Validation is novel to the present paper, an example is described here to show how the Method of Bounds-Test Validation can be used to supplement a Known Group Validation study. In this example, a hypothetical Known Group Validation study was previously conducted with a reference test assumed to be infallible. The validity of the reference test is known, and the researchers want to describe the extent to which the fallible nature of the reference test may have affected their results.

This example relates to the validity of test scores for the identification of the lateralization of temporal lobe epilepsy. The reference test used in the hypothetical Known Group Validation study was clinical interpretation of electroencephalography (EEG) ictal recordings as left ($R = 1$), right ($R = 0$), or ambiguous (excluded) as reported by the examining neurologist.

EEG is often used as the infallible reference test in Known Group Validation studies of temporal lobe epilepsy (e.g., Lancman et al., 2012; Locke et al., 2010). Despite that EEG is well recognized as fallible and that sensitivity and specificity estimates for EEG are available (e.g., Brodbeck, 2011), Known Group Validation cannot account for fallibility of the reference test, and so typical diagnostic test score validation studies will assume that the reference test is infallible out of necessity. In these cases, methods reviewed here including the Method of Bounds-Test Validation are most useful as the methods facilitate incorporating the available information about the fallibility of the reference test.

The test to be validated was the Auditory Immediate Index from the Wechsler Memory

Scale-III (WMS-III; The Psychological Corporation, 1997). Participants with a score on the WMS-III Auditory Immediate Index of less than 90.5 were classified as positive (left temporal lobe epilepsy; $X = 1$), and participants with scores greater than 90.5 were classified as negative (right temporal lobe epilepsy; $X = 0$; Wilde et al., 2001). The Wechsler Memory Scales are commonly used to evaluate patients with temporal lobe epilepsy (e.g., Hermann, Seidenberg, Schoenfeld, & Davies, 1997).

The cases were patients with left temporal lobe epilepsy, in a sample of patients with left ($C = 1$) and right temporal lobe epilepsy ($C = 0$). Temporal lobe epilepsy lateralization was determined with hippocampal volume measured with magnetic resonance imaging (Cook et al., 1992; Watson, Jack, & Cendes, 1997). This meant that sensitivity is the proportion of left temporal lobe epilepsy patients that are diagnosed as left on the Auditory Immediate Index, and nonspecificity is the proportion of right temporal lobe epilepsy patients that are diagnosed as left on the Auditory Immediate Index.

Method

Participants. Data for 136 patients was collected from consecutive patient records at [SUPPRESSED FOR REVIEW]. Of the 136 patients, EEG data was not available for 30 patients. These patients were assumed to be missing completely at random after consulting the neurologists caring for the patients. Of the patients with EEG data, 15 had ambiguous EEG readings that precluded diagnosis of lateralization.

The left temporal lobe epilepsy patients had a mean age of 37.9 years (*SD* = 9.9), and 50% of the patients were female. The right temporal lobe epilepsy patients had a mean age of 35.8 years (*SD* = 11.1), and 56% of the patients were female. The study was approved by the [SUPPRESSED FOR REVIEW] Research and Ethics Committee.

Data analysis. The reference test was validated in a Known Group Validation design against hippocampal volume (Cook et al., 1992; Watson et al., 1997). To replicate the typical diagnostic test validation study where two groups are formed from the two diagnosis categories of the reference test, in this case the EEG results, the prevalences of the left and the right EEG groups were calculated as the positive predictive value and one minus the negative predictive value for EEG, respectively (Altman & Bland, 1994b; Meehl & Rosen, 1955). Note that this procedure to estimate the prevalences within the two groups requires that the prevalence within the sample and the sensitivity and specificity of the reference test is already known.

Table 4

Group and test characteristics for the example Method of Bound-Test Validation study

	n	Pr(*C*=1)	Pr(*X*=1)	Pr(*X*=1|*C*=1)	Pr(*X*=1|*C*=0)
EEG left (*R* = 1)	41	.93	.76	.76	.67
EEG right (*R* = 0)	50	.12	.54	.67	.52
EEG Left + Right	91	.48	.64	.75	.53

Note. EEG = electroencephalogram diagnosis, n = sample size. Probabilities for EEG left are conditional on *R* = 1 and EEG right are conditional on *R* = 0. For example, the value under the heading of "Pr(*X*=1|*C*=1)" for the column "EEG left (*R* = 1)" is the value of Pr(X=1|*R*=1,*C*=1). The value under heading of "Pr(*C*=1)" for the column "EEG right (*R* = 0)" is the value of Pr(*C*=1|*R*=0).

Table 4 shows the group characteristics and the true sensitivity and nonspecificity values for the Auditory Immediate Index scores. The Mixed Group Validation and Neighborhood model assumptions were violated in the expected direction, satisfying the assumptions of Method of Bound-Test Validation.

Table 5

Estimated bounds for the example Method of Bounds-Test Validation study, and compared to true values

	Constraint					
	Defined	Tautology	MGV	Neighbor	MoB-TV	True
Pr(*X*=1|*C*=1)					.73 to .78	.75
Lower	0	.26	-	.73		
Upper	1	1.32	.78	-		
Pr(*X*=1|*C*=0)					.51 to .55	.53
Lower	0	.30	.51	-		
Upper	1	1.24	-	.55		

Note. *Defined* constraints given by Equations 18 and 19. *Tautology* constraints given by Equations 21 and 22. MGV = constraint based on Mixed Group Validation; *MGV* constraints given by Equations 6 and 7. Neighbor = constrained based on the Neighborhood model; *Neighbor* constraints given by Equations 11 and 12. Interval calculated by taking the maximum lower bound and minimum upper bound.

Table 5 shows the Method of Bounds-Test Validation estimated bounds for sensitivity and nonspecificity. The final obtained intervals for sensitivity and nonspecificity were reasonably narrow at .05 and .04 respectively, and the true values were contained within these bounds as expected.

Based on the standard errors for Mixed Group Validation (Equations 8 and 9) and the

Neighborhood model (Equations 14 and 15), the standard errors on the lower and upper limits of sensitivity were .06 and .07, respectively. The standard errors on the lower and upper limits of nonspecificity were .08 and .06, respectively. Subtracting two times the standard error from the lower limits and adding two times the standard errors to the upper limit gives an interval of .62-.92 for sensitivity, and .35-.68 for nonspecificity.

References

Altman, D. G., & Bland, J. M. (1994a). Diagnostic tests 1: Sensitivity and specificity. *British Medical Journal*, *308*, 1552.

Altman, D. G., & Bland, J. M. (1994b). Diagnostic tests 2: Predictive values. *British Medical Journal*, *309*, 102.

Brodbeck, V., Spinelli, L., Lascano, A. M., Wissmeier, M., Vargas, M. I., Vulliemoz, S., ... & Seeck, M. (2011). Electroencephalographic source imaging: a prospective study of 152 operated epileptic patients. *Brain*, *134*, 2887-2897.

Cook, M. J., Fish, D. R., Shorvon, S. D., Straughan, K., & Stevens, J. M. (1992). Hippocampal volumetric and morphometric studies in frontal and temporal lobe epilepsy. *Brain*, *115*, 1001-1015.

Hermann, B. P., Seidenberg, M., Schoenfeld, J., & Davies, K. (1997). Neuropsychological characteristics of the syndrome of mesial temporal lobe epilepsy. *Archives of Neurology*, *54*, 369-376.

Lancman, G., Vazquez-Casals, G. A., Perrine, K., Feoli, E., & Myers, L. (2012). Predictive value of Spanish neuropsychological testing for laterality in patients with epilepsy. *Epilepsy & Behavior*, *23*, 142-145.

Locke, D. E., Fakhoury, T. A., Berry, D. T., Locke, T. R., & Schmitt, F. A. (2010). Objective evaluation of personality and psychopathology in temporal lobe versus extratemporal lobe epilepsy. *Epilepsy & Behavior*, *17*, 172-177.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting score. *Psychological Bulletin*, *52*, 194-216.

The Psychological Corporation. (1997). *WAIS-III/WMS-III technical manual*. San Antonio, TX:

Author.

Watson, C., Jack, C. R., & Cendes, F. (1997). Volumetric magnetic resonance imaging: clinical applications and contributions to the understanding of temporal lobe epilepsy. *Archives of Neurology*, *54*, 1521-1531.

Wilde, N., Strauss, E., Chelune, G. J., Loring, D. W., Martin, R. C., Hermann, B. P., Sherman, E. M. S., & Hunter, M. (2001). WMS-III performance in patients with temporal lobe epilepsy: Group differences and individual classification. *Journal of the International Neuropsychological Society*, *7*, 881-891.