# Cognadev Technical Report Series

13

*18th October, 2018*

## The Binomial Effect Size Display (BESD)

**Is this always an accurate index of effect?**

I provide the definition, some warnings of the conditions under which it may not always produce accurate results, and some worked examples demonstrating those conditions. Overall, I think it's a pretty good 'quick approximation' … but it is no substitute to having all the data at hand to calculate the actual effect/accuracy implied by a correlation/validity coefficient.


Cognadev UK Ltd., 2 Jardine House, Harrovian Business Village, Harrow, Middlesex, HA1 3EX. Registered UK # 05562662

## The BESD

The Binomial Effect Size Display (BESD) was introduced in 1982 in an article:

Rosenthal, R., & Rubin, D.R. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 2, 166-169.

From p. 166:

> "The question addressed by BESD is What is the effect on the success rate (e.g., survival rate, cure rate, improvement rate, selection rate, etc.) of the institution of a certain treatment procedure? It displays the change in success rate (e.g., survival rate, cure rate, improvement rate, selection rate, etc.) attributable to a certain treatment procedure. "

From page 17 of Rosenthal, R., Rosnow, R.L., & Rubin, D.R. (2000) Contrasts and effect sizes in behavioral research: A correlational approach. Cambridge UK: Cambridge University Press. ISBN: 0-521-65980-9…

> "Instead of concentrating on $r^2$, we recommend using the point-biserial itself to create a display of the practical importance of the particular magnitude of effect (Rosenthal & Rubin, 1982). This is done simply by recasting r as a 2 x 2 contingency table, in which the rows correspond to the independent variable displayed as a table, in which the rows correspond to the independent variable displayed as a dichotomous predictor (e.g. experiment vs control) and the columns correspond to the dependent variable displayed as a dichotomous outcome (e.g. improved vs not-improved). The correlation between these two dichotomous variables is set to equal the obtained point-biserial r. The specific question addressed by this binomial effect size display (BESD) is: What is the effect on the success rate of the implementation of a certain procedure?
>
> Table 2.4 {below} illustrates the BESD based on an r of .32, which was reported to be the average size of the effect of psychotherapy in an early report of a meta-analysis (Glass, 1976). To find the psychotherapy success rate of 66%, we computed .50+ r/2, and to find the control success rate of 34%, we computed .50 - r/2. In other words, r = .32 is equivalent to increasing the success rate from 34% to 66% (which in another case might mean, for example, reducing an illness rate or a death rate from 66% to 34%). Notice that the difference between the rate of improvement in the psychotherapy group and that in the control group (i.e., 66% - 34% = 32%) corresponds to the value of r times 100. These percentages should not, of course, be mistaken for the raw percentages in the actual data, but they can be interpreted as "standardized" percentages in order for all the margins to be equal. Another way of saying this is that an r of .32 (or an $r^2$ of .10) will amount to a difference between rates of improvement of 34% and 66% if half the population received psychotherapy and half did not, and if half the population improved and half did not. "

**TABLE 2.4**
BESD for $r = .32$

| Condition | Treatment outcome: Improved | Treatment outcome: Not improved | Totals |
|---|---|---|---|
| Psychotherapy | 66 | 34 | 100 |
| Control | 34 | 66 | 100 |
| Totals | 100 | 100 | 200 |

In short, the BESD is the probability of a random outcome (0.5) plus one-half of a *point-biserial* correlation coefficient *(which is mathematically equivalent to a Pearson correlation and the conventional phi coefficient calculated from 2x 2 contingency tables).*

But, as Hsu, L.M. (2004) Biases of success rate differences shown in Binomial Effect Size Displays. *Psychological Methods*, 9, 2, 183-197 points out ..

"**Abstract:** The intent of a binomial effect size display (BESD) is to show "the [real-world] importance of [an] effect indexed by a correlation [r]" (R. Rosenthal, 1994, p. 242) by re-expressing this correlation as a success rate difference (SRD) (e.g., treatment group success rate - control group success rate). However, SRDs displayed in BESDs generally overestimate real-world SRDs implied by correlations of (a) dichotomous X and Y variables ($\phi$ coefficients), (b) dichotomous X and continuous Y variables (point-biserial coefficients [$r_{pb}$s]), and (c) continuous X and Y variables ($r_{xy}$s). Furthermore, overestimation biases are larger for $r_{xy}$s than for $r_{pb}$s. Differences in the sizes of biases linked to different correlations suggest that BESD SRDs reported for different correlations are not comparable. The stochastic difference index (N. Cliff, 1993; A. Vargha & H. D. Delaney, 2000) is recommended as an alternative to the BESD."

As Hsu says on p. 183 ..

"What may not be apparent from Rosenthal and Rubin's illustration is that the equality of the BESD SRD (calculated from Equation 1) and actual SRD of a 2 x 2 table does not generalize to 2 x 2 tables that do not have uniform marginal distributions."

Eq. 1 is: $$BESD\_SRD = \left[\left(.5 + \frac{r}{2}\right) - \left(.5 - \frac{r}{2}\right)\right] = r$$

And this gentle statement on page 18 from Rosenthal, R., Rosnow, R.L., & Rubin, D.R. (2000) Contrasts and effect sizes in behavioral research: A correlational approach. Cambridge UK: Cambridge University Press.

"It can be shown that the BESD is most appropriate when the variances within the two conditions are similar, as they are assumed to be whenever we compute the usual t statistics and its associated p-value."

Ok – all this is fine and dandy as it goes – but it's still a bit disconnected from what we really want to use the BESD for .. converting a correlation computed from a 2 x 2 decision-table to an index of classification accuracy. So, if we observe a correlation of 0.32 between a predictor (dichotomized into high-low say using a cut-score) and binary outcome (success, failure), then we can express that

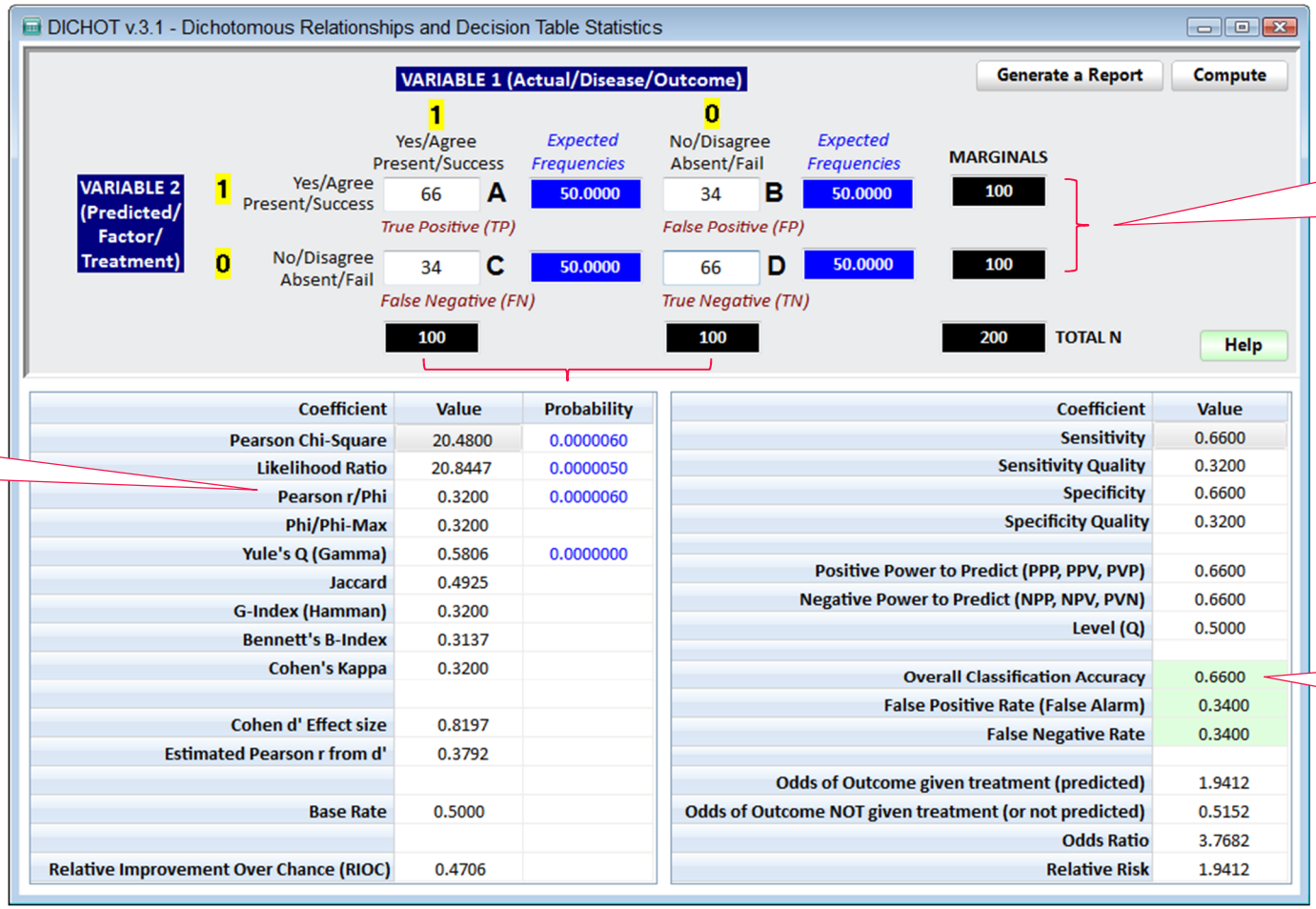
correlation as how well we can classify outcomes above a chance level of 50%. That chance level is actually no more than a base-rate of positive outcome of .5 (50%).

For example, if we observe a correlation between those scoring above 70 and below 70 on a competency, and rated job success after 1 year, of .4, our overall classification accuracy would be

$$BESD = \left(.5 + \frac{r}{2}\right) = \left(.5 + \frac{.4}{2}\right) = .7 \text{ or } 70\% \text{ classification accuracy.}$$

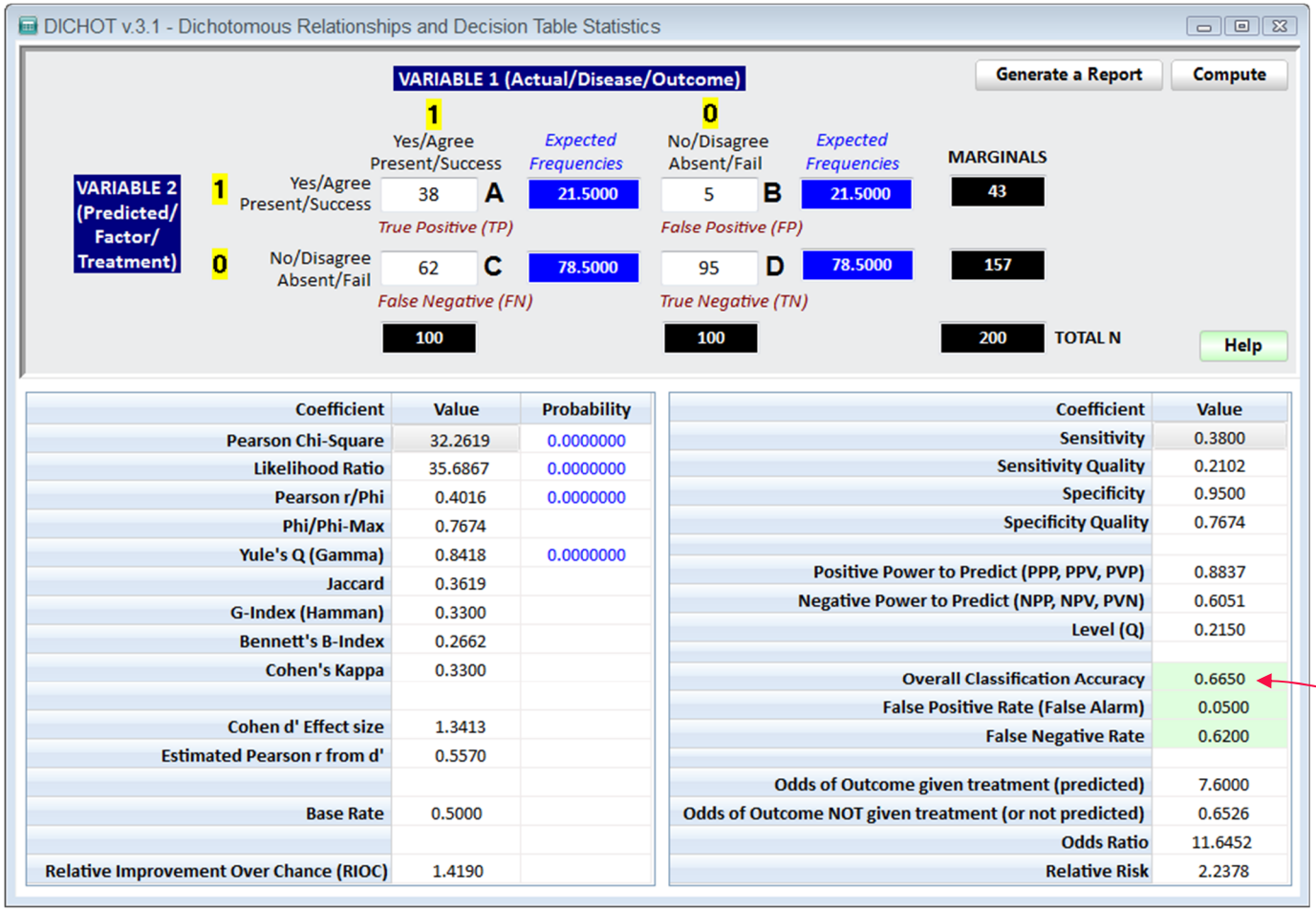
So, let's see a few 'full feature' worked examples to really get to grips with the BESD, warts and all. I'm using my Dichot 3.1 software – which is freely available for download from the web:
http://www.pbarrett.net/Dichot3/Dichot3.html
and includes a pdf manual and web-help embodies in the program.

## The Rosenthal and Rubin example presented in their Table 2.4 (on my page 1 above) …



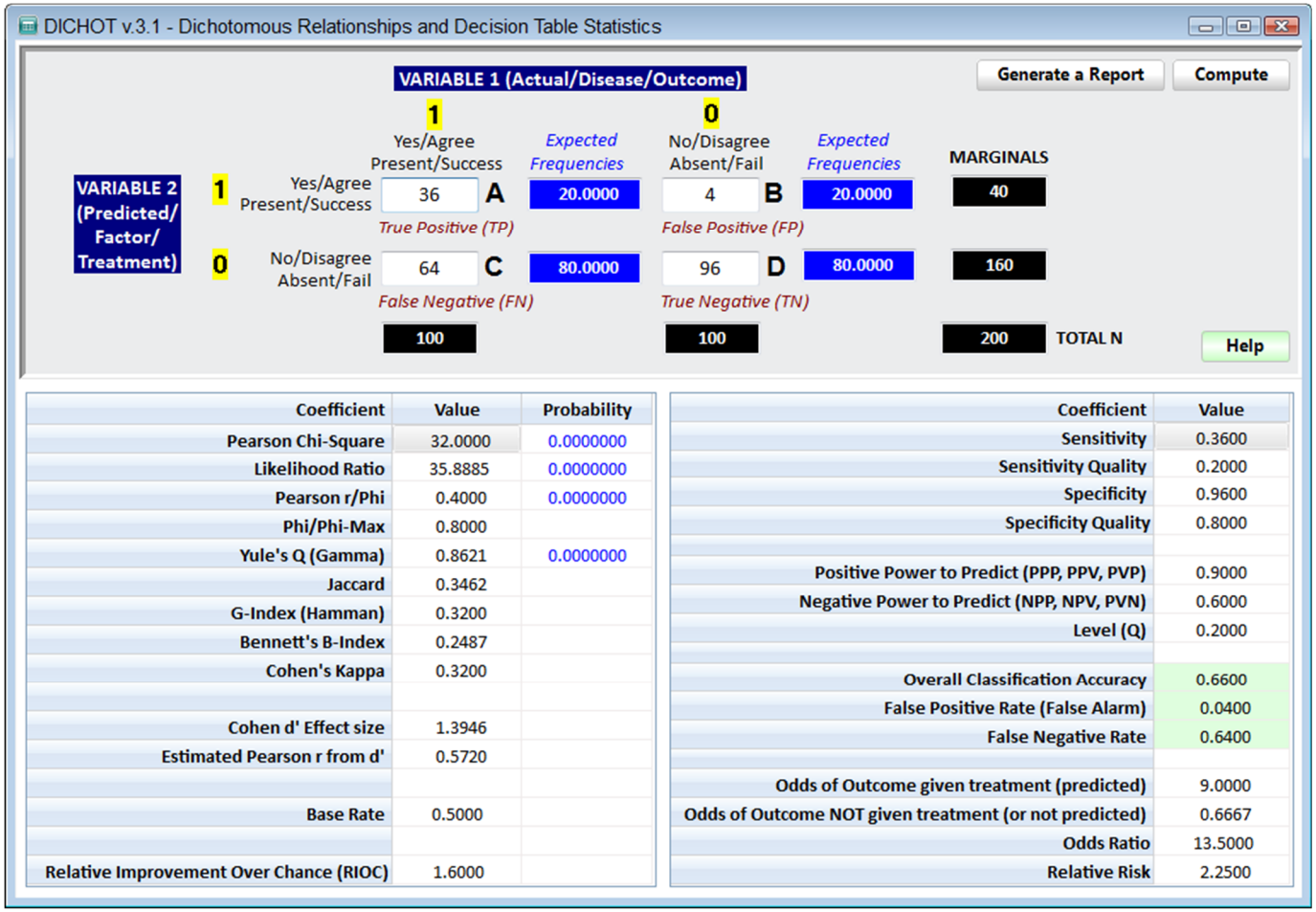

For an explanation & formulae for all the other coefficients reported here – download the program manual (if you haven't already downloaded the software).
http://www.pbarrett.net/Dichot3/Dichot3_1_program_manual.pdf

## Where the Base Rate remains at 0.5 (50%), but the error-rates are no longer balanced (unequal marginals)




The $BESD = \left(.5 + \frac{.4016}{2}\right) = .7008$ or $70.08\%$ effect display is not equal to the overall classification accuracy of 66.50%.

## Where the Base Rate remains at 0.5 (50%), but the error-rates are no longer balanced (unequal marginals)




The $BESD = \left(.5 + \frac{.4}{2}\right) = .7 \text{ or } 70.0\%$ effect display is not equal to the overall classification accuracy (effect) of 66%.

## Where the Base Rate is just above chance 0.537 (54%), and the error-rates are not balanced (unequal marginals)

DICHOT v.3.1 - Dichotomous Relationships and Decision Table Statistics

Generate a Report | Compute | Help

| | | VARIABLE 1 (Actual/Disease/Outcome) 1 Yes/Agree Present/Success | Expected Frequencies | 0 No/Disagree Absent/Fail | Expected Frequencies | MARGINALS |
|---|---|---|---|---|---|---|
| VARIABLE 2 (Predicted/ Factor/ Treatment) | 1 Yes/Agree Present/Success | 36 A True Positive (TP) | 19.8704 | 1 B False Positive (FP) | 17.1296 | 37 |
| | 0 No/Disagree Absent/Fail | 80 C False Negative (FN) | 96.1296 | 99 D True Negative (TN) | 82.8704 | 179 |
| | | 116 | | 100 | | 216 TOTAL N |

| Coefficient | Value | Probability |
|---|---|---|
| Pearson Chi-Square | 34.1269 | 0.0000000 |
| Likelihood Ratio | 42.9326 | 0.0000000 |
| Pearson r/Phi | 0.3975 | 0.0000000 |
| Phi/Phi-Max | 0.9416 | |
| Yule's Q (Gamma) | 0.9561 | 0.0000000 |
| Jaccard | 0.3077 | |
| G-Index (Hamman) | 0.2500 | |
| Bennett's B-Index | 0.1781 | |
| Cohen's Kappa | 0.2848 | |
| | | |
| Cohen d' Effect size | 1.8355 | |
| Estimated Pearson r from d' | 0.6752 | |
| | | |
| Base Rate | 0.5370 | |
| | | |
| Relative Improvement Over Chance (RIOC) | 3.0991 | |

| Coefficient | Value |
|---|---|
| Sensitivity | 0.3103 |
| Sensitivity Quality | 0.1678 |
| Specificity | 0.9900 |
| Specificity Quality | 0.9416 |
| | |
| Positive Power to Predict (PPP, PPV, PVP) | 0.9730 |
| Negative Power to Predict (NPP, NPV, PVN) | 0.5531 |
| Level (Q) | 0.1713 |
| | |
| Overall Classification Accuracy | 0.6250 |
| False Positive Rate (False Alarm) | 0.0100 |
| False Negative Rate | 0.6897 |
| | |
| Odds of Outcome given treatment (predicted) | 36.0000 |
| Odds of Outcome NOT given treatment (or not predicted) | 0.8081 |
| Odds Ratio | 44.5500 |
| Relative Risk | 2.1770 |

The $BESD = \left(.5 + \frac{.3975}{2}\right) = .6988$ or $69.88\%$ effect display is not equal to the overall classification accuracy (effect) of 62.5%.

## Where the Base Rate is substantively above chance 0.5941 (59%), and the error-rates are not balanced (unequal marginals)

DICHOT v.3.1 - Dichotomous Relationships and Decision Table Statistics

Generate a Report | Compute | Help

| | | VARIABLE 1 (Actual/Disease/Outcome) 1 Yes/Agree Present/Success | Expected Frequencies | 0 No/Disagree Absent/Fail | Expected Frequencies | MARGINALS |
|---|---|---|---|---|---|---|
| VARIABLE 2 (Predicted/ Factor/ Treatment) | 1 Yes/Agree Present/Success | 70 A *True Positive (TP)* | 46.5926 | 4 B *False Positive (FP)* | 27.4074 | 74 |
| | 0 No/Disagree Absent/Fail | 100 C *False Negative (FN)* | 123.4074 | 96 D *True Negative (TN)* | 72.5926 | 196 |
| | | 170 | | 100 | | 270 TOTAL N |

| Coefficient | Value | Probability |
|---|---|---|
| Pearson Chi-Square | 43.7382 | 0.0000000 |
| Likelihood Ratio | 53.1884 | 0.0000000 |
| Pearson r/Phi | 0.4025 | 0.0000000 |
| Phi/Phi-Max | 0.8541 | |
| Yule's Q (Gamma) | 0.8876 | 0.0000000 |
| Jaccard | 0.4023 | |
| G-Index (Hamman) | 0.2296 | |
| Bennett's B-Index | 0.2192 | |
| Cohen's Kappa | 0.3104 | |
| | | |
| Cohen d' Effect size | 1.5295 | |
| Estimated Pearson r from d' | 0.5941 | |
| | | |
| Base Rate | 0.6296 | |
| | | |
| Relative Improvement Over Chance (RIOC) | -0.0118 | |

| Coefficient | Value |
|---|---|
| Sensitivity | 0.4118 |
| Sensitivity Quality | 0.1897 |
| Specificity | 0.9600 |
| Specificity Quality | 0.8541 |
| | |
| Positive Power to Predict (PPP, PPV, PVP) | 0.9459 |
| Negative Power to Predict (NPP, NPV, PVN) | 0.4898 |
| Level (Q) | 0.2741 |
| | |
| Overall Classification Accuracy | 0.6148 |
| False Positive Rate (False Alarm) | 0.0400 |
| False Negative Rate | 0.5882 |
| | |
| Odds of Outcome given treatment (predicted) | 17.5000 |
| Odds of Outcome NOT given treatment (or not predicted) | 1.0417 |
| Odds Ratio | 16.8000 |
| Relative Risk | 1.8541 |

The bottom line, for me, is that the BESD is a reasonable stab at what the overall classification accuracy would be for a 2 x 2 decision-table phi (correlation/validity coefficient).

But, it can seriously mislead if the error-rates are not balanced in terms of yielding near-equal marginal, as Hsu (2004) indicates…
"BESD SRDs tend to overestimate targeted real-world SRDs in virtually all real-world applications" (p. 195)

The $BESD = \left(.5 + \frac{.4025}{2}\right) = .7013 \text{ or } 70.13\%$ effect display is not equal to the overall classification accuracy (effect) of 61.48%.