Calibration Performance Indicators for the Static-99R: 2013 Update

Author: Gregory DeClue, Sarasota, Florida gregdeclue@me.com

Author: Terence W. Campbell, Sterling Heights, Michigan

Abstract

Providing comprehensive statistical descriptions of tool performance can help give researchers, clinicians, and policymakers a clearer picture of whether structured assessment instruments may be useful in practice (Singh, 2013). We report positive predictive value (PPV), negative predictive value (NPV), number needed to detain (NND), and number safely discharged (NSD), along with associated confidence intervals (CIs), for each value of the Static-99R, for one data set. Values reported herein apply to detected sexual recidivism during a 5-year fixed follow-up for the samples that the Static-99R developers consider to be roughly representative of all adjudicated sex offenders (Phenix, Helmus, & Hanson, July 26, 2012).

Keywords: Static-99R, predictive accuracy, calibration performance indicators, number needed to detain, number safely discharged, sexual recidivism, risk assessment

In his recent methodological primer, Singh (2013) described differences between two components of predictive validity of violence risk assessments: calibration and discrimination. *Discrimination* indicators, such as the area under the receiver operating characteristic curve (AUC), address how well an instrument is able to separate those who went on to be violent from those who did not. *Calibration* indicators tell us how well a risk-assessment tool's predictions of risk agree with actual observed risk. Singh (2013) recommended that future research into the predictive validity of violence risk assessment tools include multiple performance indicators that measure different facets of predictive validity.

Once an evaluator selects a risk-assessment tool, discrimination indicators are no longer the primary focus. It is the calibration indicators that provide information about hits and misses. In this brief article, we focus on the following calibration indicators: positive predictive value (PPV), negative predictive value (NPV), number needed to detain (NND) and number safely discharged (NSD). As Singh (2013) points out, all four of these performance indicators are base-rate dependent and vary depending on the population, time at risk, and outcome of interest. The data presented here supplement Campbell (2011) by adding NND and NSD, with associated confidence intervals (CIs), along with CIs for PPV and NPV.

DeClue, G., & Campbell, T. W. (2013). Calibration performance indicators for the Static-99R: 2013 update. *Open Access Journal of Forensic Psychology, 5*, 82-88.

Using NND and NSD, an evaluator can tell a decision maker the following regarding, for example, a score of 7 on the Static-99R: *The number of people with a Static-99R score of 7 or higher that would have to be detained in order to prevent a sexual re-offense is five. That is, in order to prevent one act of sexual recidivism within five years, we would have to lock up five people with scores of 7 or higher. Conversely, the number of people with a Static-99R score of 7 or below that could be released prior to one act of sexual recidivism within five years is 18. That is, we could release up to 18 persons with scores of 7 or below prior to a single act of sexual recidivism.*

Singh (2013, p. 13) recognizes that consumers of research or clinical casework who see NND and NSD are directly faced with moral choices: "For example, some may consider the unnecessary detention of, say, five people to prevent the violent behavior of a sixth an appropriate measure to ensure public safety, whereas others may feel that the civil rights of those five unnecessarily detained individuals are of greater importance." We consider this to be an important advantage for the use of NND and NSD in forensic cases such as those involving civil commitment of sexually violent predators (SVPs). A proper role for a forensic evaluator is to investigate and then present data in a way that is quintessentially relevant to the practical and moral decisions to be made by the judge and/or jury. Evaluators need not advocate for or against a particular decision ("better that five people be unnecessarily detained . . ."), but we consider it to be important that the evaluators' risk communication helps the decision maker understand the practical impact of the decision to be made.

Because PPV, NPV, NND, and NSD are base-rate dependent, we focus on the Static-99R comparison group that the Static-99R developers consider to be "samples representing the full population of all [convicted sex] offenders."[1] In materials at the official website of the Static-99 and related instruments, www.static99.org, the developers refer to this comparison group as "Routine Correctional," "RC," or "Routine Samples." Here, we refer to the same group as FULLPOP, to keep the focus on the fact that this data set is considered to represent the full population of all convicted sex offenders.[2]

[1] Phenix, A., Helmus, L., & Hanson, R. K. (July 26, 2012, p. 32). Static-99R & Static-2002R Evaluators' Workbook. www.static99.org "This group consisted of eight samples of sex offenders from Canada, the United States, England, Austria and Sweden. These samples were relatively random (i.e., unselected) samples from a correctional system (as opposed to samples from specific institutions or subject to specific measures). In other words, they can be considered roughly representative of all adjudicated sex offenders. Some offenders in these samples would have been subsequently screened for treatment or other special measures (e.g., psychiatric admission or exceptional measures related to dangerousness), but these samples represent the full population of all offenders prior to any preselection processes. The recidivism norms for the unselected samples are the closest available to a hypothetical average of all sex offenders."

[2] If/when the local base rate is known, base-rate dependent calibration performance indicators such as PPV, NPV, NND, and NSD should be calculated using local data. When local base rate is unknown (as is often the case), it is generally preferable to use data considered to be representative of the full sample of convicted sex offenders, rather than guessing that the local base rate might be similar to that of some non-representative sample (DeClue, 2013; DeClue & Zavodny, manuscript submitted for publication).

Definitions and Descriptions

Table 1 presents a standard 2 X 2 classification table.

Table 1: Standard 2 X 2 Table

	Detected to Have Reoffended	Not Detected to Have Reoffended	
Predicted to Reoffend	**True Positive (TP)**	**False Positive (FP)**	→ Positive Predictive Value (PPV)
Not Predicted to Reoffend	**False Negative (FN)**	**True Negative (TN)**	→ Negative Predictive Value (NPV)
	↓ True Positive Rate (TPR)	↓ False Positive Rate (FPR)	

The following brief descriptions facilitate understanding of Tables 1 and 2. For more detailed descriptions, see Singh (2013).

True Positive (TP) is the number of people predicted to sexually reoffend, who are detected to have sexually reoffended.

False Positive (FP) is the number of people predicted to sexually reoffend, who are not detected to have sexually reoffended.

True Negative (TN) is the number of people predicted not to sexually reoffend, who are not detected to have sexually reoffended.

False Negative (FN) is the number of people predicted not to sexually reoffend, who are detected to have sexually reoffended.

Positive Predictive Value (PPV) identifies the accuracy with which one rules in recidivism risk. PPV is the proportion of people predicted to sexually reoffend, who are detected to have sexually reoffended. PPV = TP ÷ (TP + FP).

Negative Predictive Value (NPV) identifies the accuracy with which one rules out recidivism risk. NPV is the proportion of people predicted not to sexually reoffend, who are not detected to sexually reoffend. NPV = TN ÷ (TN + FN).

True Positive Rate (TPR), also called **Sensitivity**, is the proportion of people who are detected to have sexually reoffended, who had been predicted to sexually reoffend. TPR = TP ÷ (TP + FN).

False Positive Rate (FPR) is the proportion of people who were detected to have sexually reoffended, who were not predicted to sexually reoffend. FPR = FP ÷ (FP + TN).

Specificity is the proportion of people who were not detected to have sexually reoffended, who were predicted to not sexually reoffend. It is calculated as TN ÷ (TN + FP).

Number Needed to Detain (NND) calculates the number of individuals judged by a risk-assessment tool to be at high risk of committing a violent act who would need to be detained in order to prevent a single incident of violence from occurring in the community (Fleminger, 1997; Singh, 2013). NND = 1 ÷ PPV.

Number Safely Discharged (NSD) calculates the number of individuals judged to be at low risk who could be discharged prior to a single violent incident occurring in the community (Fazel, Singh, Doll, & Grann, 2012; Singh, 2013). NSD = [1 ÷ (1 – NPV)] - 1.

Additional notes for Table 2:

- The data set was originally reported in 2009. We accessed it from http://www.static99.org/pdfdocs/detailed_recid_tables_static99r_2009-11-15.pdf We constructed 2 X 2 tables for each Static-99R score, using the 5-year fixed follow-up data. TP, FP, TN, and FN were taken directly from the 2 X 2 tables.
- We calculated PPV, NPV, and associated confidence intervals (CIs) using Clinical Calculator 1 at http://vassarstats.net/
- We calculated NND and NSD using the formulas mentioned above, as described by Singh (2013).
- BR is base rate, which is the percentage of people in the entire group who were detected to have sexually reoffended.
- Three values were undefined because their calculation would entail division by zero. These undefined values are designated as “X.”
- We report PPV, NPV, and their associated CIs to 2 decimal places. In calculating NND, NSD, and their associated CIS, we used all 6 decimal places provided in Clinical Calculator 1 at http://vassarstats.net/
- In reporting NND and NSD, we maintain the tradition of rounding up because one cannot detain or release a fraction of a person.

Table 2: Calibration Indicators for the Static-99R FULLPOP Sample

Static-99R Score	TP	FP	TN	FN	PPV (95% CI)	NPV (95% CI)	NND (95% CI)	NSD (95% CI)
-3 & above	145	2261	0	0	.06 (.05 - .07)		17 (15 - 20)	
-3 & below	145	2221	40	0		1.0 (.89 - 1.0)		X
-2 & above	145	2221	40	0	.06 (.05 - .07)		16 (14 - 20)	
-2 & below	145	2156	105	0		1.0 (.96 - 1.0)		X
-1 & above	145	2156	105	0	.06 (.05 - .07)		16 (14 - 19)	
-1 & below	138	1903	358	7		.98 (.96 - .99)		52 (24 - 118)
0 & above	138	1903	358	7	.07 (.06 - .08)		15 (13 - 18)	
0 & below	130	1617	644	15		.98 (.96 - .99)		43 (26 - 75)
1 & above	130	1617	644	15	.07 (.06 - .09)		14 (12 - 16)	
1 & below	120	1277	984	25		.98 (.96 - .98)		40 (27 - 60)
2 & above	120	1277	984	25	.09 (.07 - .10)		12 (10 - 14)	
2 & below	106	941	1320	39		.97 (.96 - .98)		34 (25 - 48)
3 & above	106	941	1320	39	.10 (.08 - .12)		10 (9 - 12)	
3 & below	86	618	1643	59		.97 (.96 - .97)		28 (22 – 37)
4 & above	86	618	1643	59	.12 (.10 - .15)		9 (7 - 11)	
4 & below	69	358	1903	76		.96 (.95 - .97)		26 (20 - 32)
5 & above	69	358	1903	76	.16 (.13 - .20)		7 (5 - 8)	
5 & below	41	193	2068	104		.95 (.94 - .96)		20 (17 - 25)
6 & above	41	193	2068	104	.18 (.13 - .23)		6 (5 - 8)	
6 & below	27	97	2164	118		.95 (.94 - .96)		19 (16 - 23)
7 & above	27	97	2164	118	.22 (.15 - .30)		5 (4 - 7)	
7 & below	15	34	2227	130		.94 (.93 - .95)		18 (15 - 21)
8 & above	15	34	2227	130	.31 (.19 - .46)		4 (3 - 6)	
8 & below	7	14	2247	138		.94 (.93 - .95)		17 (14 - 20)
9 & above	7	14	2247	138	.33 (.15 - .57)		4 (2 - 7)	
9 & below	2	6	2255	143		.94 (.93 - .95)		17 (14 - 19)
10 & above	2	6	2255	143	.25 (.04 - .64)		4 (2 - 23)	
10 & below	0	1	2260	145		.94 (.93 - .95)		16 (14 - 19)
11 & above	0	1	2260	145	.00 (.00 - .95)		X	
11 & below	0	0	2261	145		.94 (.93 - .95)		16 (14 - 19)

Discussion

Singh (2013, p. 11, citation omitted) noted, "Reporting only the AUC, as do over half of violence risk assessment validation studies, does not provide adequate evidence of a risk assessment tool's predictive validity. The AUC measures discrimination but not calibration, meaning that it paints but half the picture. This said, available calibration indicators that could be used to describe an instrument's performance in identifying higher- versus lower-risk groups (e.g., PPV, NPV, NND, NSD) depend on a single cut-off threshold, which many modern risk assessment tools lack." Here, we overcome that barrier for one risk-assessment tool by providing PPV, NPV, NND, and NSD for each value of the Static-99R, for one group of sex offenders. That is the group that is considered by the Static-99R developers to be "roughly representative of all adjudicated sex offenders" (Phenix, Helmus, & Hanson, July 26, 2012, p. 32).

The data in Table 2 can be used to aid in understanding the predictive accuracy of the Static-99R in the FULLPOP group, using detected sexual recidivism within five years as the criterion. One could predict, for example, that everyone with a Static-99R score of 7 or higher would sexually reoffend. Actually, 22% of them were detected to have sexually reoffended, and the other 78% were not, so such predictions would be wrong 78% of the time. That is what PPV tells us. If a decision were made to detain everyone with a score of 7, how many people would have to be detained to prevent one sexually violent act? The answer is 5 (NND).

The data in Table 2 can also aid in understanding the practical impact of releasing everyone with a Static-99R score of 7 or lower, using the same criterion. We could predict that everyone with a Static-99R score of 7 or below would not sexually reoffend. Actually, 94% of them were not detected to have sexually reoffended, and the other 6% were. That is what NPV tells us. If a decision were made to release everyone with a score of 7 or lower, how many people could be released prior to one act of sexual recidivism being detected? The answer is 18 (NSD).

Singh (2013, p. 12) concludes, "Providing more comprehensive statistical descriptions of tool performance has the potential to help give researchers, clinicians, and policy-makers a clearer picture of whether structured assessment instruments may be useful in practice." We hope that this article meets some of those goals for the Static-99R.

Limitation

The primary limitation is that of any study that reports calibration indicators for any violence-risk-assessment tool. PPV, NPV, NND, and NSD are all base-rate dependent and vary depending on the population, time at risk, and outcome of interest. The data in Table 2 are relevant to detected sexual recidivism during a 5-year fixed follow-up for the samples that the Static-99R developers consider to be roughly representative of all adjudicated sex offenders.

References

Campbell, T. W. (2011). Predictive accuracy of Static-99R and Static-2002R. *Open Access Journal of Forensic Psychology, 3,* 82-106.

DeClue, G. (2013). Years of predicting dangerously. *Open Access Journal of Forensic Psychology, 5,* 16-28.

DeClue, G., & Zavodny, D. Forensic use of the Static-99R: 1. Post-hoc, ergo poppy-cock. Manuscript submitted for publication.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: Systematic review and meta-analysis. *British Medical Journal, 345,* e4692. DOI:10.1136/bmj.e4692.

Fleminger, S. (1997). Number needed to detain. *British Journal of Psychiatry, 171,* 287. DOI:10.1192/bjp.171.3.287a.

Phenix, A., Helmus, L., & Hanson, R. K. (July 26, 2012). Static-99R & Static-2002R Evaluators' Workbook. Accessed June 27, 2013, from www.static99.org

Singh, J. (2013). Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences and the Law, 31,* 8-22.

Received June 23, 2013; revision submitted June 27, 2013; accepted June 27, 2013.