Diagnostic Reliability and Sex Offender Civil Commitment Evaluations: A Reply to Wollert (2007)

Dennis M. Doren[1] & Jill S. Levenson[2]
[1]Sand Ridge Secure Treatment Center in Wisconsin
[2]Department of Health Services or Lynn University

[Sexual Offender Treatment, Volume 4 (2009), Issue 1]

Abstract

A recent article by Richard Wollert (2007) purports to demonstrate that the diagnostic inter-rater reliability of sex offender civil commitment evaluations is not high enough to be worthy for courtroom testimony. That author relies on a series of analyses to support that argument. Unfortunately, those analyses were flawed in serious ways, raising doubt about the overall conclusions drawn. This article delineates Wollert s inaccurate assumptions, addresses methodological flaws, and offers an alternative interpretation regarding the accuracy of sexual offender civil commitment assessments.

Key words: diagnostic accuracy, sexual offenders, civil committment, reliability

Background

There are currently 20 USA jurisdictions with civil commitment laws specifically pertaining to sexual offenders (Arizona, California, Florida, Illinois, Iowa, Kansas, Massachusetts, Minnesota, Missouri, North Dakota, New Hampshire, New Jersey, New York, Pennsylvania, South Carolina, Texas, Virginia, Washington, Wisconsin, and the US Federal government). The criteria for commitment is somewhat different across the states, but must conform to guidelines set forth by the U.S. Supreme Court ("Kansas v. Hendricks," 1997; "Kansas v. Crane," 2002). In order to qualify for sexually violent predator (SVP) civil commitment, individuals must demonstrate a prior history of criminal sexual behavior, a mental disorder that creates a propensity for sexual recidivism, and a likelihood of reoffense (Doren, 2002; Janus & Meehl, 1997). Typically, the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (American Psychiatric Association, 2000) is used to diagnose the requisite mental condition. Actuarial risk assessment instruments are used to estimate the likelihood of reoffense; the most commonly used is the Static-99 (Hanson & Thornton, 1999; Jackson & Hess, 2007). Expert testimony concerning both diagnostic and risk assessment issues is a regular component of sexual offender civil commitment proceedings.

Recently, Wollert (2007) purported to show that the inter-rater reliability of the diagnostic component of relevant civil commitment evaluations is poor. The question of reliability is an important one; if there is poor consistency between raters in assessments of the requisite mental conditions for commitment, even perfect inter-rater consistency in the risk assessment portion of evaluations would still leave serious questions about the accuracy of overall recommendations for or against commitments.

This article critiques the analyses described by Wollert (2007). First we clarify some improper assumptions made by Wollert (2007) with regard to inter-rater reliability of civil commitment criteria. Next, we illustrate flaws in the methodologies Wollert utilized to assess the reliability of civil

commitment diagnoses. Finally, we draw some conclusions about the reliability of diagnostic criteria, statutorily required mental abnormalities, and civil commitment recommendations. The ultimate finding is that Wollert s (2007) analyses do not add meaningfully to our understanding about the degree of evaluators consistency in opinions concerning sexual offender civil commitment matters.

Wollert s assumptions were threefold. First, he implied that the reliability of individual DSM diagnostic categories were equivalent to the reliability of the mental abnormality needed for SVP commitment. Second, he proposed that high levels of positive agreement between evaluators may not reflect diagnostic reliability, but rather the expectations that evaluators hold about encountering certain conditions associated with SVP status. Third, he postulated that the presumptive base rate for Paraphilia NOS nonconsent (PNOSN) was overstated.

In order to validate these assumptions, Wollert (2007) reported results from two studies. First, he re-analyzed previously published data (without access to the dataset) supposedly to re-calculate the inter-rater reliability of the mental abnormality criterion using Bayes Theorem. Second, he conducted an investigation for the purpose of estimating a base rate for the diagnosis of Paraphila NOS, nonconsent. His results and our critique are detailed below.

Wollert s First Assumption

Mental abnormality is equivalent to DSM diagnosis

Evaluators conducting sexual offender civil commitment assessments work within their mandated statutory schemes. Across all 20 relevant jurisdictions, commitments are based on at least three components: (a) a history of at least one (but usually more than one) conviction for a sexually violent act; (b) a mental condition that involves a propensity to commit sexually violent acts; and (c) a certain degree of risk for a recidivistic sex offense in the future (Doren, 2002; Janus & Meehl, 1997). The analyses by Wollert (2007) specifically and solely pertained to the second component, the requisite mental condition for commitment.

The DSM-IV-TR is the current manual published by the American Psychiatric Association to guide the psycho-diagnostic process in the USA. As such, it is also the primary manual used across the country to determine diagnoses of sex offenders facing potential commitment. Wollert asserted that an analysis of the inter-rater consistency of DSM-IV-TR diagnoses directly reflects the inter-rater consistency of the statutorily mandated mental abnormality. This presumption is not accurate, however.

Consider that state laws use various terms to describe the mental disorder commitment requirement. These include mental abnormality (e.g., IA, MO, NY), mental abnormality or personality disorder (e.g., FL, KS, MA, NH, WA), mental disorder (e.g., AZ, IL, WI), a serious mental illness, abnormality, or disorder (US Federal law), a sexual, personality, or other mental disorder or dysfunction (MN, ND), and behavioral abnormality (TX). Often, these laws use phraseology not contained in any diagnostic manual including the DSM-IV-TR. For instance, in numerous states (e.g., CA, FL, IA, IL, KS, MO, NH, NJ, NY, PA, WA, WI), the definition for the requisite mental condition includes language similar to a congenital or acquired condition affecting the emotional or volitional capacity that predisposes the person to commit sexually violent offenses -- phrases alien to the DSM-IV-TR. The term predisposes is contained in the definition of the requisite mental status for commitment in no fewer than 16 of the 20 jurisdictions (AZ, CA, FL, IA, IL, KS, MA, MO, NH, NJ, NY, PA, SC, TX, WA, & WI; with VA s law instead using the word

renders within the same context).

In other words, even if the statutory definition for the requisite mental condition for commitment can be equated to a psychiatric diagnosis (or a combination of diagnoses), the typical statutory definition also includes another consideration related to predisposition that clearly goes beyond the DSM-IV-TR and represents a necessary component of the criteria for commitment. With these findings, it is clear that DSM-IV-TR diagnoses alone are not equivalent to the statutorily defined mental condition required for SVP commitment.

To illustrate how important this difference is, consider the situation where (a) one evaluator opines for a paraphilia of a certain type with no personality disorder while (b) a second evaluator of the same offender opines for a certain personality disorder with no paraphilia, but (c) both opined that the diagnosed disorder contained all of the other elements necessary for the existence of a mental abnormality. In such a situation, despite completely different DSM-IV-TR diagnoses, there would be complete agreement on the existence of a mental abnormality required for civil commitment. On the other hand, both evaluators could agree on the DSM-IV-TR diagnosis, but disagree whether other "mental abnormality" components (such as a predisposition) are present and therefore whether or not the offender has a mental abnormality. This would result in perfect DSM-IV-TR diagnostic reliability but failed mental abnormality reliability.

An argument might be made that if a larger concept (such as mental abnormality) is based on a smaller concept (such as DSM-IV-TR diagnoses) that has poor inter-rater reliability, then the larger concept also must have poor inter-rater reliability as well you cannot make a building more stable than its foundation. This argument is not valid, however, as it represents a misunderstanding of the real relationship between DSM-IV-TR diagnoses and requisite mental conditions for sexual offender civil commitments.

A "mental abnormality" can involve "diagnosis 1" or "diagnosis 2" or "diagnosis 3", etc. (as long as predisposition is found). The reliability of the least reliable diagnosis does not serve as the maximum reliability for "mental abnormality" (its foundation limit), but the lower end. When the "or" across diagnoses is considered, then the inter-rater consistency concerning "mental abnormality" *cannot go down* but instead can only stay the same or go up as more diagnoses are added to the list.

An arithmetic example will demonstrate. Suppose a person is interested in the consistency a number, any number, within the range of 1-5 (as compared to any other numbers) across the following sequences:

Sequence 1: 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9
Sequence 2: 2 3 4 5 1 6 7 8 9 2 3 4 5 1 6 7 8 9

The consistency for specific pairings of numbers within the range of 1-5 between the two sequences equals 0%. The consistency for any number within the complete range of 1-5, however, is 100%. Different numbers within the range of 1-5 do not interfere with the finding of perfect consistency as long as the range of numbers consistently group together. This is analogous to the issue of studying individual diagnoses when the real commitment issue is the more inclusive concept of mental abnormality. The reliability of individual diagnoses may or may not be poor, but that situation does not prevent a finding of high consistency of the more inclusive concept of mental abnormality.

That said, we recognize that DSM-IV-TR diagnoses are conventionally applied in SVP commitment cases, and that it is customary practice for evaluators to use the DSM categories to establish the

required mental abnormality. We also acknowledge that it is important for any diagnostic criteria to be applied in a reliable fashion. Indeed, Packard and Levenson (2006) investigated the reliability of the most common DSM diagnoses employed in civil commitment proceedings using a range of measures and found a high degree of consistency between evaluators. Additionally, they found a high degree of consistency in the ultimate decision: the recommendation for commitment. Recommendations for commitment cannot be made unless the requisite mental abnormality exists.

Wollert s Study #1

Wollert s first study re-analyzed previously published data (without having access to Levenson s 2004 data set) concerning diagnostic reliability in civil commitment cases in Florida. He disputed the findings of previous results which, for the convenience of the reader, are briefly described here.

A sample of 288 Florida cases in which sex offenders were evaluated for civil commitment, each involving two evaluators, was analyzed to investigate the inter-rater reliability of SVP commitment criteria (Levenson, 2004; Packard & Levenson, 2006). The evaluators in all cases worked independently of one another, without communication to one another during the simultaneously occurring assessment period. Levenson (2004) mistakenly omitted significance levels associated with kappa coefficients and therefore erroneously interpreted reliability as fair to poor for most diagnoses and poor for the overall civil commitment recommendation, despite that all associations were statistically significant. Moreover, the limitations of Kappa were extensively reviewed in Packard and Levenson (2006). Packard and Levenson (2006) reanalyzed the data using additional measures of reliability including proportions of agreement, odds ratios, and relative risk ratios. Among the 288 cases, the concordance rate for assessments for or against the finding of a mental abnormality was approximately 82% (kappa = .54, $p < .001$) (Packard & Levenson, 2006). In other words, in 82% of cases, evaluators agreed on the presence or absence of a mental abnormality. As well, Packard and Levenson (2006) reported overall proportions of agreement for various diagnostic categories, some of which were lower than for the mental abnormality: 68% for Paraphilia NOS, 69% for Personality Disorder NOS, 71% for Substance Use Disorder, and 76% for Antisocial Personality Disorder (p. 9), but some of which were higher: 85% for Pedophilia, 97% for Sexual Sadism, 93% for other mental illness and 95% for other personality disorder. Packard and Levenson concluded that a high degree of inter-rater agreement between evaluators indicated that SVP evaluations were indeed reliable.

Wollert disputed Packard and Levenson s (2006) conclusions and described them as illusory. Specifically, he asserted that high levels of agreement were inflated and were not attributable to the reliable application of diagnostic criteria. In fact, Wollert s language was imprecise, often confusing reliability with validity. Wollert incorrectly reported that Packard and Levenson (2006) indicated that evaluators were about 73% certain that their diagnoses for each of these two disorders [substance abuse and other mental illness] were correct (p. 177). He states on page 180: the post-evaluation levels of confidence/certainty in the accuracy of these diagnoses were still not high enough to dispel reasonable uncertainty as I have defined it. These assertions seem to be about validity, not reliability. In fact, Packard and Levenson were not assessing whether diagnoses were correctly or accurately applied, but rather the likelihood that two independent evaluators would agree that a particular sex offender met criteria for the same DSM-IV-TR diagnosis. Thus, though Wollert purports to have re-analyzed Levenson s data and found poor inter-reliability between raters, his conclusions and the resultant discussion focused largely on validity rather than reliability issues.

Another example of Wollert s confusion between validity and reliability can be seen in his description of the likelihood ratio of 1.52 for the agreement between evaluators on the civil

commitment decision, which can be interpreted to mean that the interrater reliability exceeds chance (Wollert, p. 177). A likelihood ratio represents the ratio of one outcome to another, and a likelihood over 1 increases the ratio, with higher values indicating that the outcome was more likely to occur (Vogt, 2005). In the present context, likelihood ratios over 1 represent an increased likelihood that the two evaluators would agree on the diagnosis (or the commitment recommendation), not a measure of certainty that the diagnosis was correct, as Wollert describes on page 177.

When Wollert s analyses were related to inter-rater reliability, he used an estimated pre-evaluation base-rate expectation of 84% as representing the chance that an evaluator thought the client would meet criteria for SVP commitment. He estimated the positive predictive value (PPV; the probability that a specific diagnosis assigned by one evaluator would also be assigned by a second evaluator) for civil commitment recommendation to be 89%. In fact, the dual rater probability for expecting to find a mental abnormality prior to evaluating a respondent was computed by the current authors (who did have access to the original data set) to be 85.5%, with a Positive Predictive Value (PPV) of 90.5% (= 210/232), a 5% difference. This finding means that (a) the *overall* 82% concordance rate for mental abnormality is not significantly different from the base rate (85.5%) (chi square = 1.34, df = 1, $p > .05$), and also that (b) the concordance rate for a positive finding for mental abnormality showed a 34.5% relative improvement over chance [i.e., 5%/(100%-85.5%)].

This latter finding is of particular importance given Wollert s stated concern about the inter-rater reliability of information provided in court. By definition, positive findings about mental abnormality are what take cases into court -- not negative findings. Using Levenson s data, negative findings showed a far lower concordance rate (i.e., Negative Predictive Value = 40.0%; 18/45), indicating that there were proportionally fewer cases in which both evaluators agreed on the absence of a mental abnormality. Given that the screening process had already pre-selected offenders specifically thought to have mental abnormalities (Lucken & Bales, 2008; as explained in more detail below) this finding about where inconsistencies tended to exist is exactly as it should have been.

Wollert claimed, however, that his conclusions were further supported by a failure to control for halo effects (p. 184). Specifically, Wollert asserted that because sex offenders are screened for civil commitment criteria before being referred for face-to-face evaluation, evaluators had an *a priori* expectation that the client would meet criteria for SVP commitment. This leads us to a critique of Wollert s second analysis.

Wollert s Study #2

A second analysis reported by Wollert (2007) concerns the specific diagnosis of Paraphilia Not Otherwise Specified, involving sexual arousal to nonconsenting interactions (abbreviated by Wollert as PNOSN). To conduct the desired analysis, the author needed to know the base rate of this diagnosis among convicted sexual offenders in general (i.e., to have a baseline against which the rate for SVP candidates could be compared). Unfortunately, the procedure used to make this determination rendered the analysis virtually meaningless.

A fundamental issue when determining any base rate is that proper sampling is conducted. Presumed in this process is (a) that sufficient numbers of subjects are included, (b) that there is no inherent bias in the selection of subjects, and (c) that the data obtained are measured accurately. The latter two of these underlying assumptions are sorely lacking in the procedure used by Wollert.

Wollert recruited the entire staff of six clinicians from one outpatient treatment clinic as a team of

diagnosticians for determining the base rate of PNOSN. Wollert (2007) offered little justification for why these therapists should be considered good diagnosticians other than their years of experience working with sexual offenders in treatment. They had reportedly worked with sexual offenders for 1 to 12 years (mean = 8) and had treated a range of 40 to 2,000 incarcerated sex offenders. None had ever completed a psychosexual evaluation, but were familiar with actuarial assessment and rendering diagnostic opinions. They were considered by Wollert (2007) to be an informed sample whose views were not contaminated by a vested interest in diagnostic rates for PNOSN (p. 180).

Beyond the potential selection bias of using diagnosticians from one clinic, there were some glaring methodological unorthodoxies in the research procedures. Astoundingly, this study made no attempt to collect actual data on the prevalence of a PNOSN diagnosis. Records were not reviewed to calculate the proportion of sex offenders diagnosed with PNOSN, nor were the clinicians given a sample of records and asked render a diagnosis for present or former clients. Rather, the clinicians were asked only to *remember and approximate* the number of past clients to whom a PNOSN diagnosis would have been applied. Specifically, the clinicians were asked to estimate the percentage of once-incarcerated sex offenders he or she had treated whose case history or symptoms matched the referents in the questionnaire.

Determining a base rate using selected clinicians memories over 1-12 years of experience seems highly problematic at best. Now consider that three of the six clinicians reportedly worked with 1000-2000 offenders each. Moreover, the procedure by which the clinicians were asked to estimate the number of diagnoses of PNOSN was as follows: they considered up to 9 types of case history information (borrowed from Doren, 2002) that served as guidelines for diagnosing the disorder. Because the clinicians may not have had any experience previously applying this behavioral list (this was unspecified by Wollert), this exercise became one not of memory of previous diagnoses but of creating new diagnoses across up to 2000 cases without ever looking back at any case material or interviewing any clients. Most ironically, despite Wollert s major emphasis on the inter-rater reliability of diagnostic criteria, no inter-rater reliability test for these diagnoses was reported to determine if two assessors would reach the same diagnostic conclusion about a given client, (in this case, thousands of clients).

Even more confounding was that the definition of PNOSN used by Wollert in his analysis required "that their 'erotic' arousal depended on having a nonconsensual partner - that is they couldn't get aroused unless their partner was resistant to the sexual advances" (p. 182). This definition describes an exclusive paraphilic condition. The DSM-IV-TR does not require that any paraphilic condition represents exclusive arousal in order for a proper diagnosis to be made. Hence, low figures by the clinicians may simply have reflected their memory of the proportion of clients who were *exclusively* PNOSN in their sexual arousal, and not those properly diagnosed with PNOSN.

Though Wollert asserted that data from key informants are recognized as useful for conducting Bayesian analyses (p. 182), the whole procedure seems problematic, from how the disorder was defined to how it was measured to how the base rate was calculated. Given the lack of inter-rater consistency measure or test of the accuracy of any one clinician s memory or impression, the validity of this dataset is unknown at best and highly dubious at worst. It is rather surprising that a data collection procedure that seemingly fails to meet conventional minimal standards of acceptable scientific rigor was published.

Finally, even if the determined base rate were reasonably measured and accurate, it would not be proper as the baseline against which to compare the Levenson (2004) or Packard and Levenson (2006) results. The Florida data used in those analyses stemmed from a selected set of offenders already pre-screened by clinical personnel prior to the diagnostic evaluations in question. Lucken

and Bales (2008) calculated that only about 6.5% of all sexual offenders scheduled to be released from Florida prisons were referred to the set of evaluators studied by Levenson. Most importantly, the findings from Lucken and Bales indicate that of the 6.5% referred, [the referring agency believed that] all had the requisite prior diagnoses for sexual deviance disorders (p.120). This situation is quite different from Wollert s outpatient clinicians who apparently had no obvious reason to believe that the proportion of mentally disordered offenders in their sample would be higher than typical. The evaluators in Florida had strong reasons to understand that their cases were already disproportionately selected for mental abnormality and recidivism risk, by both the small proportion of offenders referred and the deliberate attempt to screen for such conditions.

This difference in expectation is exactly why a likelihood ratio correction to a simple base rate is often necessary, but that correction must be anchored in the proper base rate to obtain meaningful results. This difference in presumed diagnostic base rates, from Wollert s 5% (specifically for exclusive PNOSN) to the intended 100% (for all types of mental abnormalities inclusive of PNOSN) according to Lucken and Bales, is about as big a difference as can be. Wollert s derived 5% base rate for PNOSN does not even remotely represent the situation in which the Florida evaluators worked, and is therefore not applicable to the analysis conducted, rendering the results invalid.

To further elucidate this point, Bayesian analyses and likelihood ratios are based on the idea that if one changes the relevant base rate, then expectations must change accordingly. The best description of the proper expectation by Florida SVP evaluators is that 100% of those referred for evaluation were believed to have the requisite mental abnormality, while the best expectation for Wollert s clinicians would remain in the single digits (for PNOSN at least). A crucial variable is the determination of the proper expectation, only after which can a statistical determination be made as to whether the actual finding differs from the expectation or not.

Study 2 by Wollert (2007) represents an attempt to re-analyze published data concerning inter-rater reliability of the PNOSN diagnosis, this time including consideration of the likelihood ratio. Given that the likelihood ratio (LR) was determined using data that are completely unreliable from a population with a drastically different underlying base rate for the relevant diagnosis, we can learn nothing of value from Wollert s computations.

Conclusions

The analyses described in Wollert (2007) were based on assumptions and procedures that were seriously flawed. A main presumption was that diagnostic inter-rater reliability is synonymous with the inter-rater reliability of evaluator opinions concerning the required mental status for commitment. This was found to be both logically and measurably inaccurate. Procedures for obtaining a crucial analytical variable -- the base rate to determine a Likelihood Ratio (LR) -- were found to be profoundly lacking scientific rigor. As such, conclusions reliant on the derived LR were essentially of no value.

The main overall conclusion drawn in Wollert (2007) was that evaluator inter-rater reliability of the mental status prong for commitment is insufficient to support recommendations for SVP commitment. Given the findings herein, however, it appears that Wollert s results reflect seriously flawed methodologies and his conclusions are therefore not likely to be valid. Attempts to apply the Wollert findings specifically within the sexual offender civil commitment realm would be seriously misguided and inappropriate.

Civil commitment of sexually violent predators remains a controversial policy, and ethical

practitioners are encouraged to debate the myriad social, political, legal, and methodological issues. Nonetheless, we cannot allow our personal opinions about the constitutionality of civil commitment to obscure the scientific data surrounding its implementation. As John Monahan, a pioneering scholar in the science of violence risk assessment, opined: Disagreement with the substantive merits of sexually violent predator statutes does not justify depriving decision makers of the only kind of scientific evidence empirically validated actuarial violence risk assessment that can effectuate their statutory goals (Monahan, 2006, p. 45). Monahan s point applies as well to the diagnostic reliability of mental abnormality criteria.

References

1. American Psychiatric Association (2000). Diagnostic and statistical manual of mental disorders, 4th edition, text revision. Washington, D.C.: Author.
2. Doren, D. M. (2002). Evaluating sex offenders: A manual for civil commitments and beyond. Thousand Oaks, CA: Sage Publications.
3. Hanson, R. K., & Thornton, D. (1999). Static-99: Improving actuarial risk assessments for sex offenders. (No. 1999-02). Ottawa: Department of the Solicitor General of Canada.
4. Jackson, R. L., & Hess, D. T. (2007). Evaluation for civil commitment of sex offenders: A survey of experts. Sexual Abuse: A Journal of Research and Treatment, 19, 425-448.
5. Janus, E. S., & Meehl, P. E. (1997). Assessing the legal standard for prediction of dangerousness in sex offender commitment proceedings. Psychology, Public Policy and Law, 3(1), 33-64.
6. Kansas v. Crane, 534 U.S. 407 (U.S. Supreme Court 2002).
7. Kansas v. Hendricks, 117 S. Ct. 2072 (U.S. Supreme Court 1997).
8. Levenson, J. S. (2004). Reliability of Sexually Violent Predator Civil Commitment Criteria. Law & Human Behavior, 28(4), 357-369.
9. Lucken, K., & Bales, W. (2008). Florida s Sexually Violent Predator program. Crime & Delinquency, 54(1), 95-127.
10. Monahan, J. (2006). A Jurisprudence of Risk Assessment: Forecasting Harm among Prisoners, Predators, and Patients. Virginia Law Review, 92(3), 391-435.
11. Packard, R., & Levenson, J. (2006). Revisiting the Reliability of Diagnostic Decisions in Sex Offender Civil Commitment. Sex Offender Treatment, 1(3).
12. Vogt, W. P. (2005). Dictionary of statistics and methodology: A nontechnical guide for the social sciences (3rd ed.). Thousand Oaks, CA: Sage Publications.
13. Wollert, R. (2007). Poor diagnostic reliability, the null-Bayes logic model, and their implications for sexually violent predator evaluations. Psychology, Public Policy, and Law, 13(3), 167-203.

Author address

Dennis M. Doren, Ph.D.
Sand Ridge Secure Treatment Center
301 Troy Drive
Madison, Wisconsin 53704
dorendm@dhfs.state.wi.us

Jill S. Levenson, Ph.D.,
Associate Professor of Human Services at Lynn University
Assaf Academic Center
Lynn University

3601 N. Military Trail
Boca Raton, FL 33431
jlevenson@lynn.edu