

WPE WebPsychEmpiricist

Bayes' Theorem and the DSM:
Is it a book of definitions or a book of tests?

6/1/2010

Michael Karson
University of Denver

Abstract

A hallmark of scientific classification is the application of Bayes' Theorem to methods of assigning individuals to categories. Indeed, Bayes' Theorem, in medical circles, is often referred to simply as *diagnostic validity*. DSM-V should present the data needed to calculate, for each diagnostic criterion set, how well it does its job of distinguishing people who have the diagnosis from people who don't. Clarification of the meanings of Antisocial Personality Disorder and Psychopathy is offered.

Key words: Bayes' Theorem, Diagnostic Validity, DSM-V, Psychopathy

Bayes' Theorem and the DSM:

Is it a book of definitions or a book of tests?

6/1/2010

“When *I* use a word,” Humpty Dumpty said, in a rather scornful tone, “it means just what I choose it to mean, neither more nor less.”

The *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision* (DSM-IV-TR; APA, 2000) purports to be an “official nomenclature” (p. xxiii). As the publishers prepare for the next edition, it may be useful to clarify whether *nomenclature* means a system of terms that *define* various mental disorders or a system of terms that *describe* various mental disorders. A nomenclature that describes what it categorizes would be one that treats its subject as something that exists in nature and that needs words and categories to help scientists conceptualize and work with its subject. A nomenclature that defines what it categorizes can be useful politically or socially, but need not concern a subject that exists in nature. Thus, the classification of species in biology is a descriptive nomenclature, and along with the naming of species we will often find descriptors to help the user decide if an individual is or is not a member of a given species (i.e., the descriptors can be used as a *test* for category membership). The classification of angels would lie at the other extreme, a system that need not report to nature in any way. The official definition of the difference between a cherub and a seraph is by its own terms correct. The social or political importance of the DSM can be inferred from its charge to its work groups to form a consensus rather than to find out about nature (APA, 1994, p. xv).

The crux of the issue is that a definition, as I am using the term, is tautological and clarifying, but it is not subject to empirical validation. All other propositions that categorize individuals are subject to empirical validation. Thus, the DSM's (p.312) requirement that schizophrenia last for at least 6 months is definitional; it tells us that for the sake of clarity, we will not use the term for briefer episodes of psychosis. It is clearly an arbitrary cutoff, since there cannot possibly be an actual real-world difference between a duration of 182 days and 183 days. Propositions should be identified as definitions or descriptions, so that the latter may be subjected to the empirical validation they demand.

Any system for categorization of naturally-occurring phenomenon must answer the critical-thinking questions raised by filling in the famous 2X2 table, replicated below as Table 1.

The table allows computation of Bayes' Theorem, which calculates the probability that a positive score on a test means category membership from the probability that category membership means a positive score on the test. These questions include the following (Karson & Goodwin, 2009). How do we know the category exists in nature? What is the gold standard for assigning category membership? What is the sensitivity of the test (the percentage of people in the category who score positively)? What is the specificity of the test (the percentage of people not in the category who score negatively)? What is the base rate of the category in the population the test is being used on? It is important to note that what is meant by test is any algorithm for assigning group membership. The question, then, is whether the DSM criterion sets are tests for the diagnostic category or definitions of them.

If the DSM were primarily a book of descriptions (with a few clarifying definitional details thrown in) of categories that occur in nature, then presumably the format for developing each description would be as follows. 1. A carefully diagnosed sample of people with a disorder would be identified by experts, establishing a gold standard for classifying people as having or not having the disorder against which subsequent sets of criteria could be judged. 2. The rate of agreement among experts as to who has the disorder would be published (and this would set the cap on how good a test a criterion set could be, since the test cannot be better at categorizing people than the gold standard by which it is measured). 3. A list of criteria (i.e., a test for the disorder) would be developed that included as many people with the disorder as possible and excluded as many people without the disorder as possible, especially focusing on excluding people whom clinicians typically suspect of having the disorder but don't. (A good test for Post-Traumatic Stress Disorder (PTSD) will distinguish people with PTSD from people with borderline personality disorder and simple phobias, not just from the general population.) 4. The percentage of people with the disorder correctly identified by the test—i.e., its sensitivity—would be reported. 5. The percentage of people who are correctly identified by the test as not having the disorder—i.e., its specificity—would be reported. These data would allow users to estimate or measure their local base rates and develop validity estimates.

The evidence that the DSM is intended as a scientific book of descriptions of disorders that are found in nature (i.e., a book of tests) is minimal. Rates of agreement, sensitivities, and specificities were not reported. A carefully diagnosed sample of people with each disorder were not identified.

Therefore, it would seem to be a book of definitions, since either its propositions are or are not answerable to empirical validation. If it were a book of definitions, then there would be no need to report rates of agreement about which individuals have a particular disorder. Instead, we would want to know only whether two properly trained clinicians can apply the definitions with the same results. Also, as definitions, the sensitivity and specificity would by definition be perfect—individuals who meet the criteria have the disorder and those who don't, don't. But if it's a book of definitions, why would it need to be revised? Why would it say, "The specific diagnostic criteria included in DSM-IV are meant to serve as guidelines to be informed by clinical judgment and are not meant to be used in a cookbook fashion" (APA, 2000, p. xxxii)? This sentence implies that one's judgment about whether a person has a particular disorder may override the listed criteria, which makes sense only if the criteria do not define the disorder but only describe it.

The plain answer to this conundrum seems to be that the DSM is supposed to be a book of tests (criteria that optimally identify disorders that exist in nature), but it is a book of such bad tests that the sensitivities and specificities are not reported and instead the book takes refuge as a book of definitions.

This absurdity is stark in many areas, but perhaps nowhere moreso than in the application of the criteria for Antisocial Personality Disorder (ASPD). In Hare's (1996) criticism of the diagnosis, he repeatedly referred to it as a separate disorder rather than as a bad test of psychopathy. He wrote, "The distinction between psychopathy and ASPD is of considerable significance to the mental health and criminal justice systems. Unfortunately, it is a distinction that is often blurred, not only in the minds of many clinicians but in the latest edition of DSM-IV." Hare even notes that the DSM itself refers to ASPD as being also known as psychopathy. Hare's thinking about psychopathy was not fuzzy—instead, he was coping with the use of the DSM as a book of definitions. Indeed, it is frequently referred to as a bible. In that context, it was probably daunting to get clinicians to think of the DSM as a book of bad tests, so instead he took the approach of accepting ASPD as a disorder but questioning its utility.

One of many other examples is provided by Loving & Lee (2006, p.141), who call ASPD and psychopathy "overlapping but distinct constructs." They also note that the two labels identify a "markedly different subset of people." Again, I am not criticizing Loving & Lee. Instead, I am highlighting their convoluted effort to contend with the proposition that the DSM is never simply

wrong. Once a diagnosis is included in the DSM, its application identifies a group of people (or, rather, it might identify a group of people—there's no way to know if it does because the reliability data are not presented with the diagnosis), but whatever group identified gets reified as a diagnostic category instead of the book getting criticized for identifying the wrong group of people.

As a book of tests, the DSM is subject to Bayes' Theorem (Karson, 2006; Wood, 1996). The following example is performed largely on speculation since the sensitivities and specificities of the DSM diagnoses have never been published in one place, although some authors have investigated these data for various disorders. But the data provided below reflect an educated guess as to the sensitivities and specificities of the DSM criterion sets. For this example, suppose you are running a community outpatient clinic and you have read some impressive articles about successful treatment of individuals with Borderline Personality Disorder (BPD). You set up the treatment model described in the articles and instruct your clinicians to identify people with BPD so as to refer them for this new treatment.

The test for BPD provided by DSM-IV is two-pronged—the individual must have a personality disorder as defined by meeting all 6 criteria on the first prong (these may be summarized as an enduring pattern of dysfunction with no other explanation) and the individual must have at least 5 of 9 symptoms associated with borderline pathology on the second prong. Let us assume that the sensitivity and specificity of this test are very high. Let's assume that the sensitivity (accurately identifying people with BPD) is .95 and the specificity (accurately identifying people who don't have BPD as not having it) is .80. These would be very respectable numbers in psychiatry.

Bayes' Theorem is solved by completing the 2X2 table that crosses identification by the test with actually having the condition of interest. In Table 1, the gold standard is the independent method by which we distinguish people with, in this case, BPD and people without it. Typically, this may be a carefully diagnosed sample, although in medicine, it will often involve a genetic marker or observing an external sign (for example, pregnancy tests may be validated on women divided by a gold standard of observing a fetus via ultrasound). The cutoff score simply tells us at what point the test (or criterion set) identifies the person as having or not having the condition. Usually field trials provide data about sensitivity and specificity, but we are

instead assuming .95 and .80. Table 2 presents the data that will help our clinicians make referral decisions, *so far*.

But Bayes' Theorem also requires that we take the base rate into account before judging the accuracy of a test. DSM-IV (APA, 1994, p. 652) indicates that the base rate of BPD in an outpatient mental health clinic is about 10%. Accepting that figure, our 2X2 table must be adjusted so that 90% of our subjects do not have the condition. The resulting figures are presented in Table 3.

Table 3 presents the same sensitivity and specificity, adjusting the cells for the 10% base rate. Given our assumptions, if a patient evaluated at an outpatient clinic meets the criteria for BPD, it's about 2 to 1 against his or her actually having the disorder. This is not a trick of arithmetic. This is a function of using diagnostic tests (criterion sets) whose sensitivities and specificities are only *good* to *very good* in the context of trying to identify conditions that are fairly rare. As long as there are so many diagnoses, there will be classification problems.

Implications for DSM-V

1. Every diagnosis included in DSM-V should be reported with its coefficient of reliability (the odds that two experienced clinicians will agree that an individual meets the criteria) for different settings.
2. Every diagnosis should be reported with the sensitivity of its criterion set when applied to a carefully diagnosed independent sample and with its specificity when applied to various comparison samples of interest.
3. DSM-V should emphasize even more than DSM-IV that diagnoses can only be made by clinicians, not by technicians applying criterion sets. Put differently, even the most experienced AIDS physician will defer to the blood test when it contradicts her clinical impression (especially if the test comes out the same a second time). This is because the sensitivity and specificity of the test for HIV antibodies are astronomical. There is no diagnostic test in psychiatry that strong, and mental health clinicians should not defer to tests when making diagnoses. Clinicians should distinguish between a patient's meeting the criteria for a diagnosis, which merely indicates a positive result on one test for it, and an opinion that a patient actually has the disorder in question, which must depend on all the information available, including the local base rate of the disorder.

4. It may be argued that a dimensional approach to diagnosis will obviate the problems exposed by Bayes' Theorem. A dimensional approach has much to recommend it, not least because it is far from clear that conceptual diagnostic categories actually occur in nature. However, as long as dimensional traits are reduced to diagnostic categorizations (via the implementation of cutoff scores), they will have sensitivities and specificities that ought to be reported and used to compute the accuracy rates of diagnoses.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington: Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., Text Revision). Washington: Author.

Hare, R.D. (1996). Psychopathy and antisocial personality disorder: A case of diagnostic confusion. *Psychiatric Times*, Vol. XIII, 2. Retrieved January 23, 2009 from <http://www.psychiatrictimes.com/display/article/10168/54831>.

Karson, M. (2006). Diagnostic validity. In N. Salkind (ed.), *Encyclopedia of measurement and statistics*. Thousand Oaks, CA: Sage.

Karson, M. & Goodwin, J. Critical thinking about critical thinking. In M. Karson, *Deadly therapy: Lessons in liveliness from theater and performance theory* (pp. 159 – 175). Lanham, MD: Jason Aronson.

Loving, J.L. & Lee, A.J. (2006) Rorschach assessment of antisocial personality disorder and psychopathy. In S.K. Huprich (ed.), *Rorschach assessment of the personality disorders*. Lawrence Erlbaum Associates: Mahwah, NJ.

Wood, J.M. (1996). Weighing evidence in sexual abuse evaluations: an introduction to Bayes' theorem. *Child Maltreatment*, 1(1), 25-36

Table 1

The Basic 2 X 2 Table

		Criterion (Gold Standard)	
		In the category	Not in the category
Test results (by cutoff score)	In the category	True Negative/ Specificity	False Negative
	Not in the category	False positive	True Positive/ Sensitivity

Specificity = True Negative/ Total Normals: Identification of those without the condition

Sensitivity = True positives/Total Abnormals: Identification of those with the condition

False Negative = 1-Sensitivity: How many with the condition does the test misidentify?

False Positive = 1-Specificity : How many without the condition does misidentify?

Hit rate = True Negatives + True Positives/Total in Sample: Overall accuracy

Base rate = Total Abnormals/Total in Sample: How often does the condition occur?

Table 2

Hypothetical Sensitivity and Specificity Presented for DSM-IV and BPD

	Actually BPD	Actually not BPD
DSM says BPD	95	20
DSM says not BPD	5	80

Table 3

Hypothetical Evaluation of DSM-IV for Detecting BPD

	Actually BPD	Actually not BPD
DSM says BPD	95	180
DSM says not BPD	5	720