Neuropsychologists’ abilities to determine the predictive value of diagnostic tests

Andrew S. Labarge a,*, Robert J. McCaffrey a, Timothy A. Brown b

aDepartment of Psychology, University at Albany, State University of New York, Albany, NY, USA
bCenter for Anxiety and Related Disorders, Boston University, Boston, MA, USA

Accepted 24 October 2001

Abstract

The diagnostic utility of a test is relative to the base rate of the diagnosis in the population of interest. The extensive use of tests in neuropsychological assessment makes knowledge of the relevance of base rates in this context an important issue. Professional and associate members of the National Academy of Neuropsychology (NAN; n = 279) answered questions involving (1) a simple use of base rates in the absence of additional diagnostic information, (2) sensitivity and specificity, and (3) positive predictive value (PPV) presented in either a probability or frequency format. Although the majority of participants answered correctly the simple base rate, sensitivity, and specificity questions, only 8.6% answered correctly the PPV question presented in a probability format versus 63.0% correct using a frequency format. These results are discussed in terms of education about the importance and application of base rate information and difficulties applying base rate information in practice.

Keywords: Neuropsychological assessment; Diagnosis; Predictive value

1. Introduction

Nearly a half-century ago, Meehl and Rosen (1955) described how base rates can impact the accuracy of clinical judgments in psychology, stating that the diagnostic utility of a sign (i.e., a symptom, test, or set of tests) is based not only on the sign’s sensitivity and specificity but
also on the base rate of the diagnosis in the population of interest. Sensitivity is the probability that a sign will be positive given that the disorder is present. Specificity is the probability that a sign will be negative given that the disorder is not present. Sensitivity and specificity are not impacted by the base rate of the disorder in question. Positive predictive value (PPV) is the likelihood that an individual has a disorder given a positive sign. Negative predictive value (NPV) is the likelihood that an individual does not have a disorder given a negative sign. In contrast to sensitivity and specificity, PPV and NPV are affected by the base rate of the disorder in question.

The formulas used to determine predictive value are derived from Bayes’ (1763) theorem. Using terminology familiar to the clinical decision-maker (BR stands for “base rate” and RC stands for “remaining cases” or 1 – BR), Bayes’ theorem indicates that PPV = (sensitivity × BR)/[(sensitivity × BR) + [(1 – specificity) × RC]] and NPV = (specificity × RC)/[(specificity × RC) + [(1 – sensitivity) × BR]].

In studies investigating clinicians’ use of base rate information, participants typically overestimate PPV and often respond erroneously that the predictive value of a test is equivalent to the test’s sensitivity or specificity (e.g., Casscells, Schoenberger, & Graboys, 1978; Heller, Saltzstein, & Caspe, 1992). In the typical clinical scenario in which the base rate of the disorder in question is below 50% and the test’s specificity is less than 100%, the PPV is typically less, and frequently far less, than the sensitivity and/or specificity of the test. For example, given a 10% base rate and a test with 80% sensitivity and 90% specificity (see question 4a in Appendix A), the likelihood of disorder given a positive test score (i.e., PPV) is equal to 47%. Under similar circumstances, the clinician unaware of the impact of base rates on PPV likely would overestimate the likelihood of disorder when he or she obtains a positive test score (Faust & Nurcombe, 1989). In this example, the likelihood that an individual with a negative screening score does not have the disorder (i.e., NPV) is 98%. The relatively greater certainty provided by the NPV allows the clinician interpreting a negative test score under similar circumstances a considerable degree of certainty that the individual does not have the disorder (Gorry, Pauker, & Schwartz, 1978; Vecchio, 1966). Appendix B further illustrates the difference between sensitivity and specificity versus predictive value using a 2 × 2 contingency table.

The preceding example presented information similar to that provided by questions utilizing probability formats, which are the formats used in most of the research on both clinicians’ and nonclinicians’ abilities to estimate predictive value. The same problems can be expressed in a more intuitive or “user-friendly” way known as frequency formats (Gigerenzer, 1996; Gigerenzer & Hoffrage, 1995). The frequency format provides essentially the same information as problems presented in probability formats but does so in terms of the frequencies of events. Question 4b in Appendix A provides an example of a PPV question presented in a frequency format. In both clinical and nonclinical studies, a significantly greater percentage of participants are able to answer correctly problems presented in a frequency format versus a probability format (Cosmides & Tooby, 1996; Gigerenzer, 1994, 1996; Gigerenzer & Hoffrage, 1995). It should be noted that there are differences between single-event probabilities (i.e., likelihood estimates) and frequencies, as well as debate regarding the applicability of probability theory to single events (see Gigerenzer, 1994). However, from the perspective of the clinical application of the principle illustrated by both of these types of problems (i.e., the impact of base rates on diagnostic accuracy), the information obtained from each type of problem is equivalent.
Beginning with the classic studies conducted by Kahneman and Tversky (1972, 1973), the study of whether "humans [are] good intuitive statisticians" (Cosmides & Tooby, 1996, p. 1) has developed into a substantial literature, largely within cognitive psychology (see Wright & Ayton, 1994 for a review). The majority of these studies use as participants undergraduates typically untutored in probabilistic reasoning and for whom the ability to utilize likelihood estimations is not essential to their work. There is, however, a smaller literature investigating probabilistic reasoning by clinical professionals for whom such a skill is essential (Garb, 1998; Garb & Schramke, 1996; Wedding & Faust, 1989). These studies have found that psychologists (Ford & Widiger, 1989; Kennedy, Willis, & Faust, 1997; Shagoury & Satz, 1969) and physicians (Balla, 1980; Balla, Elstein, & Gates, 1983; Casscells et al., 1978; Christensen-Szalanski & Bushyhead, 1981; Curley, Yates, & Young, 1990; Heller et al., 1992; Wallsten, 1981) frequently ignore or misuse base rate information when considering the likelihood of disorder or determining predictive value. Consistent, however, with the findings in nonclinical studies, the accuracy of physicians’ abilities to determine predictive value increase significantly when the information is presented in a frequency versus probability format (Gigerenzer, 1996).

The extensive use of tests in neuropsychological assessment and diagnosis makes knowledge of the relevance of base rates in this context an important issue. There is a growing focus on improving neuropsychologists’ and other clinical decision-makers’ diagnostic skills through the application of probabilistic reasoning to the interpretation of diagnostic tests, such as the use of likelihood ratios (Curley et al., 1990; Sackett, Richardson, Rosenberg, & Haynes, 1997) and odds ratios (Bieliauskas, Fastenau, Lacy, & Roper, 1997; Ivnik et al., 2000, 2001), which allow clinicians to make highly informative and scientifically responsible statements regarding the probability of a particular diagnosis given the test finding. The present study was designed to investigate neuropsychologists’ abilities to determine predictive value when presented with information in two different, but basic and straightforward, formats. Although previous studies have assessed clinical and school psychologists’ and physicians’ abilities to use base rate information to determine predictive value, to our knowledge, no one has investigated neuropsychologists’ abilities in this regard.

2. Method

2.1. Research instruments

Two surveys were developed. The first section of both surveys requested demographic information. Participants indicated their terminal degree; years since obtaining the terminal degree; board certification status (ABPN, ABCN, or other); primary work setting (private or group practice, hospital, university, medical school, or other); the number of hours per week spent in neuropsychological assessment/report writing, non-neuropsychological clinical work, research, and teaching; and primary approach to neuropsychological assessment (fixed battery, flexible battery, or process approach). The second section contained four questions (see Appendix A). The first three were the same in both surveys. Question 1 examined the participant’s ability to use base rate information in the absence of other applicable diagnostic information to determine the probability that an individual has a disorder. Questions 2 and 3 utilized a multiple-choice
format to determine the participant’s knowledge of the definitions of sensitivity and specificity. Question 4 examined the participant’s ability to determine PPV. In survey Form P, the relevant information was provided in a probability format (i.e., Question 4a). In survey Form F, the relevant information was provided in a frequency format (i.e., Question 4b).

2.2. Participants

After obtaining permission to use the mailing list from the National Academy of Neuropsychology (NAN), cover letters, surveys, and return envelopes were mailed to a pseudorandom sample of 800 individuals (400 for each form of the survey) selected from all Professional and Associate NAN members within the United States. For purposes of anonymity, no coding system was used to determine whether specific recipients had replied. Eight surveys were returned as undeliverable. The total number of completed responses was 279 (140 of Form P and 139 of Form F), yielding a 35% response rate, within the lower range of the 34–64% response rate obtained in similar studies. Respondents (n = 140) completed and returned Form P, and 139 completed and returned Form F.

3. Results

3.1. Demographics

χ² and independent groups t tests were used to examine for differences between Group P (i.e., participants who completed Form P) and Group F (i.e., participants who completed Form F) demographics, as well as to examine the relationships between demographics and the ability to answer the base rate, sensitivity, specificity, and PPV questions. Because of the low number of respondents with EdD and MD degrees (3 and 1, respectively), only participants with PhD and PsyD degrees were included in analyses. Groups P and F did not differ on any demographic variable except degree type [χ²(1, n = 274) = 4.97, P < .5]. Because of the interrelated nature of sensitivity and specificity, comparisons were made based on the responses to both questions combined. That is, the definitions of both sensitivity and specificity must have been correct to receive “credit” for the sensitivity/specificity component. If the definition of one or both terms was misidentified, the response was counted as incorrect. For both Groups P and F, there were no differences between PhD’s and PsyD’s responses to the base rate or sensitivity/specificity questions (all P values > .10).

Therefore, because there were no differences between the groups on the other demographic variables, for the remaining analyses examining the relationship between demographic variables and answers to the base rate and sensitivity/specificity questions, Groups P and F were collapsed. Comparisons between demographic variables and answers to PPV questions were conducted separately for each group. A modified Bonferroni control for experimentwise error rates was applied to the comparisons within each demographic variable (Holland & Copenhaver, 1988; Jaccard, Becker, & Wood, 1984). No demographic variable was related to the ability to answer correctly the base rate, sensitivity/specificity, or PPV questions (all P values > .10).
Table 1
Survey question responses

<table>
<thead>
<tr>
<th>Question</th>
<th>Group</th>
<th>n</th>
<th>Correct (%)</th>
<th>Mode</th>
<th>Mean (S.D.)</th>
<th>Range</th>
<th>Response = sensitivity (%)</th>
<th>Response = specificity (%)</th>
<th>Response = base rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base rate</td>
<td>P</td>
<td>138</td>
<td>77.5</td>
<td>10</td>
<td>13.27 (10.76)</td>
<td>0–60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>138</td>
<td>79.0</td>
<td>10</td>
<td>11.50 (7.90)</td>
<td>0–55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity and</td>
<td>P</td>
<td>139</td>
<td>77.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>specificity</td>
<td>F</td>
<td>137</td>
<td>70.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPV</td>
<td>P</td>
<td>131</td>
<td>8.6</td>
<td>80</td>
<td>70.37 (25.59)</td>
<td>7–100</td>
<td>35.1</td>
<td>19.8</td>
<td>6.9</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>135</td>
<td>63.0</td>
<td>47</td>
<td>41.62 (21.14)</td>
<td>7–100</td>
<td>3.7</td>
<td>1.5</td>
<td>8.1</td>
</tr>
</tbody>
</table>

Group P = Participants whose form of the survey contained the PPV question in a probability format; Group F = Participants whose form of the survey contained the PPV question in a frequency format.

A. Percentages of responses to Question 4a or 4b that were equal to the values for sensitivity, specificity, or base rate respectively.

B. To be considered correct, both sensitivity and specificity questions must be correct.

3.2. Responses to base rate, sensitivity/specificity, and PPV questions

Table 1 summarizes the responses to the survey questions. There was no difference between groups regarding the ability to answer the base rate or sensitivity/specificity questions (using $\chi^2$ tests, both $P$ values > .10). The difference between the percentage of correct responses to the PPV question when it was presented in a probability versus frequency format was statistically significant [$\chi^2(1, n = 266) = 82.30, P < .00001$]. The strength of this relationship, as indexed by Cramér’s statistic, was 0.56.

3.3. Relationships between base rate, sensitivity/specificity, and PPV responses

There were significant relationships between Group P’s correct responses on the PPV question and percentage of correct responses on the base rate [$\chi^2(1, n = 130) = 4.52, P < .05$], sensitivity/specificity [$\chi^2(1, n = 131) = 4.10, P < .05$], and both the base rate and sensitivity/specificity questions [$\chi^2(1, n = 130) = 8.18, P < .01$]. The strength of the relationship between correct responses on both base rate and sensitivity/specificity questions and a correct response on the probability PPV question, as indexed by Cramér’s statistic, was 0.25. In contrast, there were no relationships between Group F’s correct responses on the PPV question and percentage of correct responses on the base rate, sensitivity/specificity, and both the base rate and sensitivity/specificity questions (using $\chi^2$ tests, all $P$ values > .10). The results for Group P indicate that if a participant was able to solve the PPV question presented in a probability format, then he or she was able to apply base rate information correctly and knows the definitions of sensitivity and specificity. The results of Group P could also be interpreted as indicating that knowledge of the application of base rate, sensitivity, and specificity information is necessary, but not sufficient, for solving PPV questions presented in a probability format. The results for Group F indicate that knowledge of the application of base rate information or
the definitions of sensitivity and specificity is not necessary to solve PPV questions presented in a frequency format.

4. Discussion

The art and science of diagnostic decision-making requires the integration of information from a variety of sources, including patient report, medical records, and test results. The clinician’s job is not only to determine what information is relevant to the diagnostic question but also to determine the relative usefulness of the information. Whether wittingly or unwittingly, the principle that the diagnostic utility of a sign is relative to the base rate of the disorder in question impacts the likelihood of accuracy of every test interpretation or diagnosis a neuropsychologist makes. Despite this fact, the majority of neuropsychologists in the present study either neglected or misused base rate information when that information was presented explicitly in a format similar to that in which neuropsychologists would be expected to encounter it (i.e., as in the probability format). Although it could be argued that those surveyed should not have been expected to know Bayes’ theorem or some other formula for PPV “off the top of their heads,” it can be countered that a working familiarity with this principle should allow for accurate calculation of PPV, and the results can be interpreted as a lack of such a working familiarity.

In contrast to the difficulty participants experienced when applying base rate information presented in a probability format, the majority of participants who received the frequency format answered correctly. Does this indicate that those participants were more aware of the role of base rates in the interpretation of test results or were more knowledgeable about the application of this information? This is doubtful. The frequency format is a more “user-friendly” format that in a sense guides the respondent through the problem. These results speak to the potential usefulness of frequency formats to teach the use of base rates in likelihood estimation. Students should be taught how to frame problems using a frequency format. These findings also suggest that authors presenting research on test characteristics such as sensitivity and specificity should consider presenting the information in frequency formats. As demonstrated by the present study, however, the ability to solve PPV problems using frequency formats does not indicate knowledge of the fundamental principles underlying likelihood estimation. In fact, 35.3% of the participants who solved correctly the frequency PPV question answered incorrectly one or more of the base rate, sensitivity, or specificity items. Therefore, when using frequency formats in the instruction of probabilistic reasoning, it is critical to go beyond the simple solutions of the problems to assure understanding of the underlying principles.

4.1. Difficulties in the application of base rate information

Clearly, a significant obstacle to the appropriate application of base rate information is lack of awareness of the principle of predictive value, insufficient knowledge regarding its application, or unwillingness to utilize this information. It is uncertain if the present findings are the result of inadequate emphasis placed on the topic in the course of training (anecdotally,
in discussing this issue with colleagues, most indicated that the topic was given very little, if any, coverage in the course of their graduate training), forgetting once-learned information due to lack of use, the relatively little attention the topic receives within the neuropsychology literature, clinicians’ reluctance to adopt “statistical” approaches to test interpretation, and diagnosis over approaches based on “clinical experience” or some combination of these factors.

Ultimately, however, the most significant factors impeding progress in the clinical use of base rate information may be the practical and theoretical difficulties encountered in attempting to apply these principles in research and practice (Finn, 1982, 1983; Grove, 1985; Morey & McNamara, 1987; Widiger, 1983). Determining an appropriate base rate presents both practical and theoretical problems (Miettinen & Caro, 1994). Base rates refer to the prevalence of a sign or disorder within a specified population. Populations, however, can be defined in many different ways. Using a widely known example, the prevalence of dementia in the general population of individuals over age 85 is approximately 20% (American Psychiatric Association, 1994). Less commonly known is information such as the base rate of dementia in the population of African American individuals over age 85 with a past medical history of hypertension and diabetes who have been referred to a dementia clinic by their primary care physician secondary to neuroimaging findings of diffuse brain atrophy greater than expected for the patients’ age and family reports of memory problems and inappropriate behavior including placing the laundry in the oven and getting out of bed and dressing to go out at 3:00 a.m. (hypothetical example).

Despite the purposeful exaggeration of the second example, the question remains regarding what information should be included when considering the impact of base rates on likelihood estimations. Miettinen and Caro (1994) point to the central issue concerning “the separation of the totality of known facts about the patient into those that bear on the ad hoc prior probabilities and those that are involved in the science-based updating of these” (p. 202). They describe three junctures at which Bayes’ theorem can be applied: the nonclinical/clinical juncture (i.e., the division between “nonclinical” patient information such as demographics versus clinical information obtained from symptom report and clinical tests), the pretest/posttest juncture (i.e., the division between all information, both nonclinical and clinical, versus the results of a single specific test), and the ignorance/facts juncture (i.e., the division between no information except the time and place at which the patient presents versus all relevant facts obtained in addition to this). The base rate of the disorder in question would be different at each of these junctures. Miettinen and Caro describe a variety of additional problems inherent in the delineation of the parameters involved at any of these junctures and conclude that the quantification of these parameters pose insurmountable epistemologic challenges. A related weakness of the present study is that although the disorder base rate was specified in the PPV problems, one could argue that it was unknown whether the test used in the problems was given to all clinic patients. In other words, if the test in the problem was given to only those patients for whom dementia was a reasonable rule out, then the base rate to be used in determining predictive value would be greater than 10%.

Although Miettinen and Caro’s concerns should not be ignored, clearly, one cannot disregard base rate considerations altogether, and it would seem incumbent upon the clinical decision-maker to consider how base rates impact the interpretation of diagnostic tests using
the best base rate estimates available. Elwood (1993) concludes that appropriate base rate estimates could be local or setting specific, based on epidemiological or clinical studies of similar samples, or even based on hospital or clinic records or reanalysis of existing research data. Rorer and Dawes (1982) suggest that reference groups be as narrow as possible, provided the estimates upon which they are based are stable.

4.2. Future directions

A simple but important goal to which every neuropsychologist can contribute is the collection of symptom and disorder base rates. This information would clearly benefit the clinicians and researchers within the setting in which the data are collected, and the publication of these data would allow others to benefit, as well. To this end, McCaffrey, Palav, Labarge, and Bryant (in preparation) have compiled a reference manual of published symptom base rates across a wide variety of psychological, neuropsychological, and neurologic disorders. An additional step along these lines would be the establishment of a national symptom and disorder base rate database providing base rate information across a wide range of settings and demographics. Sackett et al. (1997) describe an internet website in which they have begun to collect medical base rate data as well as a list of tests with high sensitivity and/or specificity to be used in ruling out or ruling in the diagnoses, respectively. A similar site for neuropsychologically relevant information would be valuable.

Acknowledgments

This article is based on the doctoral dissertation research of the first author, who would like to express great appreciation to chairperson Robert J. McCaffrey. Additional thanks are extended to committee members William L. Simmons and Jerid M. Fisher.

Appendix A. Survey questions

(1) The base rate of dementia among patients at an adult neuropsychology clinic is 10%. A 75-year-old male presents at the clinic with complaints of fatigue and sleep disturbance. In the absence of additional information, what is the probability that this individual has dementia? __ %. (Answer: 10%)

(2) The “sensitivity” of a test refers to:
(a) the probability that the test result will be positive when the disorder is present
(b) the probability that the test result will be positive when the disorder is not present
(c) the probability that the test result will be negative when the disorder is present
(d) the probability that the test result will be negative when the disorder is not present
(Answer: a)

(3) The “specificity” of a test refers to:
(a) the probability that the test result will be positive when the disorder is present
(b) the probability that the test result will be positive when the disorder is not present
(c) the probability that the test result will be negative when the disorder is present
(d) the probability that the test result will be negative when the disorder is not present
(Answer: d)

(4a) The base rate of dementia among patients at an adult neuropsychology clinic is 10%. The dementia screening used in the clinic has a sensitivity of 80% and a specificity of 90%. What is the likelihood that an individual evaluated at this clinic who obtains a positive dementia screening score actually has dementia? ___% (Answer: 47%)

(4b) Ten out of every 100 individuals seen in an adult neuropsychology clinic have dementia. Of these 10 individuals with dementia, 8 will have a positive score on a dementia screening. Of the remaining 90 individuals without dementia, 9 will still have a positive dementia screening score. In a sample of 100 individuals from this clinic, how many of the individuals who have a positive dementia screening score will actually have dementia? ___ out of ___ , or ___% (Answer: 8 out of 17 or 47%)

Appendix B. Two-by-two contingency table illustrating sensitivity, specificity, and PPVs using values from Questions 4a and 4b in Appendix A

<table>
<thead>
<tr>
<th>Test</th>
<th>Disease</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>8 a</td>
<td>9 b</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>2 c</td>
<td>81 d</td>
<td></td>
</tr>
</tbody>
</table>

\[ N = a + b + c + d = 100. \text{BR} = \frac{a + c}{a + b + c + d} = 10\%. \text{Sensitivity} = \frac{a}{a + c} = 80\%. \text{Specificity} = \frac{d}{b + d} = 90\%. \text{PPV} = \frac{a}{a + b} = 47\%. \text{NPV} = \frac{d}{c + d} = 98\%. \]

References


Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In: G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 129–161). New York: Wiley.


