Is It Time to Rethink How Neuropsychological Tests Are Used to Diagnose Mild Forms of HIV-Associated Neurocognitive Disorders? Impact of False-Positive Rates on Prevalence and Power

Ana-Claire L. Meyer, W. John Boscardin, Judith K. Kwasa, Richard W. Price

Department of Neurology, San Francisco General Hospital, and Division of Geriatrics and Division of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, Calif., USA; Center for Microbiology Research, Kenya Medical Research Institute, Nairobi, Kenya

Abstract

Background: Between 0 and 48% of normal HIV-uninfected individuals score below threshold neuropsychological test scores for HIV-associated neurocognitive disorders (HAND) or are false positives. There has been little effort to understand the effect of varied interpretations of research criteria for HAND on false-positive frequencies, prevalence and analytic estimates.

Methods: The proportion of normal individuals scoring below Z score thresholds drawn from research criteria for HAND, or false-positive frequencies, was estimated in a normal Kenyan population and a simulated normal population using varied interpretations of research criteria for HAND. We calculated the impact of false-positive frequencies on prevalence estimates and statistical power.

Results: False-positive frequencies of 2–74% were observed for asymptomatic neurocognitive impairment/mild neurocognitive disorder and 0–8% for HIV-associated dementia. False-positive frequencies depended on the definition of an abnormal cognitive domain, Z score thresholds and neuropsychological battery size. Misclassification led to clinically important overestimation of prevalence and dramatic decreases in power.

Conclusions: Minimizing false-positive frequencies is critical to decrease bias in prevalence estimates and minimize reductions in power in studies of association, particularly for mild forms of HAND. We recommend changing the Z score threshold to ≤ –1.5 for mild impairment, limiting analysis to 3–5 cognitive domains and using the average Z score to define an abnormal domain.

Key Words
Africa · HIV · Dementia · Prevalence · Power

Introduction

Considerable debate surrounds the diagnosis of mild forms of HIV-associated neurocognitive disorders (HAND) [1, 2]. Direct validation of the criteria for asymptomatic neurocognitive impairment (ANI) and mild neurocognitive disorder (MND), which rely largely or exclusively on neuropsychological testing, has been challenging (online suppl. appendix 1; for all online suppl. material, see www.karger.com/doi/10.1159/000354629) [3]. There are no adequately powered longitudinal clinical-pathological correlation studies [4, 5], and there is no gold standard antemortem biomarker or imaging finding.

The prevalence of ANI/MND ranges from 26 to 76% and that of HIV-associated dementia (HAD) from 1 to
35%. Studies of neuropsychological batteries used for HAND in normal HIV-uninfected populations suggest this heterogeneity may be due, at least in part, to measurement error [6–12]. For example, 15–22% of an HIV-uninfected control group [6] and 20% of a simulated normal population [7] had neuropsychological test scores below the threshold for HAND, which we call false-positive cases. In addition, 4–23% of a normal population had Z scores ≤ –1 and 0–6% had Z scores ≤ –2 on individual tests [Cettomai et al., unpubl. data]. While it has been argued that a 15% false-positive frequency is acceptable for neuropsychological tests [8], the impact of this measurement error on prevalence and analytic estimates for HAND has not been explored quantitatively.

Furthermore, neuropsychological batteries will have higher false-positive frequencies than individual tests because they involve multiple comparisons. Multiple comparisons occur when one compares two groups using multiple outcomes and attention is paid to the 'strongest' differences. Here, the more comparisons made, the more likely one is to find a 'strong difference', or statistically significant outcome [9]. With neuropsychological batteries, when a battery of many tests is administered and scores are compared to a normal population (multiple outcomes), and diagnoses are based on the most abnormal Z scores (the strongest differences), the more tests that are performed, the more likely it is to find 2 abnormal Z scores (i.e. diagnose a normal individual as impaired).

False-positive cases will lead to biased prevalence estimates and reductions in power for analytical estimates. A false positive on a diagnostic test, here a battery of neuropsychological tests, can lead to a form of measurement error called nondifferential misclassification [10]. Nondifferential misclassification arises from errors in classification which occur across all levels of the variable in question. For example, in a randomized clinical trial of a new medication for HAND, the treatment and control groups would have an equal probability of having false-positive cases assuming the only contributor to false-positive cases was the neuropsychological tests. In general, nondifferential misclassification will bias tests of association to the null [10, 11]. Misclassification does not affect the validity of statistical tests (type I error) but can drastically reduce their power (type II error) [12]. Using the example above, nondifferential misclassification will not lead investigators to falsely conclude a new treatment for HAND is effective, but it could lead investigators to falsely conclude a new treatment is not effective, when in truth it is effective. Thus, misclassification, or false-positive diagnoses of ANI/MND, could potentially affect a broad range of research including studies of prevalence, biomarkers, functional imaging, risk factors or the effect of treatments.

Thus, in this study it is our goal to (1) estimate misclassification (the false-positive frequency) on a neuropsychological battery for HAND using empirical and theoretical methods, (2) explore the impact of multiple comparisons (neuropsychological battery size) on the false-positive frequency, (3) explore the impact of the false-positive frequency on prevalence estimates and (4) explore the impact of the false-positive frequency on power in studies of association.

### Methods

**Misclassification: Empirical Estimation of the False-Positive Frequency**

Between October 2009 and March 2010, we enrolled 84 HIV-uninfected Kenyan adults without neurologic comorbidities, depression, substance abuse or a suspected learning disorder. Details of the population and methods can be found in our prior work [Cettomai et al., unpubl. data]. The neuropsychological test battery covered 7 cognitive domains recommended by the Frascati criteria [3] and contained a total of 17 tests which were selected to parallel the neuropsychological battery used in the CNS HIV Anti-Retroviral Therapy Effects Research (CHARTER) study [13]. We determined the proportion of normal individuals who met neuropsychological criteria for HAND using the Frascati criteria (online suppl. appendix 1) [3]: (1) ANI/MND, 2 domains ≤ 1 standard deviation (SD) below the mean, and (2) HAD, 2 domains ≤ 2 SD below the mean. Because the Frascati criteria do not define an 'abnormal domain', we explored 4 different definitions of an abnormal domain, as follows: (1) 1 abnormal test in a domain; (2) 2 abnormal tests in a domain; (3) all tests in a domain must be abnormal, and (4) the average of all test scores across a domain was abnormal. Z scores were truncated at ± 3 for the purpose of calculating averages to avoid undue influence of a single extreme test score.

**Misclassification: Theoretical Estimation of the False-Positive Frequency**

The false-positive frequency for a single neuropsychological test in a normal population is based largely on the statistical properties of the norming process (online suppl. appendix 2) [8]. Raw scores on neuropsychological tests are converted to Z scores based on representative population data [14, 15]. The distribution of Z scores should theoretically be normal [14, 15]. Thus, 15.9% of a normal population should have Z scores ≤ –1, 6.7% should have Z scores ≤ –1.5 and 2.2% should have Z scores ≤ –2. If neuropsychological test scores are independent of one another, simple multiplication could be used to calculate the expected false-positive frequency for a battery, akin to calculating the probability of two heads in a coin toss. However, neuropsychological test scores are correlated both within and between cognitive domains, which makes such calculations more complex.

Thus, we simulated a normal population using the structure of the neuropsychological test battery described above (7 cognitive
domains, 17 individual test scores). We assumed correlated normal distributions around a true average score of 0. Z scores were truncated at ±3 for the purpose of calculating averages. We explored a range of correlation structures both within and between domains and reported results for low correlation (0 between, 0 within), moderate correlation (0.5 between, 0.5 within) and high correlation (0.9 between, 0.9 within). We also analyzed results using an empirical correlation structure derived from the actual between- and within-domain correlation derived from our Kenyan sample described above. We estimated the proportion of normal individuals who would meet criteria for HAND using the previously mentioned definitions of an abnormal domain.

**Multiple Comparisons: The Effect of Neuropsychological Battery Size on the False-Positive Frequency**

We also explored the impact of different numbers of domains and tests per domain on the false-positive frequency. We varied the number of domains from 2 to 10 and the number of tests per domain from 2 to 5. To minimize the effect of test number, we used the average scores across a domain to define an abnormal domain. We performed simulations using 2 correlation structures. The first was derived from the average correlation within and between domains in the Kenyan sample (0.31 within, 0.17 between). The second was the moderate correlation structure, which was chosen as its performance was similar to that observed in the Kenyan sample (0.5 within, 0.5 between).

Simulations were performed in R (version 2.13.2, Vienna, Austria), using 100,000 repetitions, resulting in a Monte Carlo error of ±0.2–0.4% for the estimated proportions.

**Exploring the Impact of the False-Positive Frequency on Prevalence and Power**

This analysis was based on a classic paper by Bross [12] which has had wide influence over time [10, 16]. Bross derives formulas for the direct calculation of (1) bias in prevalence estimates and (2) effective sample size (online suppl. appendix 3). Reduction in power can be expressed practically as a reduction in effective sample size [12]. For example, in a study of 200 individuals, the loss in power due to misclassification might be equivalent to a ‘loss’ of 20 individuals, leaving an effective sample size of 180. The effective sample size is a function of the false-positive frequency, false-negative frequency, the true prevalence of disease and the actual sample size. We investigated a range of estimates relevant to the diagnoses of HAND. We used prevalence rates of 2, 10, 25 and 40%, chosen to be representative of HAD, MND and ANI (the latter 2 only for ANI) [13, 17]. We used false-positive frequencies in the ranges observed in our prior analyses (1–40%). We assumed a false-negative frequency of 5%. Sensitivity analyses were performed with a false-negative frequency of 25% (data not shown).

**Ethics Statement**

The protocol was approved by the University of California San Francisco Committee on Human Research and the National Ethical Review Committee of the Kenya Medical Research Institute. Written informed consent was obtained from each participant.

**Role of the Funding Source**

The funders did not have any role in study design, collection, analysis or interpretation of data, writing the report or decision to publish.

---

**Results**

**Misclassification: Empirical Estimation of the False-Positive Frequency**

Between 8 and 48% of normal Kenyan adults met the criteria for ANI/MND (Z scores ≤−1 SD in ≥2 domains; fig. 1; table 1 in online suppl. appendix 4). The highest false-positive frequencies were obtained when 1 abnormal test score defined an abnormal domain. The lowest false-positive frequencies were obtained when all the test scores in a domain were required to be abnormal. False-positive frequencies obtained using the average score or 2 abnormal tests per domain were in the middle range. A more stringent definition for mild impairment (Z scores ≤−1.5 in ≥2 domains) resulted in lower false-positive frequencies. With the exception of the most permissive criteria (1 abnormal test score), no normal individuals met the criteria for HAD (Z scores ≤−2 in ≥2 domains).

**Misclassification: Theoretical Estimation of the False-Positive Frequency**

False-positive frequencies obtained using a simulated normal population were similar to those found in the normal Kenyan adults described previously, particularly those using moderate correlation or the empirical correlation structure derived from the Kenyan population (fig. 1; table 2 in online suppl. appendix 4). Depending on the correlation structure and the definition of an abnormal domain, false-positive frequencies for ANI/MND (Z scores ≤−1) ranged from 2 to 74%. Again, the highest false-positive frequencies were observed when 1 abnormal test defined an abnormal domain, and the lowest when all tests were required to be abnormal. False-positive frequencies for average scores and 2 abnormal tests were again in the middle range. Using more stringent criteria for mild impairment (Z scores ≤−1.5), false-positive frequencies ranged from 0 to 29%, and using criteria for HAD (Z scores ≤−2), false-positive frequencies ranged from 0 to 8%.

**Multiple Comparisons: The Effect of Neuropsychological Battery Size on the False-Positive Frequency**

As expected, the larger the number of domains, the higher the false-positive frequency for each of the Z score thresholds (fig. 2; table 3 in online suppl. appendix 4). When averaging the scores across a domain, the number of tests per domain had only small effects on the false-positive frequency. As with prior analyses, the highest false-positive frequencies were observed for ANI/MND
(Z scores ≤ −1) and ranged from 4 to 41%. Stringent criteria for mild impairment (Z scores ≤ −1.5) resulted in false-positive frequencies of 1–16%, and criteria for HAD (Z scores ≤ −2) resulted in false-positive frequencies of 0–5%.

Exploring the Impact of the False-Positive Frequency on Prevalence

Higher false-positive frequencies led to greater bias in measured prevalence, or greater overestimation of prevalence (fig. 3; table 1 in online suppl. appendix 5). For example, a false-positive frequency of 1% with a true prevalence of disease of 2% (like the prevalence of HAD) would lead to an estimated or observed prevalence frequency of 3%, or a bias of 1%. The differences grew with higher false-positive frequencies such that a false-positive frequency of 20% with a true prevalence of 2% led to a measured prevalence of 22% (bias of 20%). However, the magnitude of the bias was attenuated with higher true prevalence. Thus, when the true prevalence was 10% (as for MND), the measured prevalence was 28% (bias of 18%). Similarly, with a true prevalence of 25% (as for ANI), the measured prevalence was 39% (bias of 14%).

Exploring the Impact of the False-Positive Frequency on Power

Even low false-positive frequencies had striking effects on power when the true prevalence of disease was low (2%), as for HAD (fig. 4; table 2 in online suppl. appendix 5). Here, a false-positive frequency as low as 1% had large effects on effective sample size, reducing an actual sample size of 100 to an effective sample size of 55. This reduction was magnified for larger sample sizes; an actual sample size of 1,000 was reduced to 186. Similarly dramatic effects were observed with false-positive frequencies typical of those observed for ANI/MND. A false-positive frequency of 20% and true prevalence of 10% (as for MND) reduced an actual sample size of 100 to 3 and one of 1,000 to 10. The same false-positive frequency of 20% and a higher true prevalence of 25% (as for ANI) led to comparable reductions from 100 to 9 and from 1,000 to 10. Finally, a false-positive frequency of 40% and a true prevalence of 25% (as for the highest estimate for ANI) reduced sample sizes from 100 to 2 and from 1,000 to 2.
Discussion

Current research criteria for the diagnosis of ANI/MND result in high false-positive frequencies, but this is not the case for HAD. The high false-positive frequencies are due to a combination of nondifferential misclassification and multiple comparisons, which are intrinsic properties of neuropsychological batteries and the scoring process. High false-positive frequencies lead to unacceptably high bias in prevalence estimates and a dramatic loss in power. Thus, it is critical to minimize the false-positive frequency for neuropsychological batteries. We propose striving for a false-positive frequency of less than 5% for ANI/MND, leading to moderate reductions in power at a true prevalence of 25% (from 100 to 72 and 1,000 to 227). Further, we propose striving for a false-positive frequency of less than 1% for HAD, to minimize the reductions...
Fig. 3. Measured prevalence as a function of the false-positive rate and true prevalence (assumes a false-negative rate of 5%).

Fig. 4. Effective sample size as a function of the false-positive rate, true prevalence of disease and actual sample size. a Sample size: n = 100. b Sample size: n = 1,000 (assumes a false-negative rate of 5%).
in power. Below, we make specific recommendations about the structure of neuropsychological batteries and threshold scores that will achieve these false-positive frequencies.

**Minimizing Misclassification**

High false-positive frequencies were observed in actual and simulated normal populations and varied depending on the definition of an abnormal cognitive domain. In addition, false-positive frequencies using a threshold Z score of ≤ −1 were unacceptably high regardless of the parameters used. Thus, we recommend changing the threshold Z score for ANI/MND to ≤ −1.5. A threshold Z score of ≤ −2 is acceptable for HAD. Further, we generally recommend using a definition for an abnormal cognitive domain that minimizes the false-positive frequency, though other considerations for the optimal definition are discussed in the sections that follow.

**Minimizing Multiple Comparisons**

In general, more tests per domain and a greater number of domains result in higher false-positive frequencies. Thus, to minimize the effect of test number on the false-positive frequency, we recommend defining an abnormal domain as the average Z score across that domain, even though this definition did not lead to the lowest false-positive frequencies we observed. This would also enable flexibility in designing neuropsychological batteries for different purposes and help comparability across studies. In addition, we recommend grouping individual tests into no more than 3–5 cognitive domains.

**Parallels with Mild Cognitive Impairment due to Alzheimer’s Disease**

Like HAND, the diagnosis of mild cognitive impairment (MCI) due to early Alzheimer’s disease rests primarily on neuropsychological testing with a threshold Z score between −1 and −1.5 [18, 19], and only a few small clinical-pathologic correlation studies exist [20]. Depending on the MCI definition used, there is substantial variation in the prevalence from 4 to 42% [17]. There are high false-positive frequencies for MCI; nearly half of healthy older adults will have borderline test scores (≤ −1.3 SD) [21], and 0–26% will meet criteria for possible or probable memory impairment [22]. Furthermore, in longitudinal studies of MCI, the positive predictive value for Alzheimer’s dementia varies from 5 to 48%, and frequencies of reversion to normal cognition range from 9 to 67% [17]. While the empiric data used for this analysis focused on HAND, the results of the simulations are largely independent of the demographic characteristics of the normal population or the tests comprising the neuropsychological battery. Thus, the results of this study are also likely applicable to the diagnostic criteria for MCI. It is critical to recognize that both MCI and ANI/MND are clinical classifications which may or may not be the phenotypic manifestations of the presumed underlying pathology, namely Alzheimer’s disease and HAD, respectively. Ultimately, identifying biomarkers and/or imaging findings will likely be necessary components of the definitive diagnosis of these disorders.

**Limitations**

The first major limitation of this study is that the empiric data are derived from one small population in western Kenya. However, our estimates of false-positive frequencies are comparable to those from much larger studies [6] as well as the simulations presented. The second major limitation is that this study focused only on the false-positive frequency due to criteria for neuropsychological testing. Although the definition of ANI is based exclusively on neuropsychological testing, the definitions of MND and HAD additionally require symptoms of cognitive impairment and/or impairment of function [19]. Thus, the false-positive frequencies obtained using the full criteria will likely be lower than those obtained from neuropsychological testing alone. Furthermore, this analysis does not address false-positive diagnoses from ‘confounding conditions’, such as substance abuse or depression, which can impair test performance but are unrelated to the direct effects of HIV infection [13].

While a neuropsychological battery that is both highly sensitive and specific would be ideal, this is difficult to achieve. When we arbitrarily define a cutoff point in our test (i.e. choose a Z score), specificity (low frequency of false positives) can only be improved at the expense of sensitivity [23]. As described previously, we cannot directly calculate the sensitivity of the neuropsychological batteries for HAND, because there is no antemortem gold standard biomarker or imaging finding. Thus, lowering the Z score threshold to −1.5 to improve specificity may decrease sensitivity, though it is difficult to estimate by how much. In addition, it will be important to explore in detail the statistical properties of other summary measures of neuropsychological performance such as the Global Deficit Score [24] or the average Z score and the effects of misclassification on other epidemiological parameters such as incidence.
Implications for Current Research

This analysis demonstrates that using the Frascati definition of ANI/MND results in a high frequency of false-positive diagnoses and that these false-positive diagnoses lead us to overestimate the prevalence of these disorders. Furthermore, this analysis found that including individuals with the Frascati definition of ANI/MND in studies of association dramatically decreases power. This decrease in power is applicable across study types; studies of biomarkers, functional imaging, risk factors or treatment effects will all be affected. Ultimately, this decrease in power will lead investigators to falsely conclude there is no association between HAND and a particular biomarker, imaging finding, risk factor or treatment when in fact there is a true association. Thus, there is an urgent need to both revisit the Frascati criteria and possibly reanalyze studies which include individuals with ANI/MND as part of HAND.

Conclusion

Current research criteria for the diagnosis of ANI/MND have high false-positive frequencies, while false-positive frequencies for HAD are low. False-positive frequencies are a result of Z score thresholds, neuropsychological battery size (number of tests and domains) and the definition of an abnormal cognitive domain. False-positive frequencies in the ranges we have described have dramatic impacts on statistical power and prevalence estimates for ANI/MND as well as important impacts on statistical power for HAD.

Minimizing false-positive frequencies is critical to decrease bias in prevalence estimates and decrease type II error in studies of association. We recommend (1) changing the Z score threshold to ≤−1.5 for ANI/MND, (2) limiting analysis of neuropsychological batteries to 3−5 cognitive domains and (3) using the average Z score to define an abnormal domain. Additional research to determine the impact of these changes on the sensitivity of these criteria is essential. Finally, it is critical to further explore the relevance of these findings to disorders like MCI, which use similar diagnostic criteria.

Acknowledgements

We are grateful to Jeff Martin, University of California San Francisco, for his critical review and comments on the manuscript. This study was supported by the American Academy of Neurology Practice Research Training Fellowship (A.-C.L. Meyer), the Fogarty International Center of the National Institutes of Health (K01TW008764), the Fogarty International Clinical Research Fellowship (5 R24 TW00798, 3 R24 TW00798-02S1; A.-C.L. Meyer, J.K. Kwasa) from the National Institutes of Health, Fogarty International Center through Vanderbilt University, the National Cancer Institute, the National Institute on Drug Abuse, Office of the Director, National Institutes of Health and the National Institute of Mental Health. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This paper is published with the permission of the Director of the Kenya Medical Research Institute.

Disclosure Statement

None of the authors has any conflicts of interest to report.

References