Published in final edited form as:
Psychol Assess. 2009 June ; 21(2): 235–239. doi:10.1037/a0015686.

Prevalence estimation and validation of new instruments in psychiatric research: An application of latent class analysis and sensitivity analysis

Brian Wells Pence[1,2,3,4], **William C. Miller**[4,5], and **Bradley N. Gaynes**[4,6]

[1] Department of Community and Family Medicine, Duke University

[2] Duke Global Health Institute, Duke University

[3] Health Inequalities Program, Center for Health Policy, Duke University

[4] Department of Epidemiology, School of Public Health, the University of North Carolina at Chapel Hill

[5] Division of Infectious Diseases, Department of Medicine, School of Medicine, the University of North Carolina at Chapel Hill

[6] Department of Psychiatry, School of Medicine, the University of North Carolina at Chapel Hill

Abstract

Prevalence and validation studies rely on imperfect reference standard (RS) diagnostic instruments which can bias prevalence and test characteristic estimates. We illustrate two methods to account for RS misclassification. Latent class analysis (LCA) combines information from multiple imperfect measures of an unmeasurable "latent" condition to estimate sensitivity (Se) and specificity (Sp) of each measure. Simple algebraic sensitivity analysis (SA) uses researcher-specified RS misclassification rates to correct prevalence and test characteristic estimates, and can succinctly summarize a range of scenarios with Monte Carlo simulation. We applied LCA to a validation study of a new substance use disorder (SUD) screener and a larger prevalence study. A traditional validation study analysis that assumed an error-free RS (SCID) estimated the screener had 86% Se/75% Sp. Validation study estimates from LCA were 91% Se/81% Sp (screener) and 73% Se/98% Sp (SCID). SA in the prevalence study suggested the prevalence of SUD was underestimated by 22% by assuming the SCID to be error-free. LCA and SA can assist investigators in relaxing the unrealistic assumption of perfect RSs in reporting prevalence and validation study results.

Correspondence concerning this article (and requests for an extended report of this study including a more detailed algebraic presentation of latent class analysis and sensitivity analysis) should be addressed to Brian Wells Pence PhD, Center for Health Policy, Box 90253 Duke University, Durham NC 27708 USA or bpence@aya.yale.edu.

Keywords

Methods; psychiatric measurement; prevalence; misclassification bias; latent class analysis; sensitivity analysis

Introduction

Definitive measurement of psychiatric illness remains a challenge despite extensive research (Kendell & Jablensky, 2003; McHugh, 2005). The development of the *Diagnostic and Statistical Manuals for Mental Disorders* (DSM) has introduced a new level of standardization in the definition of psychiatric diagnoses. The DSM generally specifies a set of symptoms and conditions as both necessary and sufficient for the assignment of a given diagnosis. Nevertheless, the current state of psychiatric diagnosis remains a process of attempting to measure an unobservable "true" condition, in which varying diagnostic strategies attempt to define a diagnosis with overlapping, but not identical, results (Kessler et al., 2004; Paykel, 2002).

The DSM framework has spurred the development of standardized diagnostic tools designed to generate valid, reliable, and consistent diagnoses in research settings, including the Structured Clinical Interview for DSM (SCID) and the Composite International Diagnostic Interview (CIDI) (First, Spitzer, Gibbon, & Williams, 1990; Robins et al., 1988). Diagnostic assignment based on instruments such as the SCID or CIDI is frequently used as the "gold standard" both in prevalence studies and in validation studies that estimate the sensitivity and specificity of new screening tools. Yet the inter-rater and test-retest reliability of such instruments remain in the range of 0.70–0.90 for most diagnoses, indicating substantial misclassification error (Steiner, Tebes, Sledge, & Walker, 1995).

The use of imperfect reference standards (RS) leads to bias in two of the primary goals of psychiatric measurement: prevalence estimation and validation of new screening instruments. Our goal here is to demonstrate and discuss the application of latent class analysis (LCA) and standard sensitivity analysis (SA) to the problem of psychiatric measurement. LCA is an analytical technique that incorporates information from multiple (usually ≥3) imperfect measures of a "latent" (not directly measurable) condition to estimate the prevalence of the latent condition as well as the sensitivity and specificity of each of the imperfect measures (Walter & Irwig, 1988). The latent class model combines the information from these imperfect measures in order to "triangulate" on the unobserved true condition. The latent class model generally requires an important underlying assumption, namely, that classification errors (i.e., false positive and false negative probabilities) are independent between the tests.

An alternative approach to adjusting test characteristics and prevalence estimates for an imperfect RS is simple sensitivity analysis (SA). In this approach, the investigator hypothesizes what the error rates of the imperfect RS might be and then uses simple algebra (Rothman & Greenland, 1998) to back-calculate the "true" two-by-two table (screening test by true status) that is consistent with the observed table (screening test by RS) and the assumed RS error probabilities. Corrected test characteristics and

prevalence can then be directly calculated from the "true" table. Normally the RS error probabilities are not known, and the investigator should consider a set of different scenarios (e.g.: assume RS sensitivity ranges from 80–95%). Recent extensions incorporate Monte Carlo methods to specify a distribution rather than a single value for the error probabilities and produce a 95% "simulation interval" to succinctly summarize the uncertainty in prevalence or test characteristic estimates associated with both random and systematic error (Fox, Lash, & Greenland, 2005).

In this paper, we demonstrate how the estimated test characteristics of a psychiatric screening instrument change depending on whether a traditional analysis (i.e., that assumes the RS to be error-free), LCA, or SA is applied. While LCA has been applied previously in psychiatric research, formal consideration of the role of RS misclassification in reports of prevalence or test characteristics remains the exception rather than the norm. Approaches such as LCA and SA have a useful role in relaxing the strong and untenable assumption of an error-free RS in psychiatric research.

Methods

We first consider a validation study of a new screening instrument for substance use disorders (SUD) in HIV+ patients (Pence et al., 2005). In brief, the new test is a 16-item self-report screening instrument designed to identify substance abuse, mood, and anxiety disorders in HIV+ patients (Substance Abuse and Mental Illness Symptoms Screener — SAMISS); we focus here on the 7-item SUD module. The validation study population comprised all new patients and patients never previously screened at the Infectious Diseases Clinic of an academic medical center who were HIV+, ≥18 years old, English-speaking, and mentally competent. Consenting patients completed the SAMISS with a clinic social worker or research assistant and then were administered the SCID for DSM-IV by a psychiatry research interviewer blinded to the participants' SAMISS responses. All SCID diagnoses were reviewed and confirmed by a psychiatrist (BNG). Separately, a trained chart abstractor blinded to SCID and SAMISS results reviewed participants' medical records to identify SUD diagnoses noted by the ID physician in the year before enrollment. After complete description of the study to the subjects, written informed consent was obtained. All study procedures were approved by the Institutional Review Board of the University of North Carolina at Chapel Hill.

Following standard scoring instructions, participants were considered to have a positive SAMISS screen for a probable SUD if they endorsed problematic frequency or quantity of alcohol consumption, endorsed drug use at least weekly, or endorsed any perceived problematic use of alcohol or drugs (Pence et al., 2005). We first calculated standard test characteristics with exact 95% CIs for the SAMISS (positive vs. negative) relative to the SCID as RS (any vs. no past-year SUD). To relax the assumption of an error-free RS, we then fit a latent class model using the DOS program "Latent.exe" (developed by Steven Walter, provided via personal communication, walter@mcmaster.ca) with three different SUD measures (SAMISS, SCID, and chart review — any vs. no noted SUD in past year) to

estimate the test characteristics and maximum likelihood CIs for each measure and the prevalence of SUD in the sample.

Finally, we used SA to calculate the expected test characteristics of the SAMISS under a set of specific assumptions about SCID error probabilities. We assumed that the SCID would have higher specificity than sensitivity, as the relatively stringent criteria required for most DSM diagnoses are more likely to lead to a case being overlooked than a noncase being classified as having a diagnosis. We considered scenarios in which the sensitivity of the SCID ranged from 70-90% and the specificity ranged from 96-100%. (Specificity<96% produced negative cell counts in one cell, indicating that the observed data were not consistent with SCID specificity<96%). CIs for test characteristic SA estimates were calculated using normal approximation theory. In all analyses reported here, we excluded 2 individuals who lacked chart review information.

As a second example, we consider a study designed to estimate the prevalence of DSM-IV-defined SUD in the entire patient population of the same clinic (Pence, Miller, Whetten, Eron, & Gaynes, 2006). From 2000-2002, 1,319 patients were seen in clinic who were HIV+, ≥18 years old, English-speaking, and mentally competent. Of these, 1,227 (93%) completed SAMISS screening as part of standard clinical care, and 1,125 (92% of those screened) gave informed consent for their medical information to be captured in a research database and were included in the analysis. In the SAMISS validation sample described earlier, which included both SAMISS and SCID results, we developed and assessed a logistic regression model predicting the presence of a SCID SUD diagnosis in the past year, using individual SAMISS responses and other sociodemographic and clinical variables as predictors. The final predictive model of SCID SUD diagnoses included 4 SAMISS questions: heavy alcohol use, use of drugs at least weekly, inability to cut back on alcohol or drug use, and being worried or anxious for a month or more. The area under the receiver operating characteristic curve for this model was 0.92, indicating excellent discriminative ability. The coefficients from this final model were then applied to the full sample of 1,125 clinic patients with completed SAMISS screens and were used to calculate a predicted probability of a SCID-defined SUD diagnosis for each patient, ranging from 0 to 1. These probabilities were summed across the sample to estimate the prevalence of SCID-defined SUD diagnoses in the sample, and a 95% confidence interval was calculated by bootstrapping.

In the present analysis, we consider the potential bias in this prevalence estimate introduced by measurement error in the SCID. We considered values for SCID sensitivity and specificity ranging from the lower to the upper bounds of the 95% CI of these quantities from the latent class model estimated in the validation study sample. Using simple algebra, we back-calculated the expected "true" prevalence given the observed prevalence and the various combinations of SCID sensitivity and specificity.

A list of corrected estimates from a range of different bias scenarios can be challenging to synthesize. We calculated a single summary corrected estimate with a 95% simulation interval using Monte Carlo methods with 10,000 iterations as previously described (Fox et al., 2005). At each iteration, we randomly selected a sensitivity and specificity from triangular

probability distributions defined by the LCA lower 95% CI bound, MLE, and upper 95% CI bound and back-calculated the expected "true" prevalence corresponding to those values. To incorporate random error into the distribution of corrected estimates, we randomly drew a deviate from a normal distribution with mean 0 and standard deviation (SD) equal to the SD of the prevalence estimate in the original analysis. We report the median of the distribution of the 10,000 corrected estimates as our best estimate of the true prevalence and the 2.5^{th} and 97.5^{th} percentiles as the bounds of a 95% simulation interval that reflects our uncertainty around this estimate due to both random and systematic error.

Results

In the validation study sample, 20% of participants had a SCID-defined SUD diagnosis, 22% had a SUD diagnosis noted in the chart, and 38% received a positive screen on the SAMISS for a probable SUD in the past year (Table 1). In the original analysis, assuming the SCID to have perfect sensitivity and specificity, we estimated that the SAMISS had 86% sensitivity and 75% specificity in identifying individuals with SUD (Table 2). From the latent class model, the maximum likelihood estimates of the SAMISS sensitivity and specificity were 91% (95% CI 80-100%) and 81% (73-89%), respectively; the CIs around these estimates encompassed the estimates from the original analysis. The SCID was estimated to have 73% sensitivity and 98% specificity in identifying individuals with SUD, and chart review was estimated to have 85% sensitivity and 99% specificity. The estimated prevalence of past-year SUD from LCA was 25% (17-34%). As this model was fully saturated, there were no excess degrees of freedom with which to estimate goodness-of-fit statistics.

We next applied simple SA. If the SCID had 70-90% sensitivity and 100% specificity, SAMISS sensitivity would be unaffected and specificity would be slightly underestimated by treating the SCID as the gold standard, and CIs for all estimates include the original estimates. If the SCID had 70-90% sensitivity and 96% specificity, SAMISS sensitivity would be substantially underestimated (with CIs that do not overlap with the Ci of the original estimate) and specificity would be slightly underestimated by treating the SCID as the reference standard (Table 2).

In the full clinic sample, the estimated prevalence of SCID-defined SUD was 21% (95% CI: 19-23%) (Table 3, row 1). From the latent class model above, we estimated that the SCID had 73% sensitivity and 98% specificity in identifying individuals with SUD. Using simple SA, the expected true prevalence based on these test characteristics was 27% (Table 3, row 2). The corrected prevalence ranged from 22-35% when we varied the sensitivity and specificity values over the ranges specified by the 95% CI estimates from the latent class model (56-89% and 96-100%, respectively). All considered scenarios suggested that prevalence was underestimated by between 5 and 40 percent by treating the SCID as error-free. The corrected prevalence estimate was more affected by varying the sensitivity than the specificity. Finally, from the Monte Carlo simulation, the median corrected prevalence estimate was 27% with a 95% simulation interval of 22-33%, suggesting the traditional analysis likely underestimated prevalence by nearly 25%.

Discussion

In psychiatric research, prevalence studies as well as validation studies of new psychometric instruments nearly invariably rely on RSs that are themselves subject to misclassification error. This reality produces bias in the estimates of prevalence and test characteristics. Simple SA and LCA represent two of several approaches that have been proposed to assess the potential magnitude of the resulting bias and to produce more valid estimates. In the present example, the results of both LCA and SA support the qualitative conclusion that the traditional analysis, which assumed an error-free SCID, somewhat underestimated the sensitivity and specificity of the SAMISS and the overall prevalence of SUD.

Use of LCA removes one strong assumption — that the RS is error-free — but replaces it with another, namely, that errors in the multiple imperfect measures are uncorrelated. That assumption is unlikely to strictly hold in the present example since all three measures of substance use disorders — SAMISS, SCID, and chart review — essentially rely on patient self-report. However, the assumption of an error-free RS is certain not to hold either, and the researcher is left with the task of synthesizing alternate estimates generated by different methods with different assumptions. Critics have noted that dependent tests will lead to overestimation of test characteristics by LCA (Pepe & Janes, 2007). An extensive LCA simulation exercise indicated that in scenarios such as the one considered here, with moderate prevalence and tests with moderate sensitivity and high specificity (such as the SCID), the most notable consequence of dependent errors would be an overestimate of the dependent tests' sensitivity (Torrance-Rynard & Walter, 1997). Multiple alternative approaches incorporating conditional dependence have been proposed, such as Gaussian random effects, beta binomial, and fixed mixture models, although in many practical situations assessing the relative validity of these various approaches can be challenging (Albert & Dodd, 2004).

Simple SA has the advantages of being easy to calculate and permitting the consideration of a range of different misclassification scenarios. It facilitates an understanding of the specific scenarios in which substantial bias would be expected. In our example, the original SAMISS test characteristic estimates were fairly robust to a wide range of assumptions about RS sensitivity. However, the results were quite sensitive to minor deviations from perfection in the RS specificity, since SCID specificity <100% primarily leads to reclassification of observations who were SAMISS-positive but SCID-negative, of whom there were only a small number to begin with in this example.

Simple SA may be criticized as subjective, relying as it does on the investigator's choice of likely error probabilities for RS misclassification. The choice of error probabilities should be driven by existing literature and expert consensus, and the investigator should examine a range of different assumptions. If no SA is undertaken, the investigator is making the implicit assumption that the RS is a perfect measure of disease status, a generally untenable assumption.

The results of simple SA can be challenging to present succinctly, as adjusted estimates from multiple different scenarios often must be listed with little guidance to the reader about how to weight the various estimates. We used Monte Carlo methods to produce a single "best guess" point estimate that reflected our best information about misclassification by the SCID as well as a single "95% simulation interval" that incorporated both uncertainty due to random sampling error as well as our uncertainty about the unknown values of SCID sensitivity and specificity. Such approaches currently require individualized programming but are relatively easy to implement (the Stata program used in this paper is available from the corresponding author upon request).

In the absence of a perfect measure, simple SA, LCA, and other methods represent alternative approaches to estimating unobservable values. These and other approaches should therefore by viewed as complementary rather than competitive, with the results of each interpreted in the context of its strengths and weaknesses. Both LCA and simple SA are computationally simple enough to be easily incorporated into future reports of prevalence and validation studies. In the present example, both simple SA and LCA supported the conclusion that the test characteristics of the new screening instrument were somewhat underestimated by treating the SCID as the gold standard. In interpreting the results of this validation study, it would therefore be more informative for future efforts to focus on producing an improved estimate of SCID specificity than SCID sensitivity. Thus, such analyses can help guide future research priorities and enrich our appreciation of the assumptions and limitations of our standard analytic approaches to common psychometric problems.

Acknowledgments

This research was supported in part by the University of North Carolina at Chapel Hill Center for AIDS Research (CFAR), an NIH-funded program (P30 AI50410). The content of this publication does not necessarily reflect the views or policies of NIH. Dr. Pence was supported in part by a Howard Hughes Medical Institute Pre-doctoral Fellowship. Dr. Gaynes was supported in part by an NIMH K23 Career Development Award (MH01951-03).

References

Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. Biometrics 2004;60(2):427—435. [PubMed: 15180668]

First, MH.; Spitzer, RL.; Gibbon, M.; Williams, J. Structured Clinical Interview for DSM-IV Axis I Disorders -- Research Version, Patient Edition (SCID-I/P). New York: Biometrics Research, New York State Psychiatric Institute; 1990.

Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. International Journal of Epidemiology 2005;34(6):1370—1376. [PubMed: 16172102]

Kendell R, Jablensky A. Distinguishing Between the Validity and Utility of Psychiatric Diagnoses. American Journal of Psychiatry 2003;160(1):4—12. [PubMed: 12505793]

Kessler RC, Abelson J, Demler O, Escobar JI, Gibbon M, Guyer ME, Howes MJ, Jin R, Vega WA, Walters EE, Wang P, Zaslavsky A, Zheng H. Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMHCIDI).

International Journal of Methods in Psychiatric Research 2004;13(2):122—139. [PubMed: 15297907]

McHugh PR. Striving for Coherence: Psychiatry's Efforts Over Classification. JAMA 2005;293(20):2526—2528. [PubMed: 15914753]

Paykel ES. Mood Disorders: Review of Current Diagnostic Systems. Psychopathology 2002;35(2-3):94—99. [PubMed: 12145491]

Pence BW, Gaynes BN, Whetten K, Eron JJ Jr, Ryder RW, Miller WC. Validation of a brief screening instrument for substance abuse and mental illness in HIV-positive patients. JAIDS 2005;40(4):434—444. [PubMed: 16280698]

Pence BW, Miller WC, Whetten K, Eron JJ, Gaynes BN. Prevalence of DSM-IV-defined mood, anxiety, and substance use disorders in an HIV clinic in the Southeastern United States. JAIDS 2006;42(3):298—306. [PubMed: 16639343]

Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. Biostatistics 2007;8(2):474—484. [PubMed: 17085745]

Robins LN, Wing J, Wittchen HU, Helzer JE, Babor TF, Burke J, Farmer A, Jablenski A, Pickens R, Regier DA, et al. The Composite International Diagnostic Interview. An epidemiologic Instrument suitable for use in conjunction with different diagnostic systems and in different cultures. Archives of General Psychiatry 1988;45(12):1069—1077. [PubMed: 2848472]

Rothman, KJ.; Greenland, S. Modern Epidemiology. 2nd. 1998.

Steiner JL, Tebes JK, Sledge WH, Walker ML. A comparison of the structured clinical interview for DSM-III-R and clinical diagnoses. Journal of Nervous and Mental Disease 1995;183(6):365—369. [PubMed: 7798084]

Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance. Statistics in medicine 1997;16(19):2157—2175. [PubMed: 9330426]

Walter SD, Irwig LM. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. Journal of clinical epidemiology 1988;41 (9):923—937. [PubMed: 3054000]

Table 1

Prevalence of substance use disorders in the past year in the validation study sample (n=146) and the full clinic sample (n=1,125) according to SAMISS screen, SCID interview, and chart review.

Sample and measure	n	%
Validation study sample		
SAMISS	55	37.7
SCID	29	19.9
Chart review	32	21.9
Full clinic sample		
SAMISS	371	33.0
SCID (predicted)		21.0

Table 2

Estimation of test characteristics using (1) SCID as reference standard, (2) latent class analysis, and (3) simple sensitivity analysis.

Method / **Test**	**Assumption about SCID sensitivity & specificity**	**Estimated Sensitivity (95% CI)**	**Estimated Specificity (95% CI)**
SCID as reference standard			
SAMISS	100% & 100%	86% (68-96%)	75% (66-82%)
Latent class analysis			
SAMISS	None	91% (80-100%)	81% (73-89%)
SCID	None	73% (56-89%)	98% (95-100%)
Chart review	None	85% (70-100%)	99% (97-100%)
Simple sensitivity analysis[1]			
SAMISS	90% & 100%	86% (81-92%)	76% (69-83%)
SAMISS	70% & 100%	86% (81-92%)	82% (75-88%)
SAMISS	90% & 96%	98% (96-100%)	76% (69-83%)
SAMISS	70% & 96%	98% (96-100%)	82% (75-88%)

[1]Simple sensitivity analysis of the expected test characteristics of the SAMISS under four different assumptions about true (unobserved) SCID sensitivity and specificity.

Table 3

Estimated prevalence of past-year substance abuse disorders in 1,125 HIV-positive patients, adjusted for measurement error.

Assumption about SCID sensitivity & specificity	Estimated true prevalence (95% CI)
Assuming no error in SCID	
100% & 100%	21% (19-23%)
Using latent class analysis estimates	
73% & 98%	27% (24-29%)
Varying sensitivity	
89% & 98%	22% (19-24%)
56% & 98%	35% (32-38%)
Varying specificity	
73% & 96%	25% (22-27%)
73% & 100%	29% (26-31%)
Monte Carlo simulation: Simultaneously considering range of possible values for sensitivity and specificity	
(56-89%) & (96-100%)	27% (22-33%)*

*95% simulation interval incorporates both uncertainty associated with random sampling error (as with a standard confidence interval) as well as uncertainty associated with the unknown values of SCID sensitivity and specificity.