Psychometric Properties of Measures of substance use: A systematic review and meta-analysis of reliability, validity and diagnostic test accuracy

CURRENT STATUS: UNDER REVIEW

Glenn-Milo Santos  
Glenn-Milo.Santos@ucsf.edu  
University of California San Francisco  
Corresponding Author  
ORCiD: 0000-0003-1009-5317

Steffanie Strathdee  
University of California San Diego

Nabila El-Bassel  
Columbia University

Poonam Patel  
San Francisco Department of Public Health

Divya Subramanian  
San Francisco Department of Public Health

Danielle Horyniak Horyniak  
University of California San Diego

Ryan R. Cook  
University of California Los Angeles

Charlotte McCullagh  
Columbia University

Phillip Marotta  
Columbia University

Foram Choksi  
San Francisco Department of Public Health

Brian Kang  
San Francisco Department of Public Health

Isabel Allen  
University of California San Francisco
Abstract

Background Synthesis of psychometric properties of substance use measures to identify patterns of use and substance use disorders remains limited. To address this gap, we sought to systematically evaluate the psychometric properties of measures to detect substance use and misuse. Methods We conducted a systematic review and meta-analysis of literature on measures of substance classes associated with HIV risk (heroin, methamphetamine, cocaine, ecstasy, alcohol) that were published in English before June 2016 that reported at least one of the following psychometric outcomes of interest: internal consistency (alpha), test-retest/inter-rater reliability (kappa), sensitivity, specificity, positive predictive value, and negative predictive value. We used meta-analytic techniques to generate pooled summary estimates for these outcomes using random effects and hierarchical logistic regression models. Results Findings across 387 paper revealed that overall, 65% of pooled estimates for alpha were in the range of fair-to-excellent; 44% of estimates for kappa were in the range of fair-to-excellent. In addition, 69%, 97%, 37% and 96% of pooled estimates for sensitivity, specificity, positive predictive value, and negative predictive value, respectively, were in the range of moderate-to-excellent. Conclusion We conclude that many substance use measures had pooled summary estimates that were at the fair/moderate-to-excellent range across different psychometric outcomes. Most scales were conducted in English, within the United States, highlighting the need to test and validate these measures in more diverse settings. Additionally, the majority of studies had high risk of bias, indicating a need for more studies with higher methodological quality.

Background

Substance use, including illicit drug use and alcohol, is prevalent worldwide with about 5% of adults using illicit substances (1) and 40% of adults consuming alcohol, in the past year.
Moreover, the number of people with drug use disorders was estimated at 62 million, while the number of individuals with alcohol use disorders was estimated at 100.4 million in 2016 (3). Substance use disorders are associated with substantial morbidity and mortality globally, including HIV acquisition and mortality among people living with HIV. Illicit drug use disorders were attributed to 20 million disability-adjusted life years (DALYs) lost (4) while alcohol use disorders were attributed to 85 million DALYs lost in 2012 (5).

Specific classes of substances play an important role in HIV risk (6-9). Use of heroin and stimulant drugs such as methamphetamine and cocaine, has been independently associated with HIV-related risk behaviors through direct pathways, such as needle sharing (9). Moreover, these classes of substances, along with alcohol use, have been linked to HIV risk through indirect pathways that involve substance-related sexual risk behaviors, including less condom use, condom failure, and high numbers of sexual partners (8, 10, 11). Furthermore, heroin and stimulant drugs has been linked to HIV incidence in multiple longitudinal studies (12-15). Among people living with HIV (PLWH), substance use disorders may lead to less optimal HIV care outcomes because of their associations with lower likelihood of being linked to HIV care, retained in care, receiving ART, having high ART adherence and lower likelihood of having an undetectable HIV viral load (6, 7, 16-18).

Given the role of substance use in the global burden of disease and the overlap between use of specific substances and HIV, it is important for clinicians and researchers to have tools with high reliability, validity, and diagnostic accuracy (19). Yet too few use measures with known psychometric properties when assessing substance use, especially in the setting of HIV. Currently, there are a myriad of standardized questionnaires used to screen substance use and misuse that require patients to self-report patterns of use and substance-related problems. Examples such as the Alcohol Use Disorders Identification Test and the Drug Use Disorders Identification test (20, 21) provide scores that correspond with severity of substance use and related problems. It remains that there are no biological measures that define a substance use disorder; existing biological measures are considered to be indirect correlates of use disorders (22). Examples include alcohol biomarkers like Carbohydrate-Deficient Transferrin (CDT), and Gamma Glutamyl Transferase (GGT), which are used to screen for alcohol dependence and heavy drinking, respectively (22). There is a great need to evaluate the psychometric performance of these measures and markers across studies in settings of HIV to elucidate the overall
validity, reliability, and diagnostic accuracy.

One approach to informing the use of psychometric measures in research and clinical care is pooling the psychometric characteristics of measures across studies involves the use of meta-analytic techniques, which generates summary estimates of the validity, reliability, and diagnostic accuracy of different questionnaires (23-27). However, synthesis of psychometric properties of substance use measures to identify patterns of use and substance use disorders remains limited, with few exceptions (21, 28, 29). One meta-analysis focused on the accuracy of self-reported assessments to diagnose alcohol and cannabis use disorders found that instruments had a pooled sensitivity of 0.88 and a pooled specificity of 0.90 among emergency room department pediatric patients (28). Another meta-analysis observed that studies with single questions to identify alcohol use disorders in primary care had pooled sensitivity of 0.54 and pooled specificity of 0.87 while two-question measures had a pooled sensitivity of 0.87 and a pooled specificity of 0.80 (29). More commonly, however, reviews on substance use measures present psychometric data in a descriptive fashion (19, 30, 31). Therefore, more rigorous efforts to systematically pool the psychometric properties of substance use measures are needed to establish the overall performance and accuracy of these tools and point toward their utility in future research.

To address these gaps, we conducted a systematic review and meta-analysis of literature to identify studies that have reported validity and reliability of substance use measures and pooled these measure using meta-analytic techniques. For the purposes of this review, we targeted our search for measures of substance classes previously associated with HIV risk. Specifically, we focused our review on measures for the following: alcohol, methamphetamine and amphetamine, cocaine, heroin, and ecstasy, regardless of whether the study was conducted among a population at high risk for HIV. Additionally, we included measures that evaluated substance use in general (i.e., measures that did not differentiate between classes of substances) as long as those measures were inclusive of our targeted substance classes. This study’s review questions are: What are the summary reliability, validity—as measured by alpha and kappa coefficients—and diagnostic accuracy—as measured by sensitivity, specificity, positive predictive value, and negative predictive value—of various substance and alcohol measures to screen for use and use disorders?

Methods
**Search strategy:**
We conducted a systematic review of studies published prior to June 2016 on substance use measures indexed in electronic databases including PubMed, PsycINFO, and EMBASE. We developed Boolean search terms to capture substance use measures that have been previously associated with HIV risk, in consultation with the reference librarian from the University of California San Francisco with a master’s degree in library and information science (MLIS). The following substance classes were included: alcohol, methamphetamine and amphetamine, cocaine, heroin, and 3,4-methylenedioxy-methamphetamine (MDMA; “ecstasy”). Because the focus of this study was to pool psychometric properties of measures, we also included search terms related to validity, reliability, and diagnostic accuracy (i.e., alpha, kappa, sensitivity, specificity, positive predictive value, negative predictive value). Search terms included MeSH headings related to our research question, general terms related to substance use and psychometric properties or interest, as well as specific terms referencing the names of well-known substance use measures. The search terms used are provided in the appendix. This review was registered in Prospero, the International prospective register of systematic reviews (study number: CRD42017058813).

**Primary Outcomes:**
We aimed to estimate the pooled summary estimates for the following psychometric outcomes: Cronbach’s alpha, kappa, sensitivity, specificity, positive predictive value, and negative predictive value. Descriptions for these outcomes are provided below:
| Psychometric Outcome | Description |
|----------------------|-------------|
| Cronbach’s alpha     | measure of internal consistency, that is, how closely correlated a set of scale items are, as |
| Kappa                | measure of inter-rater agreement or inter-rater reliability for qualitative (categorical) items account the possibility of the agreement occurring by chance. |
| Sensitivity,         | measure of a test/scales’ ability to correctly detect patients who do truly have the condition people who screen positive for substance use disorders according to the scale, among those substance use disorders based on an established standard (“gold standard”) such as meet for a disorder). |
| Specificity          | measure of the test/scales’ ability to correctly detect patients without a condition (i.e., screen negative for substance use disorders according to the scale, among those who truly use disorders based on an established standard such as meeting diagnostic criteria for a d |
| Positive predictive value (PPV) | the probability that persons with a positive screening result actually has the disorder. (i.e. who meet diagnostic criteria for a substance use disorder among those who screened positive for the disorder on a scale). |
| Negative predictive value (NPV) | the probability that people with a negative screening test actually do not have the disease people who meet diagnostic criteria for a substance use disorder among those who screen substance use disorder in a scale ). |

**Eligibility Criteria:**

We searched for relevant publications that met all of the following inclusion criteria: 1) studies that reported one or more of the psychometric outcomes of interest; 2) studies that examined on one or more substance use measures related to our substance classes of interest (i.e., alcohol, methamphetamine and amphetamine, cocaine, heroin, and ecstasy) or for substance use in general (i.e., some measures do not differentiate between multiple substances or assess classes of substances all together); 3) publication written in English (note: studies that administered measures that were not in English were eligible as long as the publication was written in English).

We excluded publications using the following exclusion criteria: 1) reporting insufficient information on reliability, validity and diagnostic accuracy for substance use measures/assessments (i.e., no numeric information on our psychometric outcomes, sample size); 2) articles that provide psychometric data for a measure/assessment that is not related to substance use (e.g., a study on internal consistency data on a depression scale among substance users); 3) articles and/or secondary data analyses that report
reliability and validity data from a primary outcome paper that was already included in the review; 4) reviews, commentaries, case report studies and other publications with insufficient reporting of data; 5) substance use measures/assessments that focus on aspects other than actual substance consumption, dependence or substance use disorder (e.g., a study reporting validity of a self-efficacy scale for resisting substance use; a study that examines the underlying mechanisms of substance use among those who already have a substance use disorder); and 6. studies with psychometric properties that focus on substance classes outside the scope of our review (e.g. marijuana or tobacco).

**Screening procedures:**
All citations (including their titles and abstracts) captured by the search strategy were imported into Covidence.org (Melbourne Victoria), which allowed research team members to independently review and screen citations using a centralized, online database. Each title/abstract was screened by two members of a team comprising master-, doctoral-, and post-doctoral-level researchers trained in the study protocol (co-authors PP, DH, RC, DS, CM, PM, and FC) and citations that were coded as eligible by both reviewers were moved to the full-text review phase. The same process was then repeated for full-text articles. In the event of discrepancies between reviewers in both the title and abstract phase and the full-text phase, a third team member (GMS) reviewed the relevant documents and helped reconcile the differences. Articles that were deemed eligible in the full-text review stage were included in the data extraction phase described below.

**Data extraction:**
Team members extracted data on the psychometric properties, scale and study characteristics, sample size, study sample characteristics/co-factors of interest (country where study was conducted, number of sites, language that the scale was administered, gender of participants included), cut-offs used, comparison measure/gold-standard used, and other information relevant to study, including information on study quality (32). Some papers reported multiple data points for psychometric outcomes from different study populations (e.g., disaggregated data by sex or different research sites). These data points were extracted as separate records only if the paper did not provide a single overall measure for the psychometric outcomes for the entire study sample, consistent with other analyses (24).

**Assessment of bias risk:**
For studies reporting diagnostic measures (e.g., sensitivity and specificity), reviewers rated study quality using the Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies, QUADAS-2, guidelines (32), which includes quality rating questions on the study’s patient selection, index test, reference standard, and flow and timing. For studies that did not include diagnostic accuracy measures, only relevant domains of QUADAS-2 were assessed, as appropriate (i.e., rating regarding the reference standard was not conducted). All extracted data were entered into an electronic questionnaire programmed in Qualtrics, and checked by another researcher (conducted by the same co-authors who screened citations, as well as co-author BK) to verify accuracy.

Data analyses:
We calculated separate pooled summary estimates for each of the 37 substance use measures and also fitted separate models for each of the six psychometric outcomes for validity, reliability, and accuracy. For alpha, kappa, PPV and NPV, we pooled data across studies using DerSimonian-Laird random effects models, implemented in STATA version 13 (Colleges Station, TX)(33). Random effects meta-analyses models, as opposed to fixed-effects models, are preferred for pooling data from diagnostic accuracy tests since heterogeneity is presumed to exists across these studies (34). Random effects models, which are considered the default models used in meta-analyses for diagnostic accuracy tests, synthesize the psychometric outcomes from separate studies into a weighted average effect size (pooled summary estimate), using inverse variance weighting, based on sample size, while taking into account the extent of the variability of the effect sizes observed in separate studies (34). Additionally, for sensitivity and specificity, we used hierarchical logistic regression models, implemented using the metandi command in STATA, to account for the correlation between the two measures (i.e., trade-off between sensitivity and specificity)(35-37). Since metandi requires a minimum of four observations to conduct a meta-analysis, we pooled measures with less than four records for sensitivity and specificity outcomes using the random effects models described for other outcomes, and noted this alternate approach in the results, as appropriate.

Classification and evaluation of pooled estimates
Qualitatively, pooled summary estimates for alpha and kappa were classified as “excellent” for estimates that were >0.89, “good” for estimates that were between 0.85-0.89, “moderate” for estimates that were between 0.80-0.84, “fair” for estimates that were between 0.75-0.79, or “unsatisfactory” for estimates below 0.75, consistent with
other studies (24, 38).

Pooled summary estimates for sensitivity, specificity, positive predictive value and negative predictive value were classified as “excellent” for estimates that were >0.89, “good” for estimates that were between 0.8-0.89, “moderate” for estimates that were between 0.6-0.79, and “low” for estimates that were <0.6 (24, 39).

For each pooled psychometric summary estimate, we calculated $I^2$ statistics, which represents the percentage of total variation across studies, to assess heterogeneity. We considered pooled estimates as having low heterogeneity if $I^2$ 25%, moderate heterogeneity if $I^2$ 50%, and high heterogeneity if $I^2$ 75% (40). We did not use standard meta-analyses tests for publication bias given the limitations of these tests for diagnostic test accuracy studies and due to the characteristics of our psychometric outcomes (e.g., truncated measures cannot fall below zero) (41). As indicated in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy, using these tests are inappropriate because they will likely lead to a high false-positive rate for publication bias (34).

Results

Screening and study inclusion:
Study screening and inclusion is summarized in Figure 1. In brief, in the identification stage, we initially identified 7,555 references in the initial search, of which, 208 were excluded for being duplicates. In the title and abstract review phase, reviewers excluded 5,854 studies that were deemed ineligible. Full-text reviews were conducted for 1,493 articles that were deemed eligible from title and abstract review. Of the full-text reviewed articles, 1105 studies were excluded for not meeting eligibility criteria. The most common reasons for exclusion were: scales or measures that were outside the scope of review (n=386), lack of psychometric data on scales of interests (n=140), lab or methods papers that were outside the scope of the review (n=130), non-English language publications (n=110), duplicate study (n=98), psychometric outcomes that were outside the scope of review (n=79). In total, there were 387 unique studies included in the data extraction phase containing sufficient data on the outcomes for 37 scales (Table 1).

Study Characteristics:
Table 2 presents characteristics of the studies included in this meta-analysis. As
mentioned, studies published in English were included in this review, regardless of the language in which the scales were administered. Among the 387 studies included, the most common language in which the scale/measure was conducted in was English (63%), followed by Spanish (9%), French (5%), Portuguese (3%), and Chinese (2%). A large proportion of studies were conducted in the United States (40%). The median sample size was 286 [Range=9-50,049]. The vast majority of studies (83%) included men and women (n=323). Additionally, 11% (n=42) of the studies included study sample comprised only of men, while 5% (n=20) studies included study samples comprised only of women. Most studies were published after 1999 (66%), with studies published between 2000-2009 accounting for 38% (n=148) of the studies meta-analyzed, and studies published between 2010-2017 accounting for 28% (n=110). Most studies involved a single study site 61%, while 39% were multi-site studies. Additionally, 72% of the studies involved convenience samples, 20% included random or probability based samples, and 7% had other or unclear sampling strategies.

**Assessment of bias in Study Quality:**
The risk of bias in the four QUADAS 2 domains for each study included in this meta-analysis is presented in Table 2. Of the studies included, 58% of studies had a low risk of bias with respect to the patient population; 57% has low risk of bias in the index test domain, 48% has low risk of bias in the reference standard test domain, and 72% had low risk for the flow and timing. Overall, only 16% of studies had low risk of bias across all four of these QUADAS 2 domains.

**Pooled Summary Estimates: Overall findings:**
The pooled summary estimates of psychometric properties of substance use measures (which are described in Table 1) are qualitatively summarized in Table 3. Overall, 65% of pooled estimates for alpha were in the range of fair-to-excellent; 44% of estimates for kappa were in the range of fair-to-excellent. In addition, 69%, 97%, 37% and 96% of pooled estimates for sensitivity, specificity, positive predictive value, and negative predictive value, respectively, were in the range of moderate-to-excellent.

Self-reported measures that had all pooled estimates that were fair/moderate or better include the following: Alcohol Dependence Scale; Addiction Severity Index (ASI); ASI subscale for Alcohol; ASSIST; the Composite International Diagnostic Interview; Drug Abuse Screen Test - 10 item scale; Drug Use Disorders Identification Test; Problem
Oriented Screening Instrument for Teenagers; Severity of Dependence scale; Timeline Followback; and Chemical Use, Abuse, and Dependence. Biomarkers that had all pooled estimates that were fair/moderate or better include the following: Ethyl glucuronide; Phosphatidylethanol test; and the combined used of Carbohydrate deficient transferrin and Mean corpuscular volume. In general, we also observed high heterogeneity between studies for most pooled estimates.

**Pooled Summary Estimates, by Substance Use Measure:**
The pooled estimates and 95% confidence intervals for alpha, kappa, sensitivity, specificity, positive predictive value, and negative predictive value are shown in Tables 4, 5, 6, 7, 8, and 9, respectively. Below we summarize the results of the pooled summary estimates alphabetically for each of the 37 substance use measures, grouping self-reported measures and biomarkers separately. The list of references for the studies meta-analyzed for each scale/measure is presented in Table 10.

**Self-Reported Measures:**

**Alcohol Dependence Scale (ADS)**
The pooled alpha estimate for ADS (3 data points) was good: 0.90 (95%CI=0.80-0.99) and there was high heterogeneity between studies (I² 98.9%). The pooled sensitivity estimate for ADS (2 data points) was excellent: 0.95 (95%CI=0.90-1.00) and there was low heterogeneity between studies (I² 0%). The pooled specificity estimate (2 data points) was moderate: 0.64 (95%CI=0.52-0.77) and there was moderate heterogeneity between studies (I² 60.1%). There was insufficient data to calculate the pooled PPV and NPV estimates for ADS.

**Addiction Severity Index (ASI)**
The pooled alpha estimate for ASI (3 data points) was good: 0.84 (95%CI=0.81-0.87) and there was moderate heterogeneity between studies (I² 38.5%). There was insufficient data to calculate pooled kappa, sensitivity, specificity, PPV, and NPV estimates.

**Addiction Severity Index-Alcohol (alcohol sub-scale; ASI-A)**
The pooled alpha estimate (18 data points) was moderate: 0.77 (95%CI=0.73-0.81) and there was high heterogeneity between studies (I² 94.3%). The pooled sensitivity estimate for ASI-A (6 data points) was good: 0.83 (95%CI=0.67-0.92) and there was high heterogeneity between studies (I² 87.6%). The pooled specificity estimate for ASI-A (6
data points) was moderate: 0.79 (95%CI=0.67-0.88) and there was high heterogeneity between studies ($I^2$ 91.2%). There was insufficient data to calculate pooled kappa, PPV and NPV estimates for ASI-A.

**Addiction Severity Index-Drugs (drugs sub-scale; ASI-D)**
The pooled alpha estimate for ASI-D (16 data points) was unsatisfactory: 0.68 (95%CI=0.63-0.74) and there was high heterogeneity between studies ($I^2$ 95.6%). The pooled sensitivity estimate (5 data points) was good: 0.86 (95%CI=0.83-0.89) and there was moderate heterogeneity between studies ($I^2$ 62.5%). The pooled specificity estimate (5 data points) was good: 0.85 (95%CI=0.77-0.91) and there was high heterogeneity between studies ($I^2$ 86%). There was insufficient data to calculate the pooled kappa, PPV and NPV estimates.

**The Alcohol, Smoking, and Substance Involvement Screening Test (ASSIST)**
The pooled alpha estimate (7 data points) was good: 0.85 (95%CI=0.80-0.91) and there was high heterogeneity between studies ($I^2$ 94%). The pooled sensitivity estimate (2 data points) was good: 0.83 (95%CI=0.80-0.87) and there was low heterogeneity between studies ($I^2$ 0%). The pooled specificity estimate (2 data points) was moderate: 0.73 (95%CI=0.57-0.88) and there was high heterogeneity between studies ($I^2$ 91%). There was insufficient data to calculate the pooled estimate for kappa, PPV, and NPV.

**Alcohol Use Disorders Identification Test (AUDIT)**
The pooled alpha estimate for AUDIT (80 data points) was moderate: 0.85 (95%CI=0.83-0.87) and there was high heterogeneity between studies ($I^2$ 98%). The pooled kappa estimate for AUDIT (4 data points) was unsatisfactory: 0.46 (95%CI=0.25-0.67) and there was high heterogeneity between studies ($I^2$ 0.99). The pooled sensitivity estimate for AUDIT (135 data points) was good: 0.86 (95%CI=0.84-0.88) and there was high heterogeneity between studies ($I^2$ 97%). The pooled specificity estimate for AUDIT (135 data points) was good: 0.87 (95%CI=0.85-0.89) and there was high heterogeneity between studies ($I^2$ 99%). The pooled PPV estimate for AUDIT (65 data points) was moderate: 0.61 (95%CI=0.51-0.71) and there was high heterogeneity between studies ($I^2$ 99%). The pooled NPV estimate for AUDIT (54 data points) was excellent: 0.94 (95%CI=0.93-0.95)
and there was high heterogeneity between studies ($I^2$ 96%)

**Alcohol Use Disorders Identification Test-3 (AUDIT-3)**

Alpha cannot be calculated for AUDIT-3 because it is a single-item measure. There was insufficient data to calculate the pooled estimate for kappa. The pooled sensitivity estimate for AUDIT-3 (22 data points) was good: $0.84$ (95%CI=$0.80$-$0.88$) and there was high heterogeneity between studies ($I^2$ 90%). The pooled specificity estimate for AUDIT-3 (22 data points) was good: $0.84$ (95%CI=$0.75$-$0.90$) and there was high heterogeneity between studies ($I^2$ 99%). The pooled PPV estimate for AUDIT-3 (9 data points) was moderate: $0.63$ (95%CI=$0.49$-$0.77$) and there was high heterogeneity between studies ($I^2$ 99%). The pooled NPV estimate (7 data points) was excellent: $0.94$ (95%CI=$0.90$-$0.98$) and there was high heterogeneity between studies ($I^2$ 95%).

**Alcohol Use Disorders Identification Test-C (AUDIT-C)**

The pooled alpha estimate for AUDIT-C (20 data points) was fair: $0.75$ (95%CI=$0.70$-$0.80$) and there was high heterogeneity between studies ($I^2$ 99%). The pooled kappa estimate for AUDIT-C (2 data points) was unsatisfactory: $0.41$ (95%CI=$0.39$-$0.43$) and there was low heterogeneity between studies ($I^2$ 0%). The pooled sensitivity estimate for AUDIT-C (45 data points) was good: $0.87$ (95%CI=$0.84$-$0.90$) and there was high heterogeneity between studies ($I^2$ 99%). The pooled specificity estimate for AUDIT-C (45 data points) was good: $0.84$ (95%CI=$0.81$-$0.87$) and there was high heterogeneity between studies ($I^2$ 99%). The pooled PPV estimate for AUDIT-C (22 data points) was low: $0.50$ (95%CI=$0.39$-$0.60$) and there was high heterogeneity between studies ($I^2$ 99%). The pooled NPV estimate for AUDIT-C (19 data points) was good: $0.88$ (95%CI=$0.83$-$0.92$) and there was high heterogeneity between studies ($I^2$ 99%).

**Brief Michigan Alcoholism Screening Test (B-MAST)**

There was insufficient data to calculate the pooled estimate for B-MAST’s alpha and kappa. The pooled sensitivity estimate for B-MAST (21 data points) was low: $0.50$ (95%CI=$0.38$-$0.62$) and there was high heterogeneity between studies ($I^2$ 99%). The pooled specificity estimate for B-MAST (21 data points) was excellent: $0.97$ (95%CI=$0.96$-$0.98$) and there was high heterogeneity between studies ($I^2$ 97%). The pooled PPV
estimate for B-MAST (3 data points) was moderate: 0.65 (95%CI=0.38-0.93) and there was high heterogeneity between studies ($I^2$ 99%). The pooled NPV estimate for B-MAST (2 data points) was excellent: 0.90 (95%CI=0.87-0.94) and there was moderate heterogeneity between studies ($I^2$ 33%).

**Cut down, Annoyed, Guilty, Eye-opener (CAGE)**
The pooled alpha estimate for CAGE (22 data points) was unsatisfactory: 0.70 (95%CI=0.65-0.75) and there was high heterogeneity between studies ($I^2$ 98%). The pooled kappa estimate for CAGE (3 data points) was unsatisfactory: 0.57 (95%CI=0.34-0.81) and there was high heterogeneity between studies ($I^2$ 0.97). The pooled sensitivity estimate for CAGE (139 data points) was moderate: 0.70 (95%CI=0.66-0.74) and there was high heterogeneity between studies ($I^2$ 98%). The pooled specificity estimate for CAGE (139 data points) was good: 0.90 (95%CI=0.88-0.91) and there was high heterogeneity between studies ($I^2$ 99%). The pooled PPV estimate for CAGE (61 data points) was low: 0.51 (95%CI=0.45-0.58) and there was high heterogeneity between studies ($I^2$ 99%). The pooled NPV estimate for CAGE (39 data points) was excellent: 0.91 (95%CI=0.88-0.93) and there was high heterogeneity between studies ($I^2$ 97%).

**Composite International Diagnostic Interview (CIDI)**
Alpha coefficients are not calculated for CIDI. The pooled kappa estimate for CIDI (2 data points) was moderate: 0.82 (95%CI=0.61-1.02) and there was high heterogeneity between studies ($I^2$ 0.78). The pooled sensitivity estimate for CIDI (3 data points) was moderate: 0.80 (95%CI=0.67-0.92) and there was high heterogeneity between studies ($I^2$ 80%). The pooled specificity estimate for CIDI (3 data points) was good: 0.86 (95%CI=0.77-0.95) and there was high heterogeneity between studies ($I^2$ 99%). The pooled PPV estimate for CIDI (2 data points) was moderate: 0.69 (95%CI=0.26-1.00) and there was high heterogeneity between studies ($I^2$ 99%). The pooled NPV estimate for CIDI (2 data points) was good: 0.89 (95%CI=0.68-1.00) and there was high heterogeneity between studies ($I^2$ 96%).

**Car, Relax, Alone, Forget, Friends, Trouble (CRAFFT)**
The pooled alpha estimate for CRAFFT (6 data points) was unsatisfactory: 0.69 (95%CI=0.64-0.74) and there was high heterogeneity between studies ($I^2$ 83%). There was
insufficient data to calculate the pooled estimate for kappa for CRAFFT. The pooled
sensitivity estimate for CRAFFT (10 data points) was good: 0.90 (95%CI=0.84-0.94) and
there was high heterogeneity between studies (I² 97%). The pooled specificity estimate for
CRAFFT (10 data points) was moderate: 0.76 (95%CI=0.68-0.83) and there was high
heterogeneity between studies (I² 97%). The pooled PPV estimate for CRAFFT (8 data
points) was low: 0.57 (95%CI=0.34-0.80) and there was high heterogeneity between
studies (I² 99%). The pooled NPV estimate for CRAFFT (8 data points) was good: 0.86
(95%CI=0.45-1.00) and there was high heterogeneity between studies (I² 99%)

**Drug Abuse Screen Test (DAST)**
The pooled alpha estimate for DAST (6 data points) was excellent: 0.94 (95%CI=0.93-0.95)
and there was low heterogeneity between studies (I² 0%). The pooled kappa estimate for
DAST (2 data points) was moderate: 0.83 (95%CI=0.58-1.00) and there was high
heterogeneity between studies (I² 0.98). The pooled sensitivity estimate for DAST (7 data
points) was good: 0.85 (95%CI=0.74-0.92) and there was high heterogeneity between
studies (I² 89%). The pooled specificity estimate for DAST (7 data points) was good: 0.84
(95%CI=0.68-0.93) and there was high heterogeneity between studies (I² 97%). The
pooled PPV estimate for DAST (5 data points) was low: 0.51 (95%CI=0.32-0.70) and there
was high heterogeneity between studies (I² 98%). The pooled NPV estimate for DAST (4
data points) was excellent: 0.95 (95%CI=0.89-1.00) and there was high heterogeneity
between studies (I² 81%).

**Drug Abuse Screen Test - 10-item version (DAST-10)**
The pooled alpha estimate DAST-10 (6 data points) was fair: 0.79 (95%CI=0.68-0.89) and
there was high heterogeneity between studies (I² 98%). There was insufficient data to
calculate the pooled estimate for kappa for DAST-10. The pooled sensitivity estimate for
DAST-10 (6 data points) was excellent: 0.90 (95%CI=0.75-0.97) and there was high
heterogeneity between studies (I² 95%). The pooled specificity estimate for DAST-10 (6
data points) was good: 0.82 (95%CI=0.72-0.89) and there was high heterogeneity between
studies (I² 92%). The pooled PPV estimate for DAST-10 (4 data points) was good: 0.80
(95%CI=0.70-0.91) and there was high heterogeneity between studies (I² 99%). The
pooled NPV estimate for DAST-10 (4 data points) was good: 0.86 (95%CI=0.81-0.91) and
there was moderate heterogeneity between studies ($I^2$ 40%).

**Drug Use Disorders Identification Test (DUDIT)**

The pooled alpha estimate for DUDIT (15 data points) was excellent: 0.92 (95%CI=0.90-0.95) and there was high heterogeneity between studies ($I^2$ 96%). There was insufficient data to calculate the pooled kappa estimate for DUDIT. The pooled sensitivity estimate for DUDIT (12 data points) was excellent: 0.93 (95%CI=0.89-0.96) and there was high heterogeneity between studies ($I^2$ 76%). The pooled specificity estimate for DUDIT (12 data points) was moderate: 0.79 (95%CI=0.67-0.87) and there was high heterogeneity between studies ($I^2$ 96%). The pooled PPV estimate (5 data points) was moderate: 0.61 (95%CI=0.34-0.87) and there was high heterogeneity between studies ($I^2$ 99%). The pooled NPV estimate (5 data points) was excellent: 0.92 (95%CI=0.82-1.00) and there was high heterogeneity between studies ($I^2$ 78%).

**Michigan Alcohol Screening Test (MAST)**

The pooled alpha estimate for MAST (8 data points) was moderate: 0.82 (95%CI=0.78-0.86) and there was high heterogeneity between studies ($I^2$ 83%). The pooled kappa estimate for MAST (4 data points) was unsatisfactory: 0.69 (95%CI=0.58-0.81) and there was high heterogeneity between studies ($I^2$ 0.88). The pooled sensitivity estimate for MAST (12 data points) was moderate: 0.70 (95%CI=0.58-0.80) and there was high heterogeneity between studies ($I^2$ 95%). The pooled specificity estimate for MAST (12 data points) was good: 0.85 (95%CI=0.77-0.91) and there was high heterogeneity between studies ($I^2$ 97%). The pooled PPV estimate for MAST (9 data points) was low: 0.51 (95%CI=0.30-0.71) and there was high heterogeneity between studies ($I^2$ 98%). The pooled NPV estimate for MAST (6 data points) was good: 0.88 (95%CI=0.82-0.94) and there was high heterogeneity between studies ($I^2$ 92%).

**Problem Oriented Screening Instrument for Teenagers (POSIT)**

The pooled alpha estimate for POSIT (2 data points) was good: 0.86 (95%CI=0.73-0.98) and there was high heterogeneity between studies ($I^2$ 94%). The pooled sensitivity estimate for POSIT (3 data points) was good: 0.84 (95%CI=0.72-0.96) and there was high heterogeneity between studies ($I^2$ 90%). The pooled specificity estimate for POSIT (3 data
points) was good: 0.82 (95%CI=0.75-0.90) and there was high heterogeneity between studies (I² 88%). There was insufficient data to calculate the pooled kappa, PPV, and NPV estimates for POSIT.

**Self-Administered Alcoholism Screening Test (SAAST)**
The pooled alpha estimate for SAAST (2 data points) was good: 0.89 (95%CI=0.79-0.99) and there was high heterogeneity between studies (I² 95%). The pooled sensitivity estimate for SAAST (7 data points) was low: 0.52 (95%CI=0.33-0.71) and there was high heterogeneity between studies (I² 98%). The pooled specificity estimate (7 data points) was good: 0.83 (95%CI=0.76-0.90) and there was high heterogeneity between studies (I² 98%). The pooled PPV estimate for SAAST (6 data points) was low: 0.32 (95%CI=0.22-0.42) and there was high heterogeneity between studies (I² 95%). The pooled NPV estimate for SAAST (6 data points) was excellent: 0.92 (95%CI=0.89-0.95) and there was high heterogeneity between studies (I² 92%). There was insufficient data to calculate the pooled kappa estimates for SAAST.

**Semi-Structured Assessment for Drug Dependence and Alcoholism (SSADDA)**
There are no alpha coefficients associated with semi-structures assessments such as SSADDA. The pooled kappa estimate for SSADDA (8 data points) was moderate: 0.84 (95%CI=0.77-0.91) and there was high heterogeneity between studies (I² 0.97). There was insufficient data to calculate the pooled sensitivity, specificity, PPV and NPV estimates for SSADDA.

**Severity of Dependence (SDS)**
The pooled alpha estimate for SDS (6 data points) was good: 0.86 (95%CI=0.78-0.93) and there was high heterogeneity between studies (I² 95%). The pooled sensitivity estimate for SDS (6 data points) was good: 0.83 (95%CI=0.76-0.90) and there was high heterogeneity between studies (I² 77%). The pooled specificity estimate (6 data points) was good: 0.84 (95%CI=0.78-0.89) and there was moderate heterogeneity between studies (I² 44%). The pooled PPV estimate for SDS (3 data points) was good: 0.90 (95%CI=0.86-0.94) and there was low heterogeneity between studies (I² 0%). The pooled NPV estimate for SDS (3 data points) was good: 0.83 (95%CI=0.76-0.89) and there was low...
heterogeneity between studies ($I^2 3.5\%$). There was insufficient data to calculate the pooled kappa estimate for SDS.

**Tolerance-Annoyance Cut Down Eye Opener (T-ACE)**
The pooled alpha estimate for T-ACE (2 data points) was unsatisfactory: 0.50 (95%CI=0.47-0.52) and there was high heterogeneity between studies ($I^2 29\%$). The pooled sensitivity estimate for T-ACE (8 data points) was good: 0.83 (95%CI=0.74-0.92) and there was high heterogeneity between studies ($I^2 96\%$). The pooled specificity estimate for T-ACE (8 data points) was moderate: 0.72 (95%CI=0.65-0.79) and there was high heterogeneity between studies ($I^2 98\%$). The pooled PPV estimate for T-ACE (6 data points) was low: 0.35 (95%CI=0.25-0.45) and there was high heterogeneity between studies ($I^2 99\%$). The pooled NPV estimate for T-ACE (2 data points) was good: 0.87 (95%CI=0.62-1.00) and there was high heterogeneity between studies ($I^2 97\%$). There was insufficient data to calculate the pooled estimate for kappa for T-ACE.

**Timeline Followback (TLFB)**
There are no alpha coefficients associated with TLFB. The pooled kappa estimate for TLFB (3 data points) was good: 0.86 (95%CI=0.81-0.91) and there was high heterogeneity between studies ($I^2 0.88\%$). The pooled sensitivity estimate for TLFB (4 data points) was moderate: 0.80 (95%CI=0.73-0.87) and there was moderate heterogeneity between studies ($I^2 63\%$). The pooled specificity estimate for TLFB (3 data points) was excellent: 0.97 (95%CI=0.95-0.99) and there was low heterogeneity between studies ($I^2 0\%$). There was insufficient data to calculate the pooled estimate for PPV and NPV for TLFB.

**Tolerance, Worried, Eye-Opener, Amnesia, Cut down (TWEAK)**
The pooled alpha estimate for TWEAK (3 data points) was unsatisfactory: 0.62 (95%CI=0.55-0.69) and there was high heterogeneity between studies ($I^2 86\%$). The pooled sensitivity estimate for TWEAK (36 data points) was good: 0.85 (95%CI=0.80-0.89) and there was high heterogeneity between studies ($I^2 96\%$). The pooled specificity estimate for TWEAK (36 data points) was good: 0.86 (95%CI=0.82-0.90) and there was high heterogeneity between studies ($I^2 99\%$). The pooled PPV estimate for TWEAK (5 data points) was low: 0.43 (95%CI=0.26-0.61) and there was high heterogeneity between
studies ($I^2$ 99%). The pooled NPV estimate for TWEAK (2 data points) was good: 0.88 (95%CI=0.70-1.00) and there was high heterogeneity between studies ($I^2$ 95%). There was insufficient data to calculate the pooled estimate for kappa for TWEAK.

The Chemical Use, Abuse, and Dependence (CUAD)
The pooled alpha estimate for CUAD (3 data points) was excellent: 0.96 (95%CI=0.94-0.98) and there was high heterogeneity between studies ($I^2$ 95%). There was insufficient data to calculate the pooled estimate for kappa, sensitivity, specificity, PPV, and NPV for CUAD.

Biomarkers:

**Alanine transaminase (ALT)**
The pooled sensitivity estimate for ALT (32 data points) was low: 0.32 (95%CI=0.24-0.40) and there was high heterogeneity between studies ($I^2$ 96.1%). The pooled specificity estimate for ALT (32 data points) was good: 0.88 (95%CI=0.83-0.92) and there was high heterogeneity between studies ($I^2$ 95.8%). The pooled PPV estimate for ALT (7 data points) was low: 0.37 (95%CI=0.18-0.56) and there was high heterogeneity between studies ($I^2$ 96.1%). The pooled NPV estimate for ALT (4 data points) was moderate: 0.63 (95%CI=0.42-0.85) and there was high heterogeneity between studies ($I^2$ 97.5%).

**Aspartate transaminase (AST)**
The pooled sensitivity estimate for AST (33 data points) was low: 0.48 (95%CI=0.40-0.55) and there was high heterogeneity between studies ($I^2$ 97%). The pooled specificity estimate for AST (33 data points) was good: 0.86 (95%CI=0.81-0.90) and there was high heterogeneity between studies ($I^2$ 97%). The pooled PPV estimate for AST (8 data points) was low: 0.42 (95%CI=0.27-0.57) and there was high heterogeneity between studies ($I^2$ 93%). The pooled NPV estimate for AST (6 data points) was moderate: 0.69 (95%CI=0.55-0.83) and there was high heterogeneity between studies ($I^2$ 95%).
**Aspartate transaminase, Alanine transaminase ratio (AST/ALT ratio)**
The pooled sensitivity estimate for AST/ALT ratio (6 data points) was low: 0.34 (95%CI=0.22-0.46) and there was high heterogeneity between studies (I² 96%). The pooled specificity estimate (4 data points) was moderate: 0.73 (95%CI=0.52-0.94) and there was high heterogeneity between studies (I² 98%). There was insufficient data to calculate the pooled estimate for PPV and NPV.

**Blood alcohol concentration (BAC)**
The pooled sensitivity estimate for BAC (5 data points) was moderate: 0.64 (95%CI=0.59-0.69) and there was moderate heterogeneity between studies (I² 44%). The pooled specificity estimate for BAC (5 data points) was moderate: 0.80 (95%CI=0.72-0.87) and there was high heterogeneity between studies (I² 93%). The pooled PPV estimate for BAC (3 data points) was low: 0.60 (95%CI=0.15-1.00) and there was high heterogeneity between studies (I² 98%). The pooled NPV estimate for BAC (3 data points) was moderate: 0.69 (95%CI=0.52-0.86) and there was high heterogeneity between studies (I² 93%).

**Carbohydrate deficient transferrin (CDT)**
There are no alpha and kappa coefficients associated with biomarkers such as CDT. The pooled sensitivity estimate for CDT (8 data points) was low: 0.59 (95%CI=0.43-0.73) and there was high heterogeneity between studies (I² 97%). The pooled specificity estimate for CDT (8 data points) was excellent: 0.96 (95%CI=0.93-0.98) and there was moderate heterogeneity between studies (I² 72%). The pooled PPV estimate for CDT (6 data points) was good: 0.85 (95%CI=0.74-0.97) and there was high heterogeneity between studies (I² 76%). The pooled NPV estimate for CDT (6 data points) was moderate: 0.79 (95%CI=0.73-0.85) and there was high heterogeneity between studies (I² 96%).

**Carbohydrate deficient transferrin-Tech (CDTech)**
There are no alpha and kappa coefficients associated with biomarkers such as CDTech. The pooled sensitivity estimate for CDTech (41 data points) was low: 0.54 (95%CI=0.45-0.62) and there was high heterogeneity between studies (I² 99%). The pooled specificity estimate for CDTech (41 data points) was good: 0.89 (95%CI=0.88-0.91) and there was high heterogeneity between studies (I² 88%). The pooled PPV estimate for CDTech (12
data points) was low: 0.52 (95%CI=0.37-0.67) and there was high heterogeneity between studies ($I^2$ 95%). The pooled NPV estimate for CDTech (8 data points) was moderate: 0.80 (95%CI=0.61-0.98) and there was high heterogeneity between studies ($I^2$ 99%).

**Carbohydrate deficient transferrin with Mean corpuscular volume (CDT with MCV)**

There are no alpha and kappa coefficients associated with biomarkers such as CDT and MCV. The pooled sensitivity estimate for CDT with MCV (8 data points) was moderate: 0.74 (95%CI=0.60-0.88) and there was high heterogeneity between studies ($I^2$ 98%). The pooled specificity estimate for CDT with MCV (4 data points) was excellent: 0.93 (95%CI=0.91-0.95) and there was low heterogeneity between studies ($I^2$ 0%). The pooled PPV estimate for CDT with MCV (4 data points) was moderate: 0.74 (95%CI=0.51-0.97) and there was high heterogeneity between studies ($I^2$ 98%). The pooled NPV estimate for CDT with MCV (4 data points) was excellent: 0.92 (95%CI=0.83-1.00) and there was high heterogeneity between studies ($I^2$ 95%)

**Gamma-Glutamyl Transferase (GGT)**

There are no alpha and kappa coefficients associated with biomarkers such as GGT. The pooled sensitivity estimate for GGT (76 data points) was low: 0.57 (95%CI=0.50-0.64) and there was high heterogeneity between studies ($I^2$ 99%). The pooled specificity estimate for GGT (76 data points) was good: 0.83 (95%CI=0.78-0.86) and there was high heterogeneity between studies ($I^2$ 98%). The pooled PPV estimate for GGT (30 data points) was low: 0.43 (95%CI=0.35-0.51) and there was high heterogeneity between studies ($I^2$ 97%). The pooled NPV estimate for GGT (23 data points) was good: 0.82 (95%CI=0.70-0.94) and there was high heterogeneity between studies ($I^2$ 99%).

**Gamma-Glutamyl Transferase with Mean corpuscular volume (GGT with MCV)**

There are no alpha and kappa coefficients associated with biomarkers such as GGT and MCV. The pooled sensitivity estimate for GGT with MCV (10 data points) was moderate: 0.64 (95%CI=0.38-0.84) and there was high heterogeneity between studies ($I^2$ 99%). The pooled specificity estimate for GGT with MCV (10 data points) was good: 0.87 (95%CI=0.76-0.93) and there was high heterogeneity between studies ($I^2$ 97%). The
pooled PPV estimate for GGT with MCV (6 data points) was low: 0.47 (95%CI=0.28-0.66) and there was high heterogeneity between studies (I² 98%). The pooled NPV estimate for GGT with MCV (6 data points) was good: 0.88 (95%CI=0.81-0.95) and there was high heterogeneity between studies (I² 94%).

**Ethyl glucuronide (EtG)**

There are no alpha and kappa coefficients associated with biomarkers such as EtG. The pooled sensitivity estimate for EtG (6 data points) was good: 0.83 (95%CI=0.61-0.94) and there was high heterogeneity between studies (I² 91%). The pooled specificity estimate for EtG (6 data points) was excellent: 0.95 (95%CI=0.90-0.98) and there was high heterogeneity between studies (I² 66%). The pooled PPV estimate for EtG (2 data points) was moderate: 0.61 (95%CI=0.39-0.84) and there was moderate heterogeneity between studies (I² 58%). The pooled NPV estimate for EtG (2 data points) was good: 0.86 (95%CI=0.78-0.94) and there was moderate heterogeneity between studies (I² 60%).

**Mean corpuscular volume (MCV)**

There are no alpha and kappa coefficients associated with biomarkers such as MCV. The pooled sensitivity estimate for MCV (55 data points) was low: 0.39 (95%CI=0.33-0.45) and there was high heterogeneity between studies (I² 97%). The pooled specificity estimate for MCV (55 data points) was excellent: 0.91 (95%CI=0.88-0.93) and there was high heterogeneity between studies (I² 98%). The pooled PPV estimate for MCV (28 data points) was low: 0.48 (95%CI=0.36-0.59) and there was high heterogeneity between studies (I² 98%). The pooled NPV estimate for MCV (22 data points) was moderate: 0.79 (95%CI=0.73-0.86) and there was high heterogeneity between studies (I² 99%).

**Percent Carbohydrate deficient transferrin (%CDT)**

The pooled sensitivity estimate for %CDT (40 data points) was low: 0.56 (95%CI=0.47-0.65) and there was high heterogeneity between studies (I² 98.2%). The pooled specificity estimate for %CDT (40 data points) was 0.91, which is considered as excellent (95%CI=0.88-0.94) and there was high heterogeneity between studies (I² 97%). The pooled PPV estimate for %CDT (13 data points) was low: 0.58 (95%CI=0.38-0.78) and there was high heterogeneity between studies (I² 98.5%). The pooled NPV estimate for
%CDT (13 data points) was good: 0.85 (95%CI=0.78-0.92) and there was high heterogeneity between studies ($I^2$ 97.6%).

**Phosphatidylethanol (PEth)**

There are no alpha and kappa coefficients associated with biomarkers such as PEth. The pooled sensitivity estimate for PEth (7 data points) was good: 0.87 (95%CI=0.79-0.96) and there was high heterogeneity between studies ($I^2$ 94%). The pooled specificity estimate for PEth (4 data points) was excellent: 0.94 (95%CI=0.91-0.97) and there was moderate heterogeneity between studies ($I^2$ 31%). There was insufficient data to calculate the pooled estimate for PPV and NPV for PEth.

**Discussion**

In this systematic review and meta-analysis, we identified 387 unique papers that have published data on the validity, reliability and diagnostic accuracy of 37 scales for substance classes that are associated with HIV risk. We observed based on meta-analyzable data available, that fourteen of the thirty-seven measures/scales (38%) that had all pooled estimates consistently meet criteria for acceptability (e.g., ranging between fair/moderate-to-excellent), which included the following self-reported measures:

- Alcohol Dependence Scale
- Addiction Severity Index (ASI)
- ASI subscale for Alcohol; ASSIST
- Composite International Diagnostic Interview
- Drug Abuse Screen Test - 10 item scale
- Drug Use Disorders Identification Test
- Problem Oriented Screening Instrument for Teenagers
- Severity of Dependence scale
- Timeline Followback
- Chemical Use, Abuse, and Dependence

Biomarkers that had all pooled estimates that were fair/moderate or better include the following:

- Ethyl glucuronide
- Phosphatidylethanol test
- The combined used of Carbohydrate deficient transferrin and Mean corpuscular volume.

Taken together, our findings highlight the availability of a promising range of tools for researchers and practitioners when assessing substance use, particularly those working with classes of substances associated with HIV risk, such as heroin, methamphetamine,
cocaine, ecstasy, and alcohol. Nevertheless, further research is needed to determine why some substance use measures do not consistently have acceptable psychometric properties across different studies.

Overall, while most of the self-reported scales had acceptable validity, most did not have acceptable reliability: 65% of pooled estimates for alpha were in the range of fair-to-excellent though only 44% of estimates for kappa were in the range of fair-to-excellent. Moreover, a greater proportion of the scales we identified and meta-analyzed were better at correctly identifying individuals who are truly not using substances/not problematic users among those truly without these conditions (specificity: 97% of summary estimates) and among those who were deemed as not having this condition in the scale (negative predictive value: 96%). In contrast to specificity and negative predictive value estimates, fewer scales had pooled estimates on sensitivity and positive predictive value that were in the fair-to-excellent range (69% and 37%, respectively). These may have implications in the application of these measures in different settings. For example, in the criminal justice system, it may be better to utilize measures that have high specificity and negative predictive properties if the priority is to avoid false-positive results. However, in health settings, it may be more ideal to use measures with better sensitivity and positivity to better capture individuals who may require further assessment for substance use disorder assessments and treatment referrals, as appropriate.

Overall, the studies identified in this review had administered scales in English, were conducted within in the United States, and were less commonly tested among exclusively-women samples (there were twice as many exclusively-men samples in comparison). These findings highlight the general lack of diversity in terms of language, setting, and study population for the studies reporting validity, reliability, and diagnostic accuracy on substance use measures. Given the high morbidity and mortality associated with substance use globally and for different risk populations, greater effort is needed to further evaluate the psychometric properties of substance use measures in such samples. This study also found that few papers on substance use psychometric properties are “low risk” across all QUADAS 2 domains (16%). This finding highlights the need to further study the validity, reliability, and diagnostic accuracy of substance use measures using studies designed with better methodological rigor to reduce risk of bias.

This present study has several limitations. First, our inclusion criteria may have excluded
some potentially relevant studies on the psychometric properties of substance use measures that were not published in English. Hence, although we included measures that were not administered in English as long as they were published in English, our findings may not necessarily be generalizable to the psychometric properties of non-English measures that were not published in English. It should also be noted that our eligibility criteria likely favored the inclusion of studies that were conducted in settings where English proficiency was higher, which is correlated with countries with higher gross national income per capita (42). Moreover, while our search strategy was developed to try and identify all the relevant studies, many publications that have calculated our psychometric properties of interest may not have language referencing the specific key words/terms in our strategy in their titles and/or abstracts. In particular, this may occur because the psychometric data of scales may not be considered a “primary outcome” of a study, and thus not be highlighted in the title or abstract (i.e., the relevant data are imbedded within the full-text only). Additionally, while we did not specifically seek out studies only among HIV-risk populations, per se, our study did focus on substance classes that have been associated with HIV risk, namely alcohol, stimulants (methamphetamine, amphetamine, cocaine, ecstasy), and heroin. Hence, our search may have missed studies on more general substance use measures that did not explicitly name our targeted substance classes. Therefore, our findings should be interpreted with this limitation in mind. Furthermore, we were unable to calculate pooled estimates for some psychometric outcomes of several measures due to lack of published data. Further research to fill our gaps in knowledge on the psychometric properties of these substance use measures.

To our knowledge, this is the first systematic review and meta-analysis involving the synthesis of psychometric data across different measures of substances that are associated with HIV risk. As mentioned, limited research has been conducted with respect with quantitatively pooling the psychometric characteristics of substance use measures. Our findings highlight the general strengths of many substance use measures with respect to their validity, reliability, and diagnostic accuracy across multiple studies/samples. To facilitate the dissemination of these findings, and provide researchers with a resource to identify validated, reliable, and accurate measures for substance use, we collaborated with members of the [BLINDED] Scientific Committee to develop a web-based tool, with the results of the pooled summary estimates presented in this study. The tool, named “Substance Use Measure Identification (SUMI) Tool” is available as a free resource in the [BLINDED] website (URL: [BLINDED]).
In summary, researchers in the field of substance use should endeavor to conduct more validity, reliability, and diagnostic accuracy studies on measures to identify substance use and use disorders among more diverse settings and populations, and with more rigorous study designs. Ultimately, accurate identification of substance users and problematic substance use is a critical step in identifying individuals for substance use treatment and evaluating the effectiveness of treatment strategies. Hence, further evaluation of substance use measures is of great importance not only to the field of substance use research, but also substance use treatment. Given the substantial contribution of substance use to the global burden of disease (5), having robust data on the psychometric properties of substance use measure can help researchers identify the best tools to use in research studies, further enhancing the collection of more valid, reliable, accurate data to inform evidence-based responses to substance use.

Declarations

- Ethics approval and consent to participate

This study involved only analysis of data from published scientific literature; we did not collect any primary data.

- Consent for publication

Not applicable.

- Competing interests

Authors declare no competing interests.
-Funding

This study was supported by HPTN, which receives its funding from three NIH Institutes: the National Institute of Allergy and Infectious Diseases, the National Institute of Mental Health and the National Institute on Drug Abuse (Grant # UM1 AI068619). No funding bodies had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

-Author Contributions:

GMS performed the data analysis, interpreted the data, and let the preparation of the manuscript. SAS, NE, SS, designed the study with GMS, and contributed to data interpretation and revising the manuscript critically for important intellectual content. PP, DS, DH, RC, CM, PM, FC, BK performed the systematic search, and data extraction, and contributed to data interpretation and revising the manuscript critically for important intellectual content. IA provided input on the data analysis and revise the manuscript critically for important intellectual content.

-Acknowledgements

We would like to thank Evans Whitaker, MD, MLIS from University of California San Francisco library for his assistance with the development and execution of the search strategy. We also thank the members of the HPTN Substance Use Scientific Committee for the feedback they provided on this project.
Tables

Due to technical limitations, tables are only available as a download in the supplemental files section

References

1. United Nations Office on Drugs and Crime. World Drug Report 2017. Vienna, Austria: United Nations Office on Drugs and Crime; 2017.

2. World Health Organization. Management of Substance Abuse: Alcohol. World Health Organization; 2017 [Available from: http://www.who.int/substance_abuse/facts/alcohol/en/.

3. G. B. D. Disease Injury Incidence Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017;390(10100):1211-59.

4. Degenhardt L, Whiteford HA, Ferrari AJ, Baxter AJ, Charlson FJ, Hall WD, et al. Global burden of disease attributable to illicit drug use and dependence: findings from the Global Burden of Disease Study 2010. Lancet. 2013;382(9904):1564-74.

5. G. B. D. Risk Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388(10053):1659-724.
6. Colfax G, Santos GM, Chu P, Vittinghoff E, Pluddemann A, Kumar S, et al. Amphetamine-group substances and HIV. Lancet. 2010;376(9739):458-74.

7. Santos GM, Das M, Colfax GN. Interventions for non-injection substance use among US men who have sex with men: what is needed. AIDS Behav. 2011;15 Suppl 1:S51-6.

8. Shoptaw S, Montgomery B, Williams CT, El-Bassel N, Aramrattana A, Metsch L, et al. Not just the needle: the state of HIV-prevention science among substance users and future directions. Journal of acquired immune deficiency syndromes. 2013;63 Suppl 2:S174-8.

9. Strathdee SA, Shoptaw S, Dyer TP, Quan VM, Aramrattana A, Substance Use Scientific Committee of the HIVPTN. Towards combination HIV prevention for injection drug users: addressing addictophobia, apathy and inattention. Current opinion in HIV and AIDS. 2012;7(4):320-5.

10. Rowe C, Santos GM, McFarland W, Wilson EC. Prevalence and correlates of substance use among trans female youth ages 16-24 years in the San Francisco Bay Area. Drug and alcohol dependence. 2015;147:160-6.

11. Santos GM, Coffin PO, Das M, Matheson T, DeMicco E, Raiford JL, et al. Dose-response associations between number and frequency of substance use and high-risk sexual behaviors among HIV-negative substance-using men who have sex with men (SUMSM) in San Francisco. J Acquir Immune Defic Syndr. 2013;63(4):540-4.

12. Ostrow DG, Plankey MW, Cox C, Li X, Shoptaw S, Jacobson LP, et al. Specific sex drug combinations contribute to the majority of recent HIV seroconversions among MSM in the MACS. Journal of acquired immune deficiency syndromes. 2009;51(3):349-55.

13. Koblin BA, Husnik MJ, Colfax G, Huang Y, Madison M, Mayer K, et al. Risk factors for HIV infection among men who have sex with men. AIDS. 2006;20(5):731-9.
14. Kerr T, Shannon K, Ti L, Strathdee S, Hayashi K, Nguyen P, et al. Sex work and HIV incidence among people who inject drugs. Aids. 2016;30(4):627-34.

15. Strathdee SA, Galai N, Safaiean M, Celentano DD, Vlahov D, Johnson L, et al. Sex differences in risk factors for hiv seroconversion among injection drug users: a 10-year perspective. Archives of internal medicine. 2001;161(10):1281-8.

16. Hinkin CH, Barclay TR, Castellon SA, Levine AJ, Durvasula RS, Marion SD, et al. Drug use and medication adherence among HIV-1 infected individuals. AIDS and behavior. 2007;11(2):185-94.

17. DeLorenze GN, Weisner C, Tsai AL, Satre DD, Quesenberry CP, Jr. Excess mortality among HIV-infected patients diagnosed with substance use dependence or abuse receiving care in a fully integrated medical care program. Alcoholism, clinical and experimental research. 2011;35(2):203-10.

18. Chander G, Himelhoch S, Moore RD. Substance abuse and psychiatric disorders in HIV-positive patients: epidemiology and impact on antiretroviral therapy. Drugs. 2006;66(6):769-89.

19. Dhalla S, Zumbo BD, Poole G. A review of the psychometric properties of the CRAFFT instrument: 1999-2010. Current drug abuse reviews. 2011;4(1):57-64.

20. Berman AH, Bergman H, Palmstierna T, Schlyter F. Evaluation of the Drug Use Disorders Identification Test (DUDIT) in criminal justice and detoxification settings and in a Swedish population sample. European addiction research. 2005;11(1):22-31.

21. Berner MM, Kriston L, Bentele M, Harter M. The alcohol use disorders identification test for detecting at-risk drinking: a systematic review and meta-analysis. Journal of studies on alcohol and drugs. 2007;68(3):461-73.

22. Substance Abuse and Mental Health Services Administration (SAMHSA). The Role of Biomarkers in the Treatment of Alcohol Use Disorders. SAMHSA Advisory. 2012;11(2).
23. Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. General hospital psychiatry. 2015;37(1):67-75.

24. Stockings E, Degenhardt L, Lee YY, Mihalopoulos C, Liu A, Hobbs M, et al. Symptom screening scales for detecting major depressive disorder in children and adolescents: a systematic review and meta-analysis of reliability, validity and diagnostic utility. Journal of affective disorders. 2015;174:447-63.

25. Mitchell AJ, Coyne JC. Do ultra-short screening instruments accurately detect depression in primary care? A pooled analysis and meta-analysis of 22 studies. The British journal of general practice : the journal of the Royal College of General Practitioners. 2007;57(535):144-51.

26. Scaini S, Battaglia M, Beidel DC, Ogliari A. A meta-analysis of the cross-cultural psychometric properties of the Social Phobia and Anxiety Inventory for Children (SPAI-C). Journal of anxiety disorders. 2012;26(1):182-8.

27. Newton AS, Soleimani A, Kirkland SW, Gokiert RJ. A Systematic Review of Instruments to Identify Mental Health and Substance Use Problems Among Children in the Emergency Department. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine. 2017;24(5):552-68.

28. Newton AS, Gokiert R, Mabood N, Ata N, Dong K, Ali S, et al. Instruments to detect alcohol and other drug misuse in the emergency department: a systematic review. Pediatrics. 2011;128(1):e180-92.

29. Mitchell AJ, Bird V, Rizzo M, Hussain S, Meader N. Accuracy of one or two simple questions to identify alcohol-use disorder in primary care: a meta-analysis. The British journal of general practice : the journal of the Royal College of General Practitioners. 2014;64(624):e408-18.
30. Dhalla S, Kopec JA. The CAGE questionnaire for alcohol misuse: a review of reliability and validity studies. Clinical and investigative medicine Medecine clinique et experimentale. 2007;30(1):33-41.

31. Allen JP, Reinert DF, Volk RJ. The alcohol use disorders identification test: an aid to recognition of alcohol problems in primary care patients. Preventive medicine. 2001;33(5):428-33.

32. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011;155(8):529-36.

33. Harris RJ, Bradburn MJ, Deeks JJ, Harbord RM, Altman D, Sterne JA. metan: fixed- and random-effects meta-analysis. The Stata Journal. 2008;8(1):3-28.

34. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 102010.

35. Freeman K, Taylor-Phillips S, Connock M, Court R, Tsertsvadze A, Shyangdan D, et al. Test accuracy of drug and antibody assays for predicting response to antitumour necrosis factor treatment in Crohn's disease: a systematic review and meta-analysis. BMJ open. 2017;7(6):e014581.

36. Harbord RM, . metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. The Stata Journal. 2009;9(2):211-29.

37. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8(2):239-51.

38. Ponterotto JG, Ruckdeschel DE. An overview of coefficient alpha and a reliability matrix for estimating adequacy of internal consistency coefficients with psychological research measures. Perceptual and motor skills. 2007;105(3 Pt 1):997-1014.
39. Andrews JA, Lewinsohn PM, Hops H, Roberts RE. Psychometric properties of scales for the measurement of psychosocial variables associated with depression in adolescence. Psychological reports. 1993;73(3 Pt 1):1019-46.

40. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Bmj. 2003;327(7414):557-60.

41. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. Journal of clinical epidemiology. 2005;58(9):882-93.

42. McCormick C. Correlates of Validity of Self-Reported Methamphetamine Use Among a Sample of Dependent Adults. Harvard Business Review. 2013;2013(11).

Figures
Figure 1

Study Identification, Screening, Eligibility, and Inclusion for Meta-Analysis
Figure 2: Distribution of Pooled Summary Estimates of Psychometric Outcomes

Distribution of Pooled Summary Estimates of Psychometric Outcomes
Figure 3: Overall Summary of study quality ratings from the Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies, QUADAS-2

Supplementary Files
This is a list of supplementary files associated with the primary manuscript. Click to download.

PRISMA 2009 checklist_BMCMRM.doc
HPTN Meta-Analysis Search Terms.docx
Tables.docx