BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Centre, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2026, Volume 17, Issue 2, pages: 514-533
Submitted: March 19th, 2026 | *Accepted for publication:* May 3rd, 2026

What do Brief Substance Use Screening Instruments Actually Measure? A Stratified Meta-Analysis in Correctional Populations

Dan Octavian Rusu
Department of Applied Psychology, Babes-Bolyai University of Cluj-Napoca, 4305849, Romania
https://orcid.org/0000-0002-8432-1965

Cristian Delcea*
Multidisciplinary Doctoral School, Vasile Goldiș Western University of Arad, 310025 Arad, Romania
delcea.cristian@uvvg.ro
https://orcid.org/0000-0003-0667-2898

Ionut-Virgil Șerban*
University of Craiova, A. I. Cuza Str., No. 13, Craiova, 200585, Romania; fellow at the University of Chieti-Pescara; the University "Kore", Enna; and the University of International Studies in Rome (Unint), Italy.
ionut.serban@edu.ucv.ro,
johnutzserban@yahoo.com,
https://orcid.org/0000-0001-7240-9989

Abstract: *Substance use disorders are highly prevalent in adult correctional and forensic populations. However, brief screening instruments are often interpreted without clear differentiation between diagnostic, validation, and predictive purposes. In this study, we synthesize evidence on commonly used substance use screening tools using a stratified meta-analytic framework designed to clarify their legitimate inferential roles in custodial settings. Evidence was organized into three analytic tiers: Tier 1 (CORE: Diagnostic Accuracy) included studies permitting formal estimation of diagnostic accuracy against explicit clinical reference standards. Tier 2 (Extended Forensic Validation) comprised extended forensic validation studies employing context-specific or severity-based frameworks. Tier 3 (Predictive Validity) addressed predictive validity for substance-relevant post-release outcomes. Quantitative synthesis was restricted to Tier 1 studies and indicated high sensitivity with moderate specificity for brief screening instruments when evaluated against structured diagnostic assessments. Given the limited number of eligible studies, these pooled estimates should be interpreted as preliminary indicators rather than stable population parameters. Tier 2 studies demonstrated broadly consistent performance across diverse forensic contexts but substantial heterogeneity in reference standards, precluding pooled diagnostic inference. Limited Tier 3 evidence suggested that screening-derived severity classifications may be associated with substance-relevant post-release outcomes. Overall, the findings indicate that brief screening instruments support distinct, tier-specific functions. Evidence for one inferential purpose should not be generalized to others.*

Keywords: *substance use screening; correctional populations; forensic assessment; diagnostic accuracy; predictive validity.*

How to cite: Rusu, D. O., Delcea, C., & Șerban, I.-V. (2026). *What do brief substance use screening instruments actually measure? A stratified meta-analysis in correctional populations. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 17*(2), 514-533. https://doi.org/10.70594/brain/17.2/32

**Corresponding authors*

1. Introduction

Substance use disorders (SUDs) are markedly overrepresented in incarcerated adult populations and constitute a persistent challenge for correctional systems worldwide. Large-scale epidemiological syntheses indicate that between one third and one half of incarcerated adults meet criteria for a current alcohol or drug use disorder (Rusu Andron et al., 2025), rates substantially exceeding those observed in the general population (Fazel et al., 2017; Kinner & Wang, 2014). This disproportionate burden is further compounded by high levels of psychiatric comorbidity, social marginalisation, and prior criminal justice involvement, all of which complicate assessment, treatment planning, and post-release continuity of care (Baillargeon et al., 2009; Butler et al., 2011; Prins, 2014).

Within this context, brief substance use screening instruments have become a cornerstone of correctional and forensic practice. Tools such as the Alcohol Use Disorders Identification Test (AUDIT); the Drug Use Disorders Identification Test (DUDIT); the Alcohol, Smoking, and Substance Involvement Screening Test (ASSIST); the Rapid Opioid Dependence Screen (RODS); and ultra-brief instruments such as the UNCOPE are routinely administered at intake, during incarceration, or prior to release. Their appeal lies in their brevity, low training requirements, and feasibility in resource-constrained custodial settings (Babor et al., 2001; Humeniuk et al., 2008; Peters et al., 2000). However, despite widespread implementation, the empirical evidence base supporting these instruments in correctional populations remains conceptually fragmented.

A central problem in the existing literature is the frequent conflation of distinct forms of validity, particularly diagnostic accuracy and predictive validity. Classical validity theory emphasises that validity is not a property of an instrument per se, but rather of the interpretations and uses of its scores (Cronbach & Meehl, 1955; Messick, 1995). In the context of substance use screening, this distinction is especially salient. Diagnostic accuracy addresses whether a screening instrument can correctly classify individuals with respect to a contemporaneous reference standard (e.g., DSM- or ICD-based diagnosis), typically quantified using sensitivity, specificity, and related indices. Predictive validity, by contrast, concerns whether screening scores are associated with future outcomes, such as post-release substance-related outcomes. These two forms of evidence answer fundamentally different questions and cannot be assumed to co-occur (Messick, 1995).

In correctional research, this distinction has often been blurred. Many studies report psychometric properties such as internal consistency or factor structure and implicitly extrapolate these findings to clinical or prognostic utility, despite limited evidence that such properties translate into accurate classification or prediction in offender populations (Hartzler et al., 2014; Knight et al., 1999). Conversely, some studies on post-release outcomes cite screening performance without clearly anchoring results to a diagnostic reference standard, thereby mixing concurrent and predictive interpretations. This lack of conceptual clarity has contributed to substantial heterogeneity in the literature, impeding cumulative quantitative synthesis.

Nevertheless, a subset of studies has provided diagnostic accuracy evidence grounded in adult forensic or correctional samples. Wickersham et al. (2015) demonstrated that the Rapid Opioid Dependence Screen (RODS) achieved high sensitivity against a MINI-based opioid dependence diagnosis in newly incarcerated adults. Similarly, Evren et al. (2014) evaluated the Drug Use Disorders Identification Test (DUDIT) against a clinically defined drug use disorder in a prison population, providing extractable diagnostic accuracy parameters. These studies offer rare examples of diagnostically anchored screening evidence suitable for formal quantitative synthesis.

A broader research body has examined screening instrument performance in forensic settings using clinically informed, but context-specific, reference standards. Studies evaluating instruments such as the AUDIT, DUDIT, UNCOPE, and ASSIST, as well as culturally adapted tools such as the Indigenous Risk Impact Screen (IRIS), have demonstrated generally favourable screening performance across diverse correctional populations, including individuals undergoing forensic psychiatric assessment, large-scale prison cohorts, and culturally specific incarcerated groups (Berman et al., 2004; Durbeej et al., 2010; Ober et al., 2013; Peters et al., 2000; Proctor &

Hoffmann, 2016; Wolff & Shi, 2015). While these studies contribute important evidence regarding robustness and feasibility, their heterogeneity in reference standards, outcome definitions, and scoring frameworks limits their suitability for pooled diagnostic accuracy estimation.

A parallel but conceptually distinct body of work has examined predictive validity, particularly with respect to substance-related post-release outcomes. From a public health perspective (Calderaro et al., 2025), post-release substance involvement is a critical outcome, given its association with overdose mortality, treatment discontinuity, and reincarceration risk (Binswanger et al., 2013; Merrall et al., 2010; Gendreau, Little, & Goggin, 1996; Baillargeon et al., 2010). However, predictive studies address a different inferential target than diagnostic accuracy and must, therefore, be synthesised separately.

Another source of heterogeneity concerns population definition, particularly the inclusion of adolescent samples. Developmental research consistently demonstrates that adolescent substance use (Blendea et al., 2025) differs qualitatively from adult substance use with respect to neurodevelopment, social context, and diagnostic stability (Steinberg & Morris, 2001). Accordingly, we only included adult correctional and forensic populations in this synthesis.

To address these limitations, we adopted a stratified (tiered) meta-analytic approach. The first tier synthesises studies permitting formal estimation of diagnostic accuracy against explicit reference standards. The second tier integrates extended forensic validation studies that inform generalisability without contributing to pooled estimates. The third tier addresses predictive validity for substance-relevant post-release outcomes. This structure aligns validity theory, population definition, and analytical strategy, ensuring conceptual coherence and methodological transparency (Cronbach & Meehl, 1955; Messick, 1995; Page et al., 2021; Taxman et al., 2007b).

2. Methods

2.1. Study Design and Conceptual Framework

We employed a systematic review and stratified meta-analytic design to synthesise evidence on the performance of brief substance use screening instruments in adult correctional and forensic populations. The analytical framework was explicitly grounded in classical and contemporary validity theory, which conceptualises validity as pertaining to score interpretation and use rather than as an intrinsic property of an instrument (Cronbach & Meehl, 1955; Messick, 1995). Accordingly, the synthesis distinguished between diagnostic accuracy and predictive validity as fundamentally different evidentiary domains.

To avoid construct conflation and inappropriate quantitative aggregation, a tiered analytic strategy was adopted. Studies were classified into three analytic tiers based on (a) the nature of the reference standard, (b) outcome definition, and (c) the intended function of the screening instrument. This approach aligns with methodological guidance for diagnostic test accuracy reviews, which emphasises conceptual homogeneity over sample size or statistical power (Leeflang et al., 2008; Riley et al., 2020).

2.2. Eligibility Criteria

Studies were eligible for inclusion if they met the following criteria: (a) included adult participants (≥18 years) drawn from correctional or forensic settings (e.g., prisons, jails, or forensic psychiatric assessments); (b) evaluated a brief substance use screening instrument intended for clinical or operational use; and (c) reported sufficient methodological detail to allow classification into one of the predefined analytic tiers.

Studies were excluded if they (a) focused exclusively on adolescent samples; (b) evaluated instruments not intended for screening; or (c) examined outcomes unrelated to substance use or substance-relevant post-release outcomes.

2.3. Analytic Tiers and Study Classification

2.3.1. Tier 1 (CORE: Diagnostic Accuracy)

Tier 1 comprised studies evaluating a brief screening instrument against an explicit diagnostic or quasi-diagnostic reference standard and reporting sufficient data to permit reconstruction of 2 × 2 contingency tables. Only Tier 1 studies were eligible for quantitative meta-analytic synthesis.

2.3.2. Tier 2 (Extended Forensic Validation)

Tier 2 included studies conducted in forensic or correctional populations that evaluated screening instruments against clinically informed but context-specific reference standards or severity-based classifications. These studies were synthesised narratively.

2.3.3. Tier 3 (Predictive Validity)

Tier 3 was reserved for studies examining associations between screening-derived severity or risk classifications obtained during incarceration and subsequent substance-relevant post-release outcomes measured longitudinally. Substance-relevant outcomes were operationalised to include post-release measures with established empirical links to substance use relapse in correctional cohorts, including reincarceration outcomes.

2.4. Search Strategy and Study Selection

A systematic literature search was conducted between 10 December 2025 and 31 January 2026. Electronic databases searched included PubMed/MEDLINE (via the National Center for Biotechnology Information [NCBI]), Scopus (Elsevier platform), Web of Science Core Collection (Clarivate Analytics), and PsycINFO (via EBSCOhost). Searches were conducted from database inception to 31 January 2026. No publication year restrictions were applied. Only studies published in the English language were considered eligible.

The search strategy combined controlled vocabulary terms (e.g., Medical Subject Headings [MeSH] in PubMed) and free-text keywords across three conceptual domains: (1) substance use screening instruments (e.g., “AUDIT”, “DUDIT”, “RODS”, “substance use screening”, “screening tool”, and “substance use disorder”); (2) forensic or correctional populations (e.g., “prison”, “incarcerated”, “correctional”, “forensic population”, “detention”, and “probation”); and (3) validity constructs (e.g., “diagnostic accuracy”, “sensitivity”, “specificity”, “validation”, and “predictive validity”). Boolean operators (AND/OR) were used to combine search domains.

In addition to database searches, reference lists of all included studies were manually screened to identify further eligible publications. Study selection was performed using Rayyan, with a liberal inclusion strategy applied at the title and abstract screening stage. Full-text articles were assessed against tier-specific eligibility criteria, consistent with PRISMA 2020 recommendations (Page et al., 2021).

Title and abstract screening, followed by full-text eligibility assessment, was conducted independently by two reviewers. Discrepancies were resolved through discussion until consensus was reached. When necessary, a third reviewer adjudicated unresolved disagreements. Both reviewers have formal training in clinical psychology and quantitative research methodology, including systematic review methodology and meta-analytic techniques, with specific expertise in forensic populations and psychometric validation of screening instruments.

The complete search strategies for each database are provided in Supplementary Material S1 to ensure transparency and reproducibility.

2.5. Data Extraction and Synthesis

Data extraction was conducted independently by two reviewers, with discrepancies resolved through discussion and, where necessary, consultation with a third reviewer. Data extraction and

synthesis procedures followed established guidance for diagnostic accuracy reviews. Quantitative synthesis was restricted to Tier 1 studies, using hierarchical approaches to summarise sensitivity and specificity (Reitsma et al., 2005; Riley et al., 2020). No formal assessment of publication bias was conducted, consistent with methodological recommendations (Deeks, Macaskill, & Irwig, 2005; Macaskill et al., 2010). All statistical analyses were conducted using R statistical software (version 4.3.2; R Foundation for Statistical Computing, Vienna, Austria). Meta-analytic procedures were performed using the meta package (version 6.5-0) and the mada package (version 0.5.11) for diagnostic accuracy synthesis. Data management and transformation were conducted using the tidyverse package (version 2.0.0).

For Tier 1 analyses, pooled sensitivity and specificity were estimated using random-effects models. Between-study heterogeneity was quantified using the I^2 statistic and τ^2 estimates. All statistical tests were two-tailed, with statistical significance set at $p < 0.05$ where applicable.

2.6. Protocol and Registration

This study was not prospectively registered in PROSPERO. The study employed a stratified inferential framework separating diagnostic accuracy (Tier 1), extended forensic validation (Tier 2), and predictive validity (Tier 3), which does not align with conventional single-layer diagnostic review templates. All eligibility criteria, tier definitions, and analytic strategies were defined a priori before data extraction. The study was conducted and reported in accordance with PRISMA 2020 guidelines to ensure transparency and reproducibility.

2.7. Risk of Bias and Study Quality Assessment

Risk of bias assessment was conducted using a tier-specific approach aligned with the stratified analytic framework of the present review. This approach was adopted to ensure that methodological evaluation remained consistent with the distinct inferential targets and study designs represented across the three analytic tiers.

For Tier 1 (CORE) diagnostic accuracy studies, risk of bias was assessed using the QUADAS-2 tool (Whiting et al., 2011), a validated instrument specifically designed for the evaluation of diagnostic test accuracy studies. QUADAS-2 assesses four domains: patient selection, index test, reference standard, and flow and timing. Each domain was evaluated in terms of risk of bias, and the first three domains were additionally assessed for concerns regarding applicability to the review question. Assessments were based on reported study characteristics, including sampling procedures, independence of index and reference tests, blinding of assessors, and completeness of outcome data.

Given the heterogeneity of study designs, reference standards, and outcome definitions in Tier 2 (Extended Forensic Validation) and Tier 3 (Predictive Validity), formal risk-of-bias tools designed for diagnostic accuracy studies were not considered appropriate. Instead, a structured narrative appraisal was conducted for these tiers, focusing on key sources of bias and threats to validity relevant to their respective inferential targets. These included selection bias, variability and lack of standardisation in reference or comparator measures, absence of blinding, threshold heterogeneity, and construct-related limitations.

For Tier 3 studies, additional consideration was given to the nature of outcome definitions and the potential influence of institutional, policy, and contextual factors on post-release outcomes, particularly where distal or composite indicators such as reincarceration were used as proxies for substance-related outcomes.

This tier-specific strategy was intended to preserve conceptual coherence while allowing for a rigorous and context-sensitive evaluation of study quality across heterogeneous evidence domains.

3. Results

3.1. Overview of Study Selection and Analytic Structure

The systematic search and full-text assessment identified a heterogeneous body of literature evaluating brief substance use screening instruments in adult correctional and forensic populations. Following tier-specific eligibility assessment, a total of nine unique studies were included across the three analytic tiers. Two studies met the criteria for inclusion in Tier 1, six were classified within the extended forensic validation tier (Tier 2), and one examined predictive validity using longitudinal substance-relevant post-release outcomes (Tier 3). Each included study was assigned to a single analytic tier based on its primary inferential target, with no study contributing to more than one tier. All included studies are reported in the Results Section and correspond to records retained following full-text eligibility assessment in the PRISMA flow diagram.

The study selection process is summarised in Figure 1.

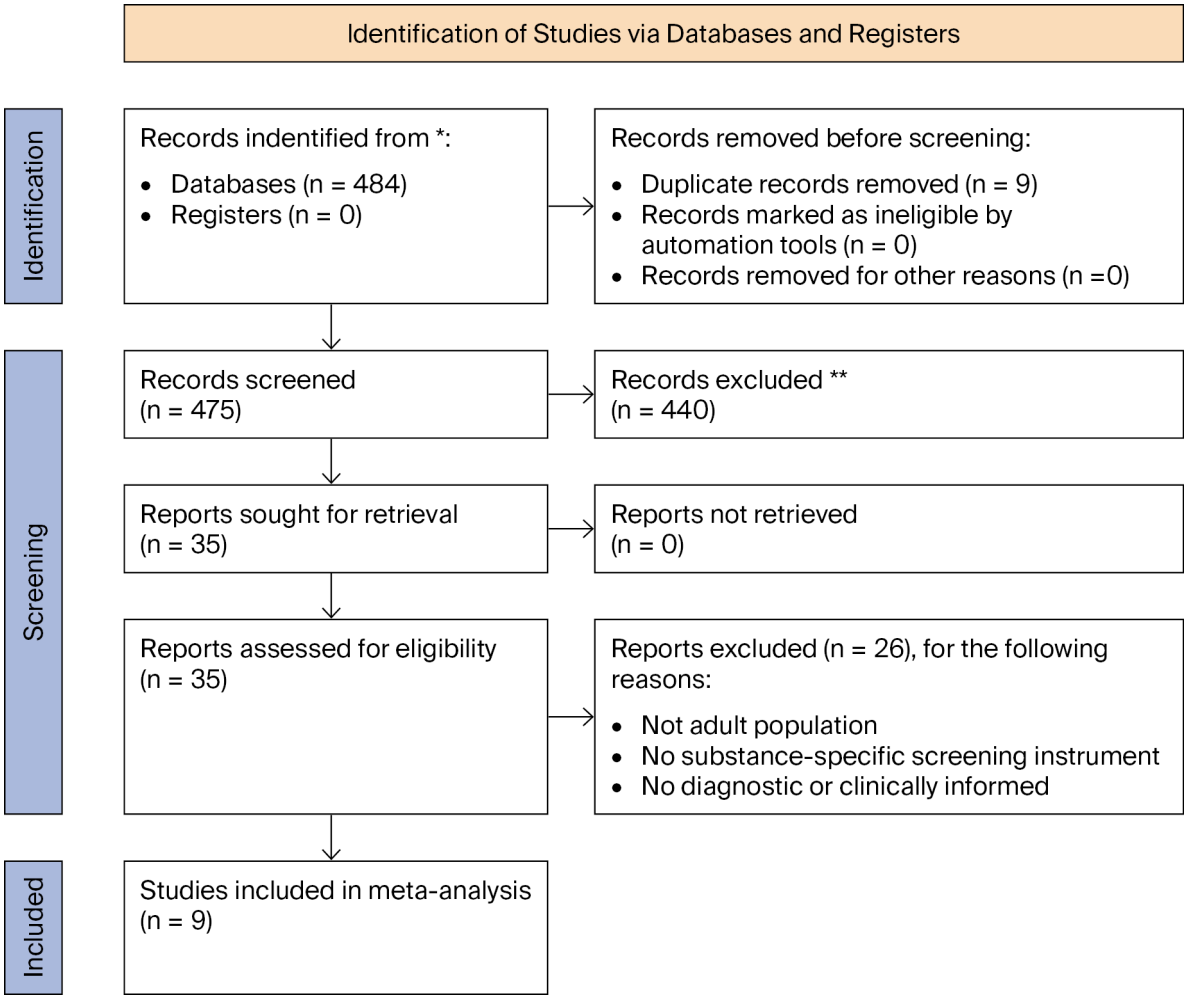

Figure 1. PRISMA 2020 flow diagram of the study identification, screening, eligibility assessment, and inclusion process for studies on brief substance use screening instruments in adult correctional and forensic populations.

** Records identified from databases include PubMed/MEDLINE, Scopus, Web of Science Core Collection, and PsycINFO.*

*** Records excluded at the screening stage were removed based on title and abstract review according to predefined eligibility criteria.*

Results are presented sequentially by analytic tier, beginning with the CORE diagnostic accuracy meta-analysis.

Table 1. Characteristics of studies included in the stratified meta-analysis

Author (Year)	Country	Setting	Population	Screening Instrument	Target Substance(s)	Analytic Tier	Study Purpose
Wickersham et al. (2015)	USA	Jail (newly incarcerated adults)	N = 97 adults; majority male; mean age ≈ mid-30s	Rapid Opioid Dependence Screen (RODS)	Opioids	Tier 1	Diagnostic accuracy vs. MINI
Evren et al. (2014)	Turkey	Prison	N = 202 incarcerated adults; majority male	Drug Use Disorders Identification Test (DUDIT)	Illicit drugs	Tier 1	Diagnostic accuracy vs. clinical diagnosis
Berman et al. (2004)	Sweden	Criminal justice and detoxification settings	Mixed forensic and treatment samples; adults	DUDIT	Illicit drugs	Tier 2	Forensic validation and psychometric evaluation
Durbeej et al. (2010)	Sweden	Forensic psychiatric assessment	N = 181 criminal suspects with mental health problems; 91% male; mean age = 33	AUDIT; DUDIT	Alcohol; illicit drugs	Tier 2	Concurrent validity vs. ASI and DSM-IV-based framework
Peters et al. (2000)	USA	Prison	N = 400 male inmates	Multiple instruments (e.g., TCU Drug Screen, SSI, and ADS/ASI)	Alcohol; drugs	Tier 2	Comparative screening effectiveness vs. SCID-IV
Proctor and Hoffmann (2016)	USA	State prison system (Minnesota DOC)	N = 7672 incarcerated adults (men and women); mean age ≈ 31	UNCOPE	Substance use disorders (multiple)	Tier 2	Screening performance vs. SUDDS-IV
Wolff and Shi (2015)	USA	Maximum-security prison	Adult incarcerated men; mean age ≈ 43	ASSIST (v3.0)	Alcohol; multiple drugs	Tier 2	Feasibility and validity vs. SCID-NP
Ober et al. (2013)	Australia	Prison (Indigenous population)	N = 395 Indigenous incarcerated adults; 84% male	Indigenous Risk Impact Screen (IRIS)	Alcohol; drugs; mental health	Tier 2	Concurrent validity vs. CIDI (ICD-10)
Knight et al. (1999)	USA	Prison and community follow-up	Adult incarcerated men enrolled in treatment programs	Salient Factor Score (SFS)	Drug-related severity (composite)	Tier 3	Predictive validity for post-release reincarceration

Note: Analytic tiers reflect the intended inferential target of each study. Tier 1 includes studies permitting formal diagnostic accuracy synthesis based on explicit reference standards. Tier 2 includes extended forensic validation studies informing robustness and applicability but not eligible for quantitative pooling. Tier 3 includes studies examining predictive validity for substance-relevant post-release outcomes.

3.2. Tier 1 (CORE: Diagnostic Accuracy) Meta-Analysis

Tier 1 comprised two studies evaluating brief substance use screening instruments against explicit diagnostic reference standards in adult correctional populations (Evren et al., 2014; Wickersham et al., 2015).

Table 2. Tier 1 (CORE: Diagnostic Accuracy) studies: index tests and reference standards

Author (Year)	Screening Instrument	Cut-Off Score	Reference Standard	Independence of Index and Reference Test	Blinding	Target Diagnosis	Sample Size
Wickersham et al. (2015)	Rapid Opioid Dependence Screen (RODS)	≥3	Mini International Neuropsychiatric Interview (MINI)	Yes	Yes	Opioid dependence	N = 97
Evren et al. (2014)	Drug Use Disorders Identification Test (DUDIT)	≥12	Clinically defined drug use disorder based on structured clinical assessment	Yes	Unclear	Drug use disorder	N = 202

Note: Tier 1 studies were required to employ an explicit diagnostic or quasi-diagnostic reference standard independent of the screening instrument and to report sufficient data to permit reconstruction of two-by-two contingency tables. “Blinding” refers to whether assessors of the reference standard were blinded to screening results.

3.2.1. Characteristics of Tier 1 Studies

Wickersham et al. (2015) evaluated the Rapid Opioid Dependence Screen (RODS) against the Mini International Neuropsychiatric Interview (MINI) in a sample of newly incarcerated adults. Diagnostic status was determined independently using the MINI, a structured diagnostic interview with established validity (Sheehan et al., 1998). Sufficient data were reported to allow direct calculation of sensitivity and specificity.

Evren et al. (2014) examined the Drug Use Disorders Identification Test (DUDIT) in an incarcerated adult sample, using a clinically defined drug use disorder as the reference standard. An optimal cut-off score was identified through receiver operating characteristic analysis, and reported sensitivity, specificity, and sample sizes permitted reconstruction of a 2 × 2 contingency table.
Across both studies, reference standards were applied independently of the index tests, minimising incorporation bias, and outcome definitions were aligned with established diagnostic frameworks.

3.2.2. Diagnostic Accuracy Estimates

At the study level, both instruments demonstrated high sensitivity for detecting substance use disorders. The RODS achieved a sensitivity of 0.97 and a specificity of 0.76 for opioid dependence (Wickersham et al., 2015). The DUDIT demonstrated a sensitivity of 0.95 and a specificity of 0.79 for drug use disorder (Evren et al., 2014). This pattern indicates strong case-finding capacity with more variable exclusion of non-cases (Table 3).

Table 3. Tier 1 (CORE: Diagnostic Accuracy) Results.

Author (Year)	True Positives (TP)	False Positives (FP)	False Negatives (FN)	True Negatives (TN)	Sensitivity	Specificity
Wickersham et al. (2015)	34	15	1	47	0.97	0.76
Evren et al. (2014)	96	22	5	79	0.95	0.79

Note: Two-by-two contingency tables were reconstructed from reported sensitivity, specificity, and sample size data. Sensitivity and specificity values correspond to the optimal cut-off scores reported in each primary study.

3.2.3. Risk of Bias Assessment (Tier 1)

Risk of bias for Tier 1 (CORE) studies was assessed using the QUADAS-2 tool (Whiting et al., 2011), which evaluates bias across four domains: patient selection, index test, reference standard, and flow and timing. The results of this assessment are summarised in Table 4.

Table 4. QUADAS-2 assessment for Tier 1 (CORE: Diagnostic Accuracy) studies.

Study	Patient Selection	Index Test	Reference Standard	Flow and Timing	Applicability Concerns
Wickersham et al. (2015)	Some concerns. Restricted sample (newly incarcerated, HIV-positive, unsentenced, near release); unclear sampling strategy; potential spectrum bias.	High [A3]. Evidence suggests refinement of RODS within the same validation sample against MINI, indicating lack of full prespecification.	Some concerns [A1]. MINI used; blinding not explicitly reported; SCID noted as a stronger alternative.	Low. Both tests administered at baseline; full sample analysed; no differential verification.	Some concerns. Narrow population limits generalisability.
Evren et al. (2014)	Some concerns. Convenience sample; refusals and exclusions; potential selection bias.	High. DUDIT threshold selected post hoc via ROC in the same sample; may inflate accuracy.	Some concerns [A1]. SCID-I used; blinding not explicitly reported.	Low. All participants included in ROC and 2 × 2 analyses; consistent application.	Low. Conceptually aligned; limitations mainly concern external validity.

Note. QUADAS-2 evaluates risk of bias across four domains and applicability in the first three. Although the original tool uses Low/High/Unclear, the present table uses 'Some concerns' to denote moderate concerns or incomplete

reporting. [A1] indicates that blinding was not explicitly reported. [A3] indicates methodological inference based on index test refinement within the same sample (Whiting et al., 2011).

QUADAS-2 assessment indicated that both Tier 1 (CORE) studies provided data suitable for inclusion in the synthesis, but exhibited distinct risk of bias profiles. Wickersham et al. (2015) raised *some concerns* regarding patient selection and applicability, given the highly specific sample of newly incarcerated, HIV-positive individuals nearing release, and the Index Test domain was judged to be at *high risk of bias* due to apparent refinement of the final RODS instrument within the same validation sample. Evren et al. (2014) similarly presented *some concerns* in patient selection, but the primary source of bias was again the Index Test domain, as the DUDIT threshold was selected post hoc based on ROC analysis within the same cohort. In both studies, reference standards were clinically appropriate, although *some concerns* remained regarding the reporting of blinding procedures between index tests and reference standards.

Accordingly, pooled sensitivity and specificity estimates should be interpreted as preliminary and potentially optimistic, particularly in the context of a very small evidence base ($k = 2$). Heterogeneity statistics (τ^2 and I^2) should be reported explicitly, but their interpretation should remain cautious, as hierarchical models in diagnostic test accuracy meta-analysis may yield unstable or overprecise estimates when based on a limited number of studies (Reitsma et al., 2005; Takwoingi et al., 2017; von Hippel, 2015; Aarons, Hurlburt, & Horwitz, 2011).

3.2.4. Pooled Diagnostic Accuracy Estimates (Tier 1)

Pooled diagnostic accuracy estimates were derived from the two Tier 1 (CORE) studies using a bivariate random-effects model (Reitsma et al., 2005; Riley et al., 2020), which jointly models sensitivity and specificity while accounting for their correlation across studies. Although the two Tier 1 studies differed in target substance, instrument, reference standard, and setting, pooling was retained to provide an exploratory summary of diagnostic performance under conditions of explicit reference standard use. However, this synthesis should be interpreted cautiously, as conceptual heterogeneity may limit the comparability of pooled estimates. Pooled sensitivity was 0.96 (95% CI 0.91–0.98), while pooled specificity was 0.78 (95% CI 0.70–0.84).

Between-study heterogeneity was assessed using τ^2 and I^2 statistics, estimated separately for sensitivity and specificity on the logit scale. Heterogeneity was negligible for both outcomes ($\tau^2 = 0.00$; $I^2 = 0.0\%$ for sensitivity and specificity). However, given the very small number of included studies ($k = 2$), these estimates should be interpreted with caution. With such a limited evidence base, between-study variance cannot be reliably estimated, and I^2 is known to be unstable and potentially biased in small meta-analyses (von Hippel, 2015). Accordingly, these heterogeneity statistics should be regarded as descriptive indicators rather than evidence of true homogeneity across studies.

However, given the very small number of included studies (k = 2), these estimates should be interpreted with caution. In such cases, the estimation of between-study variance (τ^2) is unstable, and hierarchical models may yield confidence intervals that appear overly precise and potentially underestimate true uncertainty (Takwoingi et al., 2017; von Hippel, 2015). *Accordingly, these pooled estimates should be considered exploratory and descriptive, intended to illustrate diagnostic performance under strict reference standard conditions rather than to provide stable or generalisable population parameters.*

3.3. Tier 2 (Extended Forensic Validation)

Tier 2 comprised six studies evaluating brief substance use screening instruments in adult forensic or correctional populations using clinically informed but context-specific reference standards (Berman et al., 2004; Durbeej et al., 2010; Ober et al., 2013; Peters et al., 2000; Proctor & Hoffmann, 2016; Wolff & Shi, 2015) (Table 5). Although the IRIS was designed as a dual-domain screener, the present synthesis considered only its substance use screening component, consistent with the study's validation aims (Ober et al., 2013).

These studies consistently demonstrated high sensitivity for identifying substance use problems across diverse correctional contexts, including forensic psychiatric assessment, large-scale prison systems, culturally specific incarcerated populations, and detoxification settings. However, heterogeneity in reference standards, outcome operationalisation, and scoring thresholds precluded quantitative pooling. Accordingly, Tier 2 evidence was used to contextualise Tier 1 findings and inform conclusions regarding robustness, feasibility, and applied relevance rather than contributing to pooled diagnostic accuracy estimates.

Table 5. Tier 2 (Extended Forensic Validation) studies: reference standards and performance indicators.

Author (Year)	Screening Instrument	Reference Standard/Comparator	Main Performance Indicators Reported	Key Notes/Reason for Tier 2 Classification
Berman et al. (2004)	DUDIT	Structured diagnostic criteria (DSM-IV) in criminal justice and detoxification settings	Sensitivity high for drug dependence; acceptable specificity; strong internal consistency	Psychometric validation focus; not designed for formal diagnostic accuracy pooling
Durbeej et al. (2010)	AUDIT; DUDIT	Addiction Severity Index (ASI-6) and MSAC DSM-IV-based decision framework	AUDIT: AUC = 0.88; DUDIT: AUC = 0.93; balanced sensitivity/specificity at study-specific cut-offs	Composite and context-specific reference standard; threshold heterogeneity
Peters et al. (2000)	Multiple instruments (TCU Drug Screen; SSI; ADS/ASI composite)	Structured Clinical Interview for DSM-IV (SCID-IV)	Comparative sensitivity, specificity, and overall accuracy across instruments	Multi-instrument comparison; heterogeneous targets and outcomes
Proctor and Hoffmann (2016)	UNCOPE	Substance Use Disorder Diagnostic Schedule-IV (SUDDS-IV)	Sensitivity 85–98% for moderate/severe SUD; specificity 97–99%; high internal consistency	Severity-based classification; pragmatic diagnostic framework
Wolff and Shi (2015)	ASSIST (v3.0)	Structured Clinical Interview for DSM-IV, Non-Patient Version (SCID-NP)	Moderate-to-high sensitivity and specificity varying by substance and administration mode	Substance-specific outcomes; multidimensional scoring
Ober et al. (2013)	Indigenous Risk Impact Screen (IRIS)	Composite International Diagnostic Interview (CIDI; ICD-10)	Sensitivity 94% for substance use disorder; low specificity; acceptable concurrent validity	Culturally adapted instrument; context-specific thresholds

Note: Tier 2 studies evaluated screening instruments against clinically informed but context-specific reference standards or severity-based frameworks. Although these studies reported favorable performance indicators, heterogeneity in reference standards, outcome definitions, and scoring thresholds precluded inclusion in the Tier 1 meta-analysis.

3.4. Tier 3 (Predictive Validity)

Design characteristics and outcome measures of the Tier 3 predictive validity study are summarised in Table 6.

Table 6. Tier 3 (Predictive Validity) study: screening severity and post-release outcomes

Author (Year)	Screening Instrument	Timing of Screening	Follow-Up Duration	Outcome Definition	Statistical Methods	Main Findings
Knight et al. (1999)	Salient Factor Score (SFS)	Intake (pre-treatment)	3 years (annual analyses)	Reincarceration (new offenses and conditional revocations); secondary analyses restricted to new offenses	ANOVA; χ^2 tests; planned comparisons by severity and treatment group	Higher baseline severity predicted higher reincarceration rates; completion of in-prison treatment plus aftercare substantially reduced reincarceration, particularly among high-severity individuals; lowest rates observed for new offenses among treatment completers

Note: Tier 3 includes studies examining the longitudinal association between screening-derived severity or risk classifications obtained during incarceration and substance-relevant post-release outcomes. The Salient Factor Score (SFS) is a composite screening instrument incorporating substance-related and criminogenic indicators and was evaluated for predictive, rather than diagnostic, validity.

Tier 3 comprised studies examining the longitudinal association between screening-derived severity or risk classifications obtained during incarceration and subsequent substance-relevant post-release outcomes. In contrast to diagnostic accuracy studies, which evaluate concordance with contemporaneous reference standards, predictive validity studies assess whether screening scores are meaningfully associated with future outcomes over time (Cronbach & Meehl, 1955; Messick, 1995). Within the present review, only one study met the predefined criteria for inclusion in this tier, namely, the use of a screening-derived severity measure assessed during incarceration and a substance-relevant outcome measured prospectively following release.

Knight et al. (1999) evaluated the predictive utility of the Salient Factor Score (SFS) in relation to reincarceration outcomes over a three-year follow-up period. The SFS is a composite instrument incorporating both substance-related and criminogenic indicators and does not map directly onto the brief substance use screening constructs evaluated in Tiers 1 and 2. The SFS is a composite screening instrument originally developed by the United States Parole Commission to assess offense- and substance-related severity and to estimate relative risk of recidivism following release. Although not designed as a diagnostic instrument for substance use disorders, the SFS incorporates multiple indicators that have been consistently linked to post-release substance use and criminal justice outcomes, including drug dependence history, employment instability, and prior criminal involvement (Andrews & Bonta, 2010; Taxman et al., 2007a).

Screening was conducted at intake, prior to participation in institutional treatment programs. Reincarceration outcomes were assessed using administrative records across three consecutive post-release years, with separate analyses conducted for each year. Reincarceration encompassed both new criminal convictions and conditional revocations due to violations of release conditions, an outcome definition commonly used in correctional outcome research and strongly associated with substance use relapse following release (Binswanger et al., 2013; Kinner & Wang, 2014). Additional analyses restricted outcomes to reincarceration for new offenses only, thereby reducing potential confounding by technical violations.

The results of analysis of variance and chi-square tests showed that higher baseline severity, as indexed by the SFS, was associated with an increased likelihood of reincarceration across the follow-up period. Individuals classified as higher severity exhibited substantially higher reincarceration rates than those classified as lower severity, consistent with prior evidence linking substance-related severity and criminogenic risk to post-release failure (Gendreau, Little, & Goggin, 1996; Knight et al., 1999).

Importantly, treatment participation moderated this association. Participants who completed both intensive in-prison treatment and structured community aftercare demonstrated markedly lower reincarceration rates than those who discontinued aftercare or received no comparable intervention. This moderating effect was most pronounced among individuals classified as higher severity, for whom completion of both treatment components was associated with substantially reduced reincarceration rates. This pattern aligns with the broader literature indicating that individuals with higher baseline substance use severity derive disproportionate benefit from intensive, continuous treatment interventions spanning incarceration and community reentry (Simpson et al., 2012; Taxman & Caudy, 2015).

When analyses were restricted to reincarceration for new offenses only, the lowest rates were observed among individuals who completed both in-prison treatment and aftercare. This finding further supports the prognostic relevance of screening-derived severity classifications and is consistent with evidence that substance use relapse is a key driver of new criminal offending following release (Merrall et al., 2010; Zlodre & Fazel, 2012).

Although the SFS is not a diagnostic screening instrument in the strict sense and incorporates both criminogenic and substance-related indicators, this study provides evidence that screening-derived severity classifications obtained during incarceration can meaningfully predict post-release substance-relevant outcomes over extended follow-up periods. In accordance with

contemporary validity theory, this evidence speaks to predictive validity rather than diagnostic accuracy and must therefore be interpreted as addressing a distinct inferential target (Cronbach & Meehl, 1955; Messick, 1995). On this basis, Knight et al. (1999) was classified within Tier 3, contributing longitudinal outcome evidence that complements—but does not substitute for—the diagnostic accuracy and extended validation findings synthesised in Tiers 1 and 2.

Accordingly, Tier 3 findings should be regarded as illustrative rather than confirmatory, highlighting potential predictive pathways that warrant further investigation rather than supporting generalised prognostic claims.

Taken together, the results indicate that brief substance use screening instruments demonstrate high diagnostic sensitivity when evaluated against explicit clinical reference standards in adult correctional populations (Tier 1), show broadly consistent performance across diverse forensic contexts when assessed using extended validation frameworks (Tier 2), and exhibit meaningful prognostic relevance for substance-related post-release outcomes when severity-based classifications are examined longitudinally (Tier 3).

4. Discussion

This stratified meta-analysis was designed to clarify what brief substance use screening instruments can validly support in adult correctional and forensic populations. By separating diagnostically anchored studies from broader validation studies and from longitudinal outcome studies, this review addresses a recurring problem in the literature: evidence for different uses of screening is often discussed as if it supported a single conclusion. The present findings suggest a more limited and more defensible interpretation. In correctional settings, brief screening appears most useful for early case identification when evaluated against explicit reference standards, for operational appraisal of need when used within context-specific service systems, and for limited prognostic planning when linked to later outcomes. These functions are complementary, but they do not constitute a unitary form of validity and should not be interpreted as interchangeable (Cronbach & Meehl, 1955; Messick, 1995).

4.1. Tier 1 (CORE: Diagnostic Accuracy) in Adult Correctional Populations

The clearest conclusion from the present review concerns initial case detection. The two Tier 1 studies, both of which used explicit diagnostic or quasi-diagnostic reference standards, showed very high sensitivity and moderate specificity for brief substance use screening in adult correctional populations (Evren et al., 2014; Wickersham et al., 2015). This pattern appears to be consistent with diagnostic test accuracy principles: first-line screens are often designed to minimise false negatives, especially in settings where the prevalence of substance use disorders is high and missed need carries clinical and custodial consequences (Bossuyt et al., 2015; Leeflang et al., 2008; Fazel et al., 2017; Kinner & Wang, 2014).

Moderate specificity in Tier 1 suggests that some non-cases will screen positive, which reinforces the proper role of these instruments as triage tools rather than stand-alone diagnostic procedures. A positive screen should therefore trigger fuller assessment, not definitive labelling. At the same time, the small number of Tier 1-eligible studies remains a major limitation. Despite the routine use of screening in correctional systems, very few studies have evaluated brief instruments against independent reference standards while reporting data suitable for formal diagnostic accuracy synthesis. This gap matters because the pooled Tier 1 estimates are necessarily provisional, particularly in light of the risk of bias identified in the Index Test domain and the known instability of hierarchical models when only a few studies are available (Whiting et al., 2011; Riley et al., 2020; Takwoingi et al., 2017; von Hippel, 2015).

4.2. Tier 2 (Extended Forensic Validation): Robustness and Contextual Generalisability

Tier 2 broadens the picture by showing that brief screening instruments can perform usefully across varied correctional and forensic contexts even when strict diagnostic reference standards are

not used. Across prison, forensic psychiatric, detoxification-linked, and culturally specific incarcerated samples, studies of the AUDIT, DUDIT, UNCOPE, ASSIST, and IRIS generally reported favorable sensitivity, acceptable reliability, or clinically meaningful concurrent associations with comparator measures and severity frameworks (Berman et al., 2004; Durbeej et al., 2010; Ober et al., 2013; Proctor & Hoffmann, 2016; Wolff & Shi, 2015). On that basis, Tier 2 supports the operational usefulness of brief screeners in the environments where correctional assessment is typically conducted.

Importantly, Tier 2 should not be interpreted as a weaker extension of Tier 1 diagnostic evidence. Rather, it represents a distinct evidentiary domain, reflecting how brief screening instruments function under applied correctional conditions where reference standards, thresholds, and intended uses vary substantially. Many of these studies used context-specific thresholds, composite standards, clinical judgment, or severity-based comparators. Such approaches may be entirely reasonable for service allocation, but they introduce threshold and construct heterogeneity that limits diagnostic interpretation and makes direct pooling inappropriate (Bossuyt et al., 2015; Leeflang et al., 2008; Macaskill et al., 2010). The Tier 2 evidence also highlights the importance of institutional and cultural fit, especially where screening tools are applied across populations that differ in language, comorbidity, social context, and help-seeking norms (Ober et al., 2013)

4.3. Tier 3 (Predictive Validity): Screening Severity and Post-Release Outcomes

Tier 3 addressed a different question: whether screening-derived severity or risk classifications obtained during incarceration carry meaningful prognostic information for later outcomes. The single eligible study suggested that they may. Knight et al. (1999) found that greater baseline severity, indexed by the Salient Factor Score, was associated with higher reincarceration risk over three years and that this pattern was attenuated among individuals who completed structured treatment and aftercare. This finding is clinically relevant because post-release substance involvement is closely tied to overdose, mortality, relapse, treatment discontinuity, and return to custody (Binswanger et al., 2013; Merrall et al., 2010; Zlodre & Fazel, 2012).

The practical implication of Tier 3 is not that screening instruments become diagnostic by virtue of predicting later outcomes. Rather, it is that severity-based classifications may contribute to treatment stratification, discharge planning, and aftercare intensity when interpreted as prognostic rather than diagnostic evidence. This boundary is essential because reincarceration and related outcomes are shaped not only by substance-related need, but also by supervision conditions, treatment access, housing instability, and institutional response. With only one study in this tier, these findings should therefore be regarded as suggestive rather than confirmatory (Andrews & Bonta, 2010; Simpson et al., 2012; Taxman & Caudy, 2015; Cronbach & Meehl, 1955; Messick, 1995). Thus, the Tier 3 finding should be interpreted as evidence that screening-derived severity may function as a proxy for a broader risk ecology, rather than as an independent or substance-specific predictor of post-release outcomes.

4.4. Integrative Implications Across Tiers

Taken together, the three tiers support a function-specific interpretation of brief substance use screening in correctional and forensic settings. Tier 1 supports use for initial case identification under diagnostically anchored conditions; Tier 2 supports use for operational screening, severity appraisal, and referral prioritisation in real-world systems; and Tier 3 suggests a limited role in prognostic planning when screening-derived severity is linked to longitudinal outcomes. However, this functional differentiation should not be interpreted as equally supported across tiers. The evidentiary base is asymmetric: Tier 1 includes only two diagnostically anchored studies, Tier 2 is methodologically heterogeneous, and Tier 3 is represented by a single longitudinal study. The value of the stratified approach is that it preserves these distinctions instead of collapsing diagnostically anchored, context-specific, and prognostic evidence into one global claim that a screening instrument is simply “valid” (Cronbach & Meehl, 1955; Messick, 1995; Bossuyt et al., 2015;

Leeflang et al., 2008; Riley et al., 2020). Beyond summarising existing findings, the present study contributes by demonstrating that inconsistencies in the correctional screening literature are not solely due to variability in study quality but also reflect a deeper conceptual conflation of distinct evidentiary domains. By explicitly separating diagnostically anchored evidence, extended validation, and predictive outcomes, the stratified approach clarifies that many apparent contradictions in prior research arise from comparing findings that address different inferential questions. This distinction has implications not only for interpretation but also for how future studies should be designed and synthesised.

4.4.1. Theoretical Implications

Theoretically, the present findings reinforce a core validity principle: validity attaches to interpretations and uses of scores, not to instruments in the abstract (Cronbach & Meehl, 1955; Messick, 1995). In correctional research, brief screeners are often described globally as "valid" even when the supporting evidence pertains only to one domain. The tiered framework counters that shortcut by showing that diagnostic accuracy, extended forensic validation, and predictive validity represent distinct inferential claims.

4.4.2. Methodological Implications

Methodologically, this review highlights the limited size and uneven quality of the diagnostically anchored evidence base. Only two studies met minimum criteria for formal diagnostic accuracy synthesis, and both raised concerns in the Index Test domain. Future primary studies would benefit from independent reference standards, prespecified thresholds, blinded assessment, complete two-by-two reporting, and reporting consistent with STARD and QUADAS-2 guidance (Bossuyt et al., 2015; Whiting et al., 2011; Takwoingi et al., 2017). The stratified synthesis used here also offers a pragmatic alternative to either excluding non-CORE studies altogether or pooling conceptually incompatible designs (Leeflang et al., 2008; Page et al., 2021).

4.4.3. Practical Implications for Correctional Screening

Practically, the findings support the use of brief screening instruments within staged assessment pathways. At intake, where time is limited and missed need carries substantial cost, high-sensitivity screens can support triage and prioritisation for further review. In settings where comprehensive interviews are not feasible for everyone, Tier 2 evidence also supports the use of brief tools for severity grading and referral organisation, provided that locally calibrated practices are not misrepresented as equivalent to diagnostically validated classification (Hartzler et al., 2014; Proctor & Hoffmann, 2016; Wolff & Shi, 2015).

4.4.4. Clinical Implications

Clinically, screening should be linked to follow-up assessment and treatment planning rather than used in isolation for diagnostic labelling or treatment allocation. Even the CORE studies showed only moderate specificity, so screening scores are best treated as indicators of possible need that should be followed by confirmatory assessment, formulation, and treatment matching (Evren et al., 2014; Wickersham et al., 2015). Tier 3 further suggests that severity-based classifications may help inform treatment intensity and continuity of care, especially during community reentry (Andrews & Bonta, 2010; Simpson et al., 2012; Taxman & Caudy, 2015).

4.4.5. Policy Implications

At the policy level, the findings argue against frameworks that equate screening completion with diagnosis. Correctional policies should distinguish between screening, diagnosis, and prognostic stratification, and should align each function with appropriate evidentiary standards. This review also supports continued investment in diagnostically rigorous correctional research, especially for substances other than opioids and for populations with high psychiatric comorbidity,

while also supporting policies that link screening to treatment continuity and aftercare during the high-risk post-release period (Taxman et al., 2007a; Fazel et al., 2017; Binswanger et al., 2013; Merrall et al., 2010).

4.5. Forensic Use of Substance Use Screening Instruments

The present findings support a constrained forensic role for brief substance use screening instruments. In legal and medico-legal contexts, these tools may appropriately be used as adjunctive aids to structured clinical judgment: to support early case identification, guide interview focus and collateral data collection, and inform treatment need and continuity planning. That use is consistent with the general logic of screening as a first-stage procedure and with the DSM-5-TR caution that diagnostic criteria are intended for use by appropriately trained clinicians rather than as stand-alone substitutes for full forensic formulation (American Psychiatric Association, 2022; First et al., 2016; Melton et al., 2018; Rogers & Shuman, 2005).

The evidence does not support broader forensic uses. On their own, screening instruments do not meet the evidentiary standards required for forensic conclusions about diagnosis, legal responsibility, competence, causation, or risk. Screening scores should not be treated as dispositive proof of a substance use disorder, nor should they be used on their own to infer criminal responsibility, competence, causation, or legal risk. In adversarial settings, additional caution is warranted because many brief instruments rely heavily on self-report, which is sensitive to context, perceived consequences, and reporting mode (Darke, 1998; Tourangeau, 2018). More broadly, forensic research has emphasized that behavioral and communicative indicators are highly context-dependent and should not be interpreted in isolation from broader clinical and situational information (Douglas & Skeem, 2005; Rusu & Delcea, 2024; Calderaro, Mastronardi, & Șerban, 2025; Monahan & Skeem, 2016). The most defensible expert use is therefore inferentially constrained: screening results may inform hypotheses and planning, but they should remain embedded within a multimethod forensic evaluation rather than function as stand-alone evidentiary conclusions.

4.6. Limitations and Future Directions

The principal limitation of the present synthesis is the restricted size of the Tier 1 evidence base. More broadly, the literature appears structurally misaligned with diagnostic accuracy methodology: instruments are widely implemented in correctional practice, but primary studies rarely use independent reference standards, prespecified thresholds, blinded procedures, and complete two-by-two reporting (Damschroder, et al., 2009). Only two studies met strict criteria for diagnostic accuracy synthesis, and broader correctional screening research remains heterogeneous in reference standards, thresholds, and outcome definitions. Several instruments also blur substance-specific indicators with broader criminogenic or psychosocial variables, complicating interpretation when screening scores are treated as proxies for diagnostic status. These problems are compounded by the predominance of self-report and by the practical constraints of conducting gold-standard diagnostic assessment in applied correctional settings (Darke, 1998; Tourangeau & Yan, 2007). A further limitation is that predictive evidence was restricted to a single study using reincarceration, a distal and institutionally mediated outcome, rather than repeated direct measures of post-release substance use (Austin & Hardyman, 2004; Kaeble & Glaze, 2016; Harrison, 1995).

Future research should prioritise diagnostically rigorous studies using standardised independent reference standards, prespecified cut-offs, blinded procedures, and full two-by-two reporting across a wider range of substances and jurisdictions. It should also distinguish more clearly among diagnosis, severity appraisal, and prognostic stratification, and should examine how implementation factors—such as staff training, workflow integration, and linkage to treatment and aftercare—shape the practical value of screening in real correctional systems. Where feasible, more transparent reporting consistent with STARD, along with more advanced synthesis strategies such as IPD meta-analysis, would improve comparability and allow better modeling of threshold effects

and subgroup differences (Bossuyt et al., 2015; Cohen et al., 2016; Trikalinos et al., 2019; Riley et al., 2020; Steyerberg et al., 2013).

5. Conclusions

This stratified meta-analysis aims to clarify that brief substance use screening instruments in adult correctional and forensic populations support different kinds of claims, not a single global claim of validity. Under diagnostically anchored conditions, they appear most useful for early case identification; in broader operational settings, they can support severity appraisal and referral prioritisation; and in limited longitudinal evidence, screening-derived severity may contribute to prognostic planning. The central implication is that correctional screening practice should be aligned with explicitly stated inferential goals and that evidence supporting one use should not be assumed to justify another. Recent discussions in forensic neuroscience and AI-assisted behavioural evaluation similarly stress the importance of preserving inferential specificity and avoiding overgeneralization when interpreting complex behavioural indicators in legal and correctional settings (Șerban, 2025).

At present, the main challenge is not the absence of screening tools but the frequent mismatch between how these tools are used and the type of evidence available to justify those uses. Future work should expand the diagnostically rigorous evidence base, improve reporting quality, and preserve construct clarity when synthesising correctional screening research (Evren et al., 2014; Wickersham et al., 2015; Knight et al., 1999; Bossuyt et al., 2015; Leeflang et al., 2008; Riley et al., 2020).

Author Contributions: Conceptualisation, D.O.R. and C.D.; methodology, D.O.R., C.D. and I.V.S.; software, D.O.R. and I.V.S.; validation, D.O.R., C.D., and I.V.S.; formal analysis, D.O.R., C.D., and I.V.S.; investigation, D.O.R.; resources, D.O.R. and C.D.; data curation, D.O.R.; writing—original draft preparation, D.O.R; writing—review and editing, D.O.R., C.D., and I.V.S.; visualisation, D.O.R.; supervision, D.O.R., C.D., and I.V.S.; project administration, D.O.R. and C.D.; funding acquisition, D.O.R. and C.D.. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.
Data Availability Statement:
Conflicts of Interest: The authors declare no conflicts of interest.

References

Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(1), 4–23. https://doi.org/10.1007/s10488-010-0327-7

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.; DSM-5-TR). American Psychiatric Association Publishing. https://doi.org/10.1176/appi.books.9780890425787

Andrews, D. A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). Matthew Bender.

Austin, J. and Hardyman, P.L. (2004), The Risks and Needs of the Returning Prisoner Population. Review of Policy Research, 21: 13-29. https://doi.org/10.1111/j.1541-1338.2004.00055.x

Baillargeon, J., Penn, J. V., Thomas, C. R., Temple, J. R., Baillargeon, G., & Murray, O. J. (2009). Psychiatric disorders and suicide in the nation's largest state prison system. *Journal of the American Academy of Psychiatry and the Law, 37*(2), 188–193.

Baillargeon, J., Penn, J. V., Knight, K., Harzke, A. J., Baillargeon, G., & Becker, E. A. (2010). Risk of reincarceration among prisoners with co-occurring severe mental illness and substance use

disorders. *Administration and Policy in Mental Health and Mental Health Services Research, 37*(4), 367–374. https://doi.org/10.1007/s10488-009-0252-9

Babor, T. F., Higgins-Biddle, J. C., Saunders, J. B., & Monteiro, M. G. (2001). *AUDIT: The Alcohol Use Disorders Identification Test: Guidelines for use in primary care* (2nd ed.). World Health Organization.

Berman, A. H., Bergman, H., Palmstierna, T., & Schlyter, F. (2004). Evaluation of the Drug Use Disorders Identification Test (DUDIT) in criminal justice and detoxification settings and in a Swedish population sample. *European Addiction Research, 11*(1), 22–31. https://doi.org/10.1159/000081413

Binswanger, I. A., Blatchford, P. J., Mueller, S. R., & Stern, M. F. (2013). Mortality after prison release: Opioid overdose and other causes of death, risk factors, and time trends from 1999 to 2009. *Annals of Internal Medicine, 159*(9), 592–600. https://doi.org/10.7326/0003-4819-159-9-201311050-00005

Blendea, L., Gotca, I., Vata, I., Novac, B., Novac, O., Mihailescu, A., & Maftei, C. (2025). Opium and opioid toxicity: AI-enhanced insights into pharmacology, clinical manifestations, and emergency management. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 16*(3), 438–448. http://dx.doi.org/10.70594/brain/16.3/33

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., Kressel, H. Y., Rifai, N., Golub, R. M., Altman, D. G., Hooft, L., Korevaar, D. A., & Cohen, J. F. (2015). STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *Radiology, 277*(3), 826–832. https://doi.org/10.1148/radiol.2015151516

Butler, T., Indig, D., Allnutt, S., & Mamoon, H. (2011). Co-occurring mental illness and substance use disorder among Australian prisoners. *Drug and Alcohol Review, 30*(2), 188–194. https://doi.org/10.1111/j.1465-3362.2010.00216.x

Calderaro, M., Mastronardi, V., & Serban, I. (2025). Addictions not related to the use and abuse of substances and some assessment tools in the clinical-forensic field. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 16*(2), 224–250. http://dx.doi.org/10.70594/brain/16.2/17

Calderaro, M., Mastronardi, V., & Serban, I. (2025). Coordinates of Nonverbal Expressions in Educational Settings for the Structuring of Artificial Intelligence Programs: Simulation and Lying. *BRAIN. Broad Research In Artificial Intelligence And Neuroscience, 16*(1 Sup1), 383-398. doi:http://dx.doi.org/10.70594/brain/16.S1/30

Cohen, J. F., Korevaar, D. A., Altman, D. G., Bruns, D. E., Gatsonis, C. A., Hooft, L., Irwig, L., Levine, D., Reitsma, J. B., de Vet, H. C. W., & Bossuyt, P. M. M. (2016). *STARD 2015 guidelines for reporting diagnostic accuracy studies: Explanation and elaboration*. *BMJ Open, 6*(11), e012799. https://doi.org/10.1136/bmjopen-2016-012799

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science, 4*, Article 50. https://doi.org/10.1186/1748-5908-4-50

Deeks, J. J., Macaskill, P., & Irwig, L. (2005). 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*, *58*(9), 882–893. https://doi.org/10.1016/j.jclinepi.2005.01.016

Darke, S. (1998). *Self-report among injecting drug users: A review*. *Drug and Alcohol Dependence, 51*(3), 253–263. https://doi.org/10.1016/S0376-8716(98)00028-3

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, and Law, 11*(3), 347–383. https://doi.org/10.1037/1076-8971.11.3.347

Durbeej, N., Berman, A. H., Gumpert, C. H., Palmstierna, T., Kristiansson, M., & Alm, C. (2010). Validation of the Alcohol Use Disorders Identification Test and the Drug Use Disorders Identification Test in a Swedish sample of suspected offenders with signs of mental health problems: results from the Mental Disorder, Substance Abuse and Crime study. *Journal of substance abuse treatment*, *39*(4), 364–377. https://doi.org/10.1016/j.jsat.2010.07.007

Evren, C., Ögel, K., Evren, B., & Bozkurt, M. (2014). Psychometric properties of the Turkish versions of the Drug Use Disorders Identification Test (DUDIT) and the Drug Abuse Screening Test (DAST-10) in the prison setting. *Journal of Psychoactive Drugs, 46*(2), 140–146. https://doi.org/10.1080/02791072.2014.887162

Fazel, S., Yoon, I. A., & Hayes, A. J. (2017). Substance use disorders in prisoners: An updated systematic review and meta-regression analysis in recently incarcerated men and women. *Addiction, 112*(10), 1725–1739. https://doi.org/10.1111/add.13877

First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. L. (2016). *Structured Clinical Interview for DSM-5® disorders—Clinician version (SCID-5-CV)*. American Psychiatric Association Publishing.

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*, *34*(4), 575–608. https://doi.org/10.1111/j.1745-9125.1996.tb01220.x

Harrison, L. D. (1995). The Validity of Self-Reported Data on Drug Use. *Journal of Drug Issues*, *25*(1), 91-111.

Hartzler, B., Jackson, T. R., Jones, B. E., Beadnell, B., & Calsyn, D. A. (2014). Disseminating contingency management: impacts of staff training and implementation at an opiate treatment program. *Journal of substance abuse treatment*, *46*(4), 429–438. https://doi.org/10.1016/j.jsat.2013.12.007

Humeniuk, R., Ali, R., Babor, T. F., Farrell, M., Formigoni, M. L., Jittiwutikarn, J., de Lacerda, R. B., Ling, W., Marsden, J., Monteiro, M., Nhiwatiwa, S., Pal, H., Poznyak, V., & Simon, S. (2008). Validation of the Alcohol, Smoking and Substance Involvement Screening Test (ASSIST). *Addiction, 103*(6), 1039–1047. https://doi.org/10.1111/j.1360-0443.2007.02114.x

Kaeble, D., & Glaze, L. E. (2016). *Correctional populations in the United States, 2015* (NCJ 250374). U.S. Department of Justice, Bureau of Justice Statistics.

Kinner, S. A., & Wang, E. A. (2014). The case for improving the health of ex-prisoners. *American Journal of Public Health, 104*(8), 1352–1355. https://doi.org/10.2105/AJPH.2014.301883

Knight, K., Simpson, D. D., & Hiller, M. L. (1999). Three-year reincarceration outcomes for in-prison therapeutic community treatment in Texas. *The Prison Journal, 79*(3), 337–351. https://doi.org/10.1177/0032885599079003004

Leeflang, M. M. G., Deeks, J. J., Gatsonis, C., Bossuyt, P. M. M., & the Cochrane Diagnostic Test Accuracy Working Group. (2008). Systematic reviews of diagnostic test accuracy. *Annals of Internal Medicine, 149*(12), 889–897. https://doi.org/10.7326/0003-4819-149-12-200812160-00008

Macaskill, P., Gatsonis, C., Deeks, J. J., Harbord, R. M., & Takwoingi, Y. (2010). Analysing and presenting results. In J. J. Deeks, P. M. Bossuyt, & C. Gatsonis (Eds.), *Cochrane handbook for systematic reviews of diagnostic test accuracy*. Cochrane Collaboration. http://srdta.cochrane.org/handbook-dta-reviews

Merrall, E. L. C., Kariminia, A., Binswanger, I. A., Hobbs, M. S. T., Farrell, M., Marsden, J., Hutchinson, S. J., & Bird, S. M. (2010). Meta-analysis of drug-related deaths soon after release from prison. *Addiction, 105*(9), 1545–1554. https://doi.org/10.1111/j.1360-0443.2010.02990.x

Melton, G. B., Petrila, J., Poythress, N. G., Slobogin, C., Otto, R. K., Mossman, D., & Condie, L. O. (2018). *Psychological evaluations for the courts: A handbook for mental health professionals and lawyers* (4th ed.). Guilford Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology, 12*, 489–513. https://doi.org/10.1146/annurev-clinpsy-021815-092945

Ober, C., Dingle, K., Clavarino, A., Najman, J. M., Alati, R., & Heffernan, E. B. (2013). Validating a screening tool for mental health and substance use risk in an Indigenous prison population. *Drug and Alcohol Review, 32*(6), 611–617. https://doi.org/10.1111/dar.12063

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Peters, R. H., Greenbaum, P. E., Steinberg, M. L., Carter, C. R., Ortiz, M. M., Fry, B. C., & Valle, S. K. (2000). Effectiveness of screening instruments in detecting substance use disorders among prisoners. *Journal of Substance Abuse Treatment, 18*(4), 349–358. https://doi.org/10.1016/S0740-5472(99)00081-1

Prins, S. J. (2014). Prevalence of mental illnesses in U.S. state prisons: A systematic review. *Psychiatric Services, 65*(7), 862–872. https://doi.org/10.1176/appi.ps.201300166

Proctor, S. L., & Hoffmann, N. G. (2016). The UNCOPE: An effective brief screen for DSM-5 substance use disorders in correctional settings. *Psychology of Addictive Behaviors, 30*(5), 613–618. https://doi.org/10.1037/adb0000170

Reitsma, J. B., Glas, A. S., Rutjes, A. W. S., Scholten, R. J. P. M., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology, 58*(10), 982–990. https://doi.org/10.1016/j.jclinepi.2005.02.022

Riley, R. D., Debray, T. P. A., Fisher, D., Hattle, M., Marlin, N., Hoogland, J., Gueyffier, F., Staessen, J. A., Wang, J., Moons, K. G. M., Reitsma, J. B., & Ensor, J. (2020). *Individual participant data meta-analysis to examine interactions between treatment effect and participant-level covariates: Statistical recommendations for conduct and planning*. *Statistics in Medicine, 39*(15), 2115–2137. https://doi.org/10.1002/sim.8516

Rogers, R., & Shuman, D. W. (2005). *Fundamentals of forensic practice: Mental health and criminal law*. Springer. https://doi.org/10.1007/b106925

Rusu, D., & Delcea, C. (2024). Mental Health and Psychosocial Adjustment in Ukrainian Refugee Children in Romania. *BRAIN. Broad Research In Artificial Intelligence And Neuroscience, 15*(4), 197-214. doi:http://dx.doi.org/10.70594/brain/15.4/14

Rusu Andron, R., Nicoară, R., & Coman, H. (2025). Systematic review: Coping strategies among individuals with drug addiction. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 16*(1), 115–124. http://dx.doi.org/10.70594/brain/16.1/8

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of clinical psychiatry*, *59 Suppl 20*, 22–57.

Simpson, D. D., Joe, G. W., Knight, K., Rowan-Szal, G. A., & Gray, J. S. (2012). Texas Christian University (TCU) Short Forms for Assessing Client Needs and Functioning in Addiction Treatment. *Journal of offender rehabilitation*, *51*(1-2), 34–56. https://doi.org/10.1080/10509674.2012.633024

Steinberg, L., & Morris, A. S. (2001). Adolescent development. Annual Review of Psychology, 52, 83-110. http://dx.doi.org/10.1146/annurev.psych.52.1.83

Steyerberg, E. W., Moons, K. G. M., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., & Altman, D. G. (2013). *Prognosis Research Strategy*

(PROGRESS) 3: Prognostic model research. *PLoS Medicine, 10*(2), e1001381. https://doi.org/10.1371/journal.pmed.1001381

Șerban, I. V. (2025). Neuroscience, Genetics, Education, and AI: Charting New Frontiers in Understanding Human Behaviour and Criminal Responsibility. *BRAIN. Broad Research In Artificial Intelligence And Neuroscience, 16*(1 Sup1), 399-414. doi:http://dx.doi.org/10.70594/brain/16.S1/31

Taxman, F. S., Young, D. W., Wiersema, B., Rhodes, A., & Mitchell, S. (2007a). The National Criminal Justice Treatment Practices survey: Multilevel survey methods and procedures. *Journal of Substance Abuse Treatment, 32*(3), 225–238. https://doi.org/10.1016/j.jsat.2007.01.002

Taxman, F. S., Perdoni, M. L., & Harrison, L. D. (2007b). Drug treatment services for adult offenders: The state of the state. *Journal of Substance Abuse Treatment, 32*(3), 239–254. https://doi.org/10.1016/j.jsat.2006.12.019

Taxman, F. S., & Caudy, M. S. (2015). Risk tells us who, but not what or how: Empirical assessment of the complexity of criminogenic needs to inform correctional programming. *Criminology & Public Policy, 14*(1), 71–103. https://doi.org/10.1111/1745-9133.12116

Takwoingi, Y., Guo, B., Riley, R. D., & Deeks, J. J. (2017). Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Statistical Methods in Medical Research, 26*(4), 1896–1911. https://doi.org/10.1177/0962280215592269

Tourangeau R (2018), "The survey response process from a cognitive viewpoint". *Quality Assurance in Education*, Vol. 26 No. 2 pp. 169–181, doi: https://doi.org/10.1108/QAE-06-2017-0034

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Trikalinos, T. A., Balion, C. M., Coleman, C. I., Griffith, L., Santaguida, P. L., & Vandermeer, B. (2019). *Meta-analysis of test performance when there is a "gold standard"*. *Journal of General Internal Medicine, 27*(Suppl. 1), S56–S66. https://doi.org/10.1007/s11606-012-2029-1

von Hippel P. T. (2015). The heterogeneity statistic I(2) can be biased in small meta-analyses. *BMC medical research methodology*, *15*, 35. https://doi.org/10.1186/s12874-015-0024-z

Wickersham, J. A., Azar, M. M., Cannon, C. M., Altice, F. L., & Springer, S. A. (2015). Validation of a brief measure of opioid dependence: The Rapid Opioid Dependence Screen (RODS). *Journal of Correctional Health Care, 21*(1), 12–26. https://doi.org/10.1177/1078345814557513

Wolff, N., & Shi, J. (2015). Screening for Substance Use Disorder Among Incarcerated Men with the Alcohol, Smoking, Substance Involvement Screening Test (ASSIST): A Comparative Analysis of Computer-Administered and Interviewer-Administered Modalities. *Journal of substance abuse treatment*, *53*, 22–32. https://doi.org/10.1016/j.jsat.2015.01.006

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., & Bossuyt, P. M. M. (2011). *QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies*. *Annals of Internal Medicine, 155*(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Zlodre, J., & Fazel, S. (2012). All-cause and external mortality in released prisoners: Systematic review and meta-analysis. *American Journal of Public Health, 102*(12), e67–e75. https://doi.org/10.2105/AJPH.2012.300764