HHS Public Access

Author manuscript

Alcohol Clin Exp Res. Author manuscript; available in PMC 2023 March 01.

Published in final edited form as:

Alcohol Clin Exp Res. 2022 March ; 46(3): 458–467. doi:10.1111/acer.14778.

Practical assessment of DSM-5 alcohol use disorder criteria in routine care: High test-retest reliability of an Alcohol Symptom Checklist

Kevin A. Hallgren, PhD[1,2,3], **Theresa E. Matson, MPH**[2,3], **Malia Oliver, BA**[2], **Ryan M. Caldeiro, MD**[4], **Daniel Kivlahan, PhD**[5], **Katharine A. Bradley, MD MPH**[2,3,6]

[1]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, United States

[2]Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States

[3]University of Washington, Department of Health Systems and Population Health, Seattle, WA, United States

[4]Mental Health and Wellness, Kaiser Permanente of Washington, Renton, WA.

[5]Center of Innovation for Veteran-Centered and Value-Driven Care, Health Services Research and Development, Veteran Affairs Puget Sound HealthCare System, Seattle, WA, United States

[6]Department of Medicine, University of Washington, Seattle, WA, United States

Abstract

Background: Alcohol use disorder (AUD) is underdiagnosed and undertreated in medical settings, in part due to a lack of AUD assessment instruments that are reliable and practical for use in routine care. This study evaluates the test-retest reliability of a patient-report Alcohol Symptom Checklist questionnaire when it is used in routine care, including primary care and mental health specialty settings.

Methods: We performed a pragmatic test-retest reliability study using electronic health record (EHR) data from Kaiser Permanente Washington, an integrated health system in Washington state. The sample included 454 patients who reported high-risk drinking on a behavioral health screen and completed two Alcohol Symptom Checklists 1–21 days apart. Subgroups who completed both checklists in primary care (n=271) or mental health settings (n=79) were also examined. The primary measure was an Alcohol Symptom Checklist on which patients self-report whether they have experienced each of the 11 AUD criteria within the past year, as defined by the Diagnostic and Statistical Manual of Mental Disorders-5th edition (DSM-5).

Results: Alcohol Symptom Checklists completed in routine care and documented in EHRs had excellent test-retest reliability for measuring AUD criteria counts (ICC=0.79, 95% CI: 0.76–0.82). Test-retest reliability estimates were also high and not significantly different for the subsamples of patients who completed both checklists in primary care (ICC=0.82, 95% CI: 0.77–0.85) or mental

Correspondence: Kevin A. Hallgren, PhD. Address: Department of Psychiatry and Behavioral Sciences, University of Washington Medical Center, Box 356560, 1959 NE Pacific St., Seattle, WA 98195, United States. khallgre@uw.edu. Phone: +1 (206) 616-2906. Fax: +1 (206) 543-9520.

health settings (ICC=0.74, 95% CI: 0.62–0.83). Test-retest reliability was not moderated by having a past two-year AUD diagnosis, nor by the age or sex of the patient completing it.

Conclusions: Alcohol Symptom Checklists can reliably and pragmatically assess AUD criteria in routine care among patients who screen positive for high-risk drinking. The Alcohol Symptom Checklist may be a valuable tool in supporting AUD-related care and monitoring AUD criteria longitudinally in routine primary care and mental health settings.

Keywords

alcohol use disorder; assessment; measurement-based care; primary care; symptom checklist

INTRODUCTION

Excessive drinking accounts for 5% of years of life lost due to early mortality and 3% of deaths in the United States (US; Rehm et al., 2009, 2014). In the US, 39.0% of adults drink above recommended limits (Chen et al., 2016) and 13.9% of adults meet past-year criteria for alcohol use disorder (AUD; Grant et al., 2015). Most adults with AUD utilize primary care (Mintz et al., 2021). However, for the vast majority of patients with AUD who visit primary care, AUD goes undiagnosed (Williams et al., 2014) and untreated (Glass et al., 2016; Hallgren et al., 2020; Rieckmann et al., 2016; Williams et al., 2017) even though effective AUD treatments can be delivered from primary care and specialty mental health settings (Jonas et al., 2014; Oslin et al., 2014; Watkins et al., 2017).

Currently, there is a lack of tools for medical providers to practically and reliably assess AUD criteria, which impedes their ability to diagnose and treat AUD. Practical and reliable alcohol *screening* measures, such as the Alcohol Use Disorders Identification Test-Consumption version (AUDIT-C; Bradley et al., 2003; Bush et al., 1998), are increasingly used in routine care and higher scores on these screening measures are associated with a higher probability of AUD. However, the AUDIT-C does not assess criteria required for an AUD diagnosis as defined by the Diagnostic and Statistical Manual of Mental Disorders-5th Edition (DSM-5; American Psychiatric Association, 2013) and it does not provide information about the number of AUD criteria that are present, which is required for determining the severity of an AUD diagnosis when it is present (i.e., mild, moderate, or severe AUD). Moreover, within a general US population sample, only two-thirds of people with the highest AUDIT-C scores (i.e., 12 points on a 0–12 scale) meet criteria for AUD and the association between AUDIT-C scores and AUD criteria varies across age groups (Rubinsky et al., 2013). Assessing AUD criteria via direct patient report can provide the information that is needed for correctly diagnosing AUD and determining its severity. Moreover, assessing AUD criteria, all of which constitute negative consequences attributable to alcohol, can provide opportunities for clinicians to engage patients in discussions about those criteria and other potential harms caused by drinking, which can be helpful for introducing and discussing treatment options. Although well-validated interviews for diagnosing AUD are available for use in research (e.g., Hasin et al., 2020), they require considerable time and training to administer, making them impractical for assessing AUD criteria in most routine health care settings.

Starting in 2015, Kaiser Permanente (KP) Washington, an integrated health system in Washington state, began implementing universal alcohol screening, followed by AUD symptom assessment using an Alcohol Symptom Checklist (Bobb et al., 2017; Sayre et al., 2020) among patients who reported high-risk drinking on the AUDIT-C (scores of 7–12; Rubinsky et al., 2013). As a result of this implementation, most patients with high-risk drinking complete Alcohol Symptom Checklists (78% of eligible patients as of November 2021), supporting the feasibility of implementing AUD symptom assessment with Alcohol Symptom Checklists after patients report high-risk drinking in a large health system. The Alcohol Symptom Checklist was primarily designed to help engage patients and providers in clinical discussions about alcohol use and associated consequences and to provide information that could help providers determine if AUD is present (e.g., if ≥2 AUD criteria are reported by the patient and clinician determines they are recurrent). The Alcohol Symptom Checklist could also support providers in determining whether the AUD was mild (2–3 criteria), moderate (4–5 criteria), or severe (≥6 criteria; Bobb et al., 2017; Bradley et al., 2019; Marsden et al., 2019; Sayre et al., 2020).

These Alcohol Symptom Checklists have been completed in routine care by over 11,000 patients who reported high-risk drinking on alcohol screening. Previous cross-sectional analyses have shown that patients with high-risk drinking who complete Alcohol Symptom Checklists in routine care frequently endorse AUD symptoms at levels that warrant offering AUD treatment (Sayre et al., 2020). More recent analyses showed that the Alcohol Symptom Checklist's items are able to discriminate AUD severity in the expected manner and that the checklist measures AUD severity along a scaled, unidimensional continuum of severity, which is consistent with current diagnostic conceptualizations of AUD (American Psychiatric Association, 2013; Hallgren et al., 2021a) and supports the Alcohol Symptom Checklist's construct validity. These analyses also showed that the checklist discriminates AUD severity in a consistent manner regardless of the age, sex, race, or ethnicity of the patient completing it (Hallgren et al., 2021a).

However, no studies have evaluated the test-retest reliability of the Alcohol Symptom Checklist when it is used in routine care, even though such an evaluation is critical for understanding how the measure performs when it is administered repeatedly over time. For example, patients may complete multiple Alcohol Symptom Checklists over time as part of clinical monitoring and measurement-based care (Bradley et al., 2019; Marsden et al., 2019), and test-retest reliability analyses can help providers distinguish whether changes in a patient's self-reported AUD criteria are likely reflective of a reliable change in AUD criteria rather than an artifact of poor test-retest reliability. In the current study, we report the Alcohol Symptom Checklist's test-retest reliability and thresholds for reliable change when it is completed as part of routine care. To maximize external validity in this study, we utilize data from Alcohol Symptom Checklists that were completed as part of routine clinical care (i.e., the sample was not recruited for research and checklists were completed in real-world routine care conditions). To enhance internal validity in the study, we evaluate Alcohol Symptom Checklists completed within a 1- to 21-day test-retest window, which reflects an assessment window in which past-year AUD criteria are unlikely have substantial changes, and therefore most of the observed changes in Alcohol Symptom Checklist scores

within this window are likely to be attributable to test-retest reliability issues and are less likely to be confounded with actual changes in patients' AUD criteria.

MATERIALS AND METHODS

Study Setting

This is a pragmatic test-retest reliability study using clinical data from electronic health records (EHRs) from KP Washington.

Alcohol Screening and Assessment Procedures

As part of an effort to integrate behavioral health across the health system, 33 KP Washington primary care practices implemented annual alcohol screening via the AUDIT-C starting in 2015 (Glass et al., 2018; Yeung et al., 2020). When patients report high-risk drinking on the AUDIT-C (score 7–12; Rubinsky et al., 2013), the EHR prompts medical staff to administer an Alcohol Symptom Checklist. The EHR only prompts one Alcohol Symptom Checklist per year, but checklists may be completed more frequently if a clinician chooses to administer the checklist ad hoc or if treatment plans include monitoring symptoms over time.

In primary care settings, procedures for completing Alcohol Symptom Checklists are standardized: for patients who have screened positive for high-risk drinking within the past year (including on the date of the primary care appointment) the EHR prompts medical assistants to give a paper form with the Alcohol Symptom Checklist. Patients complete the Alcohol Symptom Checklist in writing and medical assistants enter the results into the EHR. Primary care medical assistants received training in procedures for administering the checklist.

The AUDIT-C is also included on a mental health monitoring tool used in specialty mental health settings. During the current study's observation period, when a patient reported high-risk drinking on the AUDIT-C in a mental health setting, the EHR prompted clinicians to administer an Alcohol Symptom Checklist; however, procedures for administering the checklist were less standardized in these settings and paper Alcohol Symptom Checklists were not routinely stocked or used to our knowledge (e.g., clinicians in mental health settings could have administered the checklist in ad hoc ways, including by reading the checklist questions aloud to patients). Further, there was no standardized training in procedures for administering Alcohol Symptom Checklists in mental health settings.

Patient Population

Patients were eligible for this study if they (a) had ≥1 visit to a KP Washington primary care setting between October 1, 2015 and February 29, 2020, (b) screened positive for high-risk drinking on the AUDIT-C (score 7–12), (c) completed two Alcohol Symptom Checklists 1–21 days apart, with the first being within 365 days after the positive AUDIT-C screen, and (d) were at least 18 years old when the checklists were completed. The study was approved by the KP Washington Health Research Institute's Institutional Review Board with a waiver of consent and HIPAA authorization to use existing EHR data.

Measures

AUDIT-C—The AUDIT-C is a brief measure validated as a screen for unhealthy alcohol use and AUD (Bradley et al., 2003; Bush et al., 1998). It has three items that assess the frequency of drinking, typical drinks per drinking day, and frequency of heavy drinking. Items are answered on a 5-point scale (0–4 points) then summed (total score 0–12 points).

Alcohol Symptom Checklist—The Alcohol Symptom Checklist (available in supplement; Hallgren et al., 2021a; Sayre et al., 2020) is an 11-item questionnaire that asks patients to self-report whether they have experienced each of the 11 DSM-5 AUD criteria within the past year, a timeframe consistent with the DSM-5. Total scores reflect summed criteria counts (0–11) that may assist with determining if an AUD diagnosis is present (≥2 criteria endorsed) and its severity designation (i.e., mild: 2–3 criteria; moderate: 4–5 criteria; severe: 6–11 criteria).

Descriptive measures and covariates—Age, sex, race, and ethnicity were obtained from EHR data. Medicaid and Medicare insurance status were obtained from enrollment records. AUD diagnoses by healthcare providers up to two years before completing Alcohol Symptom Checklists were identified using ICD-9 and ICD-10 codes from EHRs and insurance claims.

Analytic Approach

Test-retest reliability was evaluated for the full sample of patients with two Alcohol Symptom Checklists then within stratified subsamples of patients who completed both Alcohol Symptom Checklists in primary care or in mental health settings. We hypothesized that test-retest reliability would be higher when checklists were completed in primary care settings where self-administered paper forms are used, compared to when checklists were completed in mental health settings that had non-standard administration, often without paper self-administration.

Statistical analyses were modeled after a recent study by Hasin et al. (2020), which evaluated test-retest reliability of confidential, clinician-administered diagnostic research interviews in people who screened positive for past 30-day substance use and substance use disorder criteria. Mean changes in the number of criteria endorsed from T1 to T2 (i.e., change scores) were characterized using paired sample *t*-tests and Cohen's *d* effect size coefficients. Test-retest reliability coefficients were estimated for four composite measures derived from the number of AUD criteria reported on the checklist, including total scores reflecting AUD symptom counts (0–11), estimated using a one-way single-measures agreement intraclass correlation coefficient (ICC; McGraw & Wong, 1996); AUD severity designation (none, mild, moderate, severe), estimated using weighted kappa (Cohen, 1968); and binary measures indicating criteria consistent with AUD (2+ criteria) versus no AUD (0 or 1 criteria) or criteria consistent with moderate or severe AUD (4+ criteria) verse no AUD or mild AUD (0–3 criteria), estimated using kappa (Cohen, 1960). Test-retest reliabilities for each of the 11 individual AUD criteria were estimated using kappa. Reliability coefficients were interpreted using cutoffs described by Cicchetti (1994) to indicate excellent (0.75–1.00), good (0.60–0.74), fair (0.40–0.59), or poor reliability (≤0.39). Testing for significant

differences between reliability coefficients obtained in the primary care versus mental health subsamples was conducted using parametric bootstrapping with 10,000 resampled coefficient estimates.

Additional analyses aimed to identify predictors of test-retest reliability by testing whether rates of agreement on the checklists completed at the first (T1) and second (T2) time points differed based on patient age, sex, or past 2-year AUD diagnosis by a healthcare provider. For these analyses, a dichotomous code was created to reflect that T1 and T2 checklists agreed (agreement=1) or disagreed (agreement=0), with agreements defined by both checklists reflecting moderate or severe AUD (4–11 criteria) or both reflecting mild or no AUD (0–3 criteria). This definition of agreement was based on KP Washington's alcohol decision support tools that suggest offering AUD medications to patients with moderate or severe AUD. Log-likelihood ratio chi-square tests from logistic regression models were used to test for differences in the odds of discordance across the levels of each covariate.

We used the reliable change index (RCI; Jacobson & Truax, 1992) to identify how large of a change in the number of AUD criteria *an individual patient* would need to report for their medical provider to conclude that they had a statistically reliable change in their AUD criteria at the $p < .20$, $p < .10$, or $p < .05$ level (Wise, 2004). The RCI indicates whether the magnitude of change in test scores for a single patient is significantly larger than the degree of variation that would be expected due to test-retest-related measurement error alone. The RCI is calculated as:

$$RCI = \frac{X_2 - X_1}{S_{diff}}$$

where X_2-X_1 reflects the change score for a single patient and S_{diff} is the standard error of the difference in test scores, computed as:

$$S_{diff} = \sqrt{2SD^2\left(1 - r_{xx}\right)}$$

In the formula above, SD is the sample standard deviation of checklist scores at T1 and r_{xx} is the sample test-retest reliability coefficient for the number of criteria endorsed. We computed the magnitude of the changes in scores (X_2-X_1) that would be required to produce RCIs >1.28, 1.65, and 1.96, as these indicate that the difference score (X_2-X_1) is significantly larger than what would be expected by test-retest related measurement error alone at the $p < 0.20$, $p < 0.10$, and $p < 0.05$ levels, respectively (Wise, 2004). These three thresholds for reliable change are presented to provide a range of thresholds for reliable change associated with varying levels of confidence for concluding whether reliable change has occurred. A larger change score is needed to exceed the threshold of $p < 0.05$, and using this conservative threshold may result in some patients who experience "true" changes in their AUD criteria not having large enough change scores to conclude that reliable change has occurred at the $p < 0.05$ level, resulting in higher type-II error rates about whether individual patients have experienced reliable change (i.e., clinician fails to detect changes in AUD symptoms, even though symptoms have changed). In contrast, a smaller change

score is needed to exceed the threshold of $p < 0.20$, and using this more liberal threshold may result in some patients exceeding the threshold for concluding that reliable change has occurred even if they did not experience "true" changes in their AUD symptoms, resulting in higher type-I error rates about whether individual patients have experienced reliable change (i.e., clinician concludes that AUD symptoms have reliably changed, even though changes in scores may be due to test-retest-related measurement error rather than "true" changes in AUD symptoms). We present a range of thresholds associated with different confidence levels because the degree of confidence desired for concluding reliable change may vary from context to context and because each threshold carries a different balance of risk for type-I and type-II errors. We therefore encourage clinicians to utilize the reliable change thresholds presented here as one source of evidence about the likelihood of reliable change being present, rather than using them as firmly conclusive indicators about whether a patient has or has not experienced change.

RESULTS

Patient Characteristics

There were 11,452 patients who completed one or more Alcohol Symptom Checklist during the study period. Among them, 454 patients completed two checklists 1–21 days apart and comprised the current study sample. Differences between these patients and the 10,998 patients not included in the sample are reported in the Supplement; in brief, the test-retest reliability sample had a higher proportion of patients who were less than 45 years old, were female, reported criteria consistent with severe AUD, had a past two-year AUD diagnosis from a healthcare provider, and completed symptom checklists in mental health settings rather than in primary care settings.

Within the study sample (N=454), 271 patients completed both checklists in primary care settings (primary care subsample) and 79 completed both checklists in mental health settings (mental health subsample). The remaining 104 patients who completed T1 and T2 checklists in different settings were not included in subgroup analyses. In the full study sample and both study subsamples (Table 1), most patients were age 25–44, white, non-Hispanic, and reported criteria consistent with severe AUD (6+ criteria). Compared to the primary care subsample, the mental health subsample had a higher proportion of patients who were female, aged 18–24, and diagnosed with AUD by healthcare provider in the past two years. The primary care and mental health clinic subsamples did not differ by race, ethnicity, or T1 Alcohol Symptom Checklist scores.

AUD Criterion Assessments

In the full test-retest reliability sample, patients completed checklists a mean of 9.31 (SD=5.92) days apart and reported a mean of 6.25 (SD=3.40) AUD criteria at T1 and 5.89 (SD=3.59) criteria at T2, a significant difference with a small effect size (mean difference=−0.35 criteria, SD=2.23, t_{453}=−3.37, p=.001, d=−0.16). In the primary care clinic subsample, patients completed checklists a mean of 9.44 (SD=5.99) days apart and reported a mean of 6.24 (SD=3.39) AUD criteria at T1 and 5.98 (SD=3.60) criteria at T2 (mean difference=−0.27 criteria, SD=2.10, t_{270}=−2.08, p=.04, d=−0.12). In the mental health

clinic subsample, patients completed checklists a mean of 10.47 (*SD*=6.11) days apart and reported a mean of 5.53 (*SD*=3.60) AUD criteria at T1 and 4.89 (*SD*=3.81) criteria at T2 (mean difference=−0.65 criteria, SD=2.61, t_{70}=−2.20, p=.03, d=−0.25).

Test-Retest Reliability

In the full sample, test-retest reliability was excellent for the number of AUD criteria (ICC=0.79), good for the AUD severity designation (severe, moderate, mild, or no AUD; weighted kappa=0.74), fair for the binary indicator of any AUD (vs. no AUD; kappa=0.57), and good for the binary indicator of moderate/severe AUD (vs. no or mild AUD; kappa=0.63). Reliability coefficients ranged from 0.44 to 0.70 for the 11 individual AUD criteria (Table 2).

In the primary care subsample, test-retest reliability was excellent for the number of AUD criteria (ICC=0.82), excellent for the AUD severity designation (weighted kappa=0.75), fair for the binary indicator of any AUD (kappa=0.58), and good for the binary indicator of moderate or severe AUD (kappa=0.62). Coefficients ranged from 0.42 to 0.76 for the 11 individual AUD criteria (Table 2).

In the mental health subsample, test-retest reliability was good for the number of AUD criteria (ICC=0.74), good for the AUD severity designation (weighted kappa=0.73), fair for the binary indicator of any AUD (kappa=0.51), and good for the binary indicator of moderate or severe AUD (kappa=0.62). Reliability coefficients ranged from 0.36 to 0.60 for the 11 individual AUD criteria (Table 2). Although most of the reliability coefficients were nominally higher in the primary care subsample, parametric bootstrapping analyses indicated that there were no significant differences between the primary care and mental health subsamples for any of the four full-scale reliability coefficients (Table 2).

Predictors of Test-Retest Discordance

Log-likelihood ratio tests examined predictors of test-retest discordance, defined as checklist scores that suggested moderate or severe AUD (4–11 criteria) at one occasion but not the other. These analyses were stratified within each subsample to reduce the influence of setting as a potential confounder, given that some predictors of test-retest discordance also differed between the primary care and mental health subsamples.

In both subsamples, rates of discordance were not associated with age[1], sex, Medicaid insurance, Medicare insurance, or a past two-year AUD diagnosis from a healthcare provider (Table 3).

Reliable Change Thresholds

Table 4 shows the magnitude of change in total scores (AUD criteria counts) required for concluding that an individual patient has experienced reliable change in criteria at the $p < 0.20$, $p < 0.10$, and $p < 0.05$ levels. For the full sample (i.e., not accounting for primary care or mental health setting), patients would need to report a change of at least 3, 4, or 5 criteria

[1]For the mental health clinic subsample, individuals age 65+ were excluded from this analysis due to an insufficient sample size.

to conclude reliable change with $p<0.20$, $p<0.10$ or $p<0.05$, respectively. In the primary care subsample, patients would need to report a change of at least 3 criteria to conclude reliable change with $p<0.20$ or at least 4 criteria to conclude reliable change with $p<0.10$ or $p<0.05$. In the mental health subsample, a change of at least 4, 5, or 6 criteria was necessary to conclude reliable change with $p<0.20$, $p<0.10$, or $p<0.05$, respectively.

DISCUSSION

The current study found that a pragmatic, self-report Alcohol Symptom Checklist had good-to-excellent test-retest reliability when it was completed by patients as part of routine care after screening positive for high-risk drinking. Test-retest reliability estimates for the primary care subsample (n=271), where procedures for administering the checklists were most standardized, were only modestly lower than the test-retest reliability estimates obtained from a recent study by Hasin et al. (2020), who evaluated gold-standard semi-structured diagnostic interviews administered by clinicians with several years of addiction clinical experience who were well-trained and well-supervised in administering the diagnostic interviews. That study found excellent test-retest reliability estimates for AUD criteria counts (ICC=0.90) and AUD severity (ICC=0.87), and good test-retest reliability for a binary measure reflecting any AUD versus no AUD (kappa=0.69; Hasin et al., 2020). Our findings suggest that test-retest reliability is also high for Alcohol Symptom Checklists completed on paper forms in routine primary care, with reliability coefficients observed here only being modestly lower than that of clinician-administered diagnostic interviews (Hasin et al., 2020) that used a gold-standard research interviews that are not practical for most real-world healthcare settings (Bradley et al., 2019).

In the current study, test-retest reliability was nominally (but non-significantly) lower for checklists completed in specialty mental health settings, where procedures for completing checklists were not standardized, paper forms were not routinely stocked, and staff received little or no training for administering the checklist. This non-significant difference contrasts our hypothesis and could be due in part to limited power for subsample analyses. A prior study that observed medical staff as they administered alcohol screening measures in Veteran Affairs settings found that verbal reading of alcohol screening questions and non-standardized administration (e.g., adapting question wording, omitting questions, suggesting or inferring responses) were common practices that could likely reduce the accuracy of results; therefore, the use of paper Alcohol Symptom Checklist forms may still be a helpful standard practice to promote standardized administration and more accurate patient responding (Williams et al., 2015).

Our findings suggest that the Alcohol Symptom Checklist may be a reliable tool for assessing AUD criteria and for supporting clinicians in diagnosing AUD and determining its severity (mild, moderate, or severe) in routine care settings. In particular, good-to-excellent test-retest reliability observed in the full sample and the primary care subsample indicate that the measure may be a viable tool for monitoring whether patients are experiencing *changes* in their AUD criteria over time, for example, when checklists are used for clinical monitoring or measurement-based care (Bradley et al., 2019; Hallgren et al., 2021a, 2021b). Reliable change analyses indicated that an individual primary care patient who increases

or decreases the number of AUD criteria reported by at least 3 criteria may be concluded to have experienced statistically reliable change (i.e., change that is unlikely attributable to measurement error alone) at the $p < .20$ level, and that a primary care patient who increases or decreases the number of AUD criteria by at least 4 criteria may be concluded to have experienced reliable change at the $p < .05$ level.

Limitations and Strengths

Our study has noteworthy limitations. Because we evaluated the performance of Alcohol Symptom Checklists completed in routine care, we were unable to measure or control for many factors that could impact symptom reporting and reliability (e.g., patient experiences with medical assistants). We also were unable to know why the patients in our study completed two Alcohol Symptom Checklists 1–21 days apart, and it is likely there was selection bias within the test-retest sample that we were unable to measure. For example, patients in this sample may have had providers who opted to re-assess AUD criteria for clinical monitoring due to their high frequency of criteria consistent with severe AUD. Nonetheless, the likelihood that many patients were undergoing clinical monitoring is also a strength of the study, as this sample may more accurately reflect patients for whom test-retest reliability analyses are highly relevant. It is possible that patients reported fewer AUD criteria at both time points than they would have in confidential research interviews due to the checklists being completed in routine healthcare settings and the results being shared with medical providers and entered into the EHR. It is also possible that patients who completed checklists for clinical monitoring experienced actual changes in AUD criteria during the test-retest period and those changes would have been modeled as measurement error in the current study, producing lower estimates of test-retest reliability and higher estimates of thresholds for reliable change. Thus, it is possible the results observed here are conservative underestimates of the Alcohol Symptom Checklist's test-retest reliability. Our sample was limited by being predominantly white and non-Hispanic, and sample sizes for racial and ethnic minority subgroups were insufficient to test for differences in test-retest reliability across these subgroups; additional efforts to evaluate test-retest reliability of Alcohol Symptom Checklists among patients of color is warranted.

Our study also had several strengths. Test-retest reliability was evaluated in real-world routine-care conditions, which provides high external validity for the procedures used when completing the checklists and reduces the likelihood of biasing our sample to only include people who are willing to participate in AUD-related research. The 1- to 21-day test-retest window reflected a period in which past-year AUD criteria have a low likelihood of substantively changing, and thus the study design was more capable of isolating sources of measurement error (i.e., test-retest unreliability) while minimizing confounding that could be attributable to actual changes in AUD criteria, which improves the internal validity of the study. The large sample size allowed us to obtain reasonably precise reliability estimates and to test for factors that may increase or decrease test-retest reliability in practice. The high test-retest reliability observed here strengthens recent findings supporting the construct validity and measurement invariance of the Alcohol Symptom Checklist (Hallgren et al., 2021a). The Alcohol Symptom Checklist was designed for use in routine real-world care, is

highly pragmatic, and can be easily administered in routine primary care settings with paper and pencil forms.

Conclusion

AUD criteria can be reliably assessed using an Alcohol Symptom Checklist in routine primary care among patients who screen positive for high-risk drinking. Using Alcohol Symptom Checklists in routine care can help providers diagnose AUD, determine its severity, and monitor changes in AUD symptoms over time.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

Research reported in this publication was supported by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) of the National Institutes of Health (NIH) under award numbers R21AA028073, R33AA028073, and K01AA024796. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIAAA or the NIH.

REFERENCES

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub.

Bobb JF, Lee AK, Lapham GT, Oliver M, Ludman E, Achtmeyer C, Parrish R, Caldeiro RM, Lozano P, Richards JE, & Bradley KA (2017). Evaluation of a Pilot Implementation to Integrate Alcohol-Related Care within Primary Care. International Journal of Environmental Research and Public Health, 14(9). 10.3390/ijerph14091030

Bradley KA, Bush KR, Epler AJ, Dobie DJ, Davis TM, Sporleder JL, Maynard C, Burman ML, & Kivlahan DR (2003). Two brief alcohol-screening tests From the Alcohol Use Disorders Identification Test (AUDIT): Validation in a female Veterans Affairs patient population. Archives of Internal Medicine, 163(7), 821–829. 10.1001/archinte.163.7.821 [PubMed: 12695273]

Bradley KA, Caldeiro RM, Hallgren KA, & Kivlahan DR (2019). Making measurement-based care for addictions a reality in primary care. Addiction, 114(8), 1355–1356. [PubMed: 31037777]

Bush K, Kivlahan DR, McDonell MB, Fihn SD, & Bradley KA (1998). The AUDIT alcohol consumption questions (AUDIT-C): An effective brief screening test for problem drinking. Ambulatory Care Quality Improvement Project (ACQUIP). Alcohol Use Disorders Identification Test. Archives of Internal Medicine, 158(16), 1789–1795. [PubMed: 9738608]

Chen CM, Slater ME, Castle IJP, & Grant BF (2016). Alcohol use and alcohol use disorders in the United States: Main findings from the 2012–2013 National Epidemiologic Survey on Alcohol and Related Conditions–III (NESARC-III). In US Alcohol Epidemiologic Data Reference Manual, Volume 10, April 2016, NIH Publication No. 16-AA-8020. National Institute on Alcohol Abuse and Alcoholism, Bethesda, MD.

Cicchetti DV (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6(4), 284–290. 10.1037/1040-3590.6.4.284

Cohen J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, 20(1), 37–46. 10.1177/001316446002000104

Cohen J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213. [PubMed: 19673146]

Glass JE, Bobb JF, Lee AK, Richards JE, Lapham GT, Ludman E, Achtmeyer C, Caldeiro RM, Parrish R, Williams EC, Lozano P, & Bradley KA (2018). Study protocol: A cluster-randomized

trial implementing Sustained Patient-centered Alcohol-related Care (SPARC trial). Implementation Science: IS, 13(1), 108. 10.1186/s13012-018-0795-9 [PubMed: 30081930]

Glass JE, Bohnert KM, & Brown RL (2016). Alcohol Screening and Intervention Among United States Adults who Attend Ambulatory Healthcare. Journal of General Internal Medicine, 31(7), 739–745. 10.1007/s11606-016-3614-5 [PubMed: 26862079]

Grant BF, Goldstein RB, Saha TD, Chou SP, Jung J, Zhang H, Pickering RP, Ruan WJ, Smith SM, Huang B, & Hasin DS (2015). Epidemiology of DSM-5 Alcohol Use Disorder: Results From the National Epidemiologic Survey on Alcohol and Related Conditions III. JAMA Psychiatry, 72(8), 757–766. 10.1001/jamapsychiatry.2015.0584 [PubMed: 26039070]

Hallgren KA, Holzhauer CG, Epstein EE, McCrady BS, & Cook S. (2021b). Optimizing the length and reliability of measures of mechanisms of change to support measurement-based care in alcohol use disorder treatment. Journal of Consulting and Clinical Psychology, 89(4), 277. [PubMed: 34014690]

Hallgren KA, Matson TE, Oliver M, Witkiewitz K, Bobb JF, Lee AK, Caldeiro RM, Kivlahan D, & Bradley KA (2021a). Practical Assessment of Alcohol Use Disorder in Routine Primary Care: Performance of an Alcohol Symptom Checklist. Journal of General Internal Medicine. 10.1007/s11606-021-07038-3

Hallgren KA, Witwer E, West I, Baldwin L-M, Donovan D, Stuvek B, Keppel GA, Mollis B, & Stephens KA (2020). Prevalence of documented alcohol and opioid use disorder diagnoses and treatments in a regional primary care practice-based research network. Journal of Substance Abuse Treatment, 110, 18–27. 10.1016/j.jsat.2019.11.008 [PubMed: 31952624]

Hasin D, Shmulewitz D, Stohl M, Greenstein E, Roncone S, Aharonovich E, & Wall M. (2020). Test-retest reliability of DSM-5 substance disorder measures as assessed with the PRISM-5, a clinician-administered diagnostic interview. Drug and Alcohol Dependence, 216, 108294. 10.1016/j.drugalcdep.2020.108294

Jacobson NS, & Truax P. (1992). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research.

Jonas DE, Amick HR, Feltner C, Bobashev G, Thomas K, Wines R, Kim MM, Shanahan E, Gass CE, Rowe CJ, & Garbutt JC (2014). Pharmacotherapy for adults with alcohol use disorders in outpatient settings: A systematic review and meta-analysis. JAMA, 311(18), 1889–1900. 10.1001/jama.2014.3628 [PubMed: 24825644]

Marsden J, Tai B, Ali R, Hu L, Rush AJ, & Volkow N. (2019). Measurement-based care using DSM-5 for opioid use disorder: Can we make opioid medication treatment more effective? Addiction (Abingdon, England). 10.1111/add.14546

McGraw KO, & Wong SP (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods, 1(1), 30.

Mintz CM, Hartz SM, Fisher SL, Ramsey AT, Geng EH, Grucza RA, & Bierut LJ (2021). A cascade of care for alcohol use disorder: Using 2015–2019 National Survey on Drug Use and Health data to identify gaps in past 12-month care. Alcoholism: Clinical and Experimental Research, 45(6), 1276–1286. 10.1111/acer.14609

Oslin DW, Lynch KG, Maisto SA, Lantinga LJ, McKay JR, Possemato K, Ingram E, & Wierzbicki M. (2014). A Randomized Clinical Trial of Alcohol Care Management Delivered in Department of Veterans Affairs Primary Care Clinics Versus Specialty Addiction Treatment. Journal of General Internal Medicine, 29(1), 162–168. 10.1007/s11606-013-2625-8 [PubMed: 24052453]

Rehm J, Dawson D, Frick U, Gmel G, Roerecke M, Shield KD, & Grant B. (2014). Burden of Disease Associated with Alcohol Use Disorders in the United States. Alcoholism: Clinical and Experimental Research, 38(4), 1068–1077. 10.1111/acer.12331

Rehm J, Mathers C, Popova S, Thavorncharoensap M, Teerawattananon Y, & Patra J. (2009). Global burden of disease and injury and economic cost attributable to alcohol use and alcohol-use disorders. The Lancet, 373(9682), 2223–2233. 10.1016/S0140-6736(09)60746-7

Rieckmann T, Muench J, McBurnie MA, Leo MC, Crawford P, Ford D, Stubbs J, O'Cleirigh C, Mayer KH, Fiscella K, Wright N, Doe-Simkins M, Cuddeback M, Salisbury-Afshar E, & Nelson C. (2016). Medication-assisted treatment for substance use disorders within a national community health center research network. Substance Abuse, 37(4), 625–634. 10.1080/08897077.2016.1189477 [PubMed: 27218678]

Rubinsky AD, Dawson DA, Williams EC, Kivlahan DR, & Bradley KA (2013). AUDIT-C scores as a scaled marker of mean daily drinking, alcohol use disorder severity, and probability of alcohol dependence in a U.S. general population sample of drinkers. Alcoholism, Clinical and Experimental Research, 37(8), 1380–1390. 10.1111/acer.12092

Sayre M, Lapham GT, Lee AK, Oliver M, Bobb JF, Caldeiro RM, & Bradley KA (2020). Routine Assessment of Symptoms of Substance Use Disorders in Primary Care: Prevalence and Severity of Reported Symptoms. Journal of General Internal Medicine, 35(4), 1111–1119. 10.1007/s11606-020-05650-3 [PubMed: 31974903]

Watkins KE, Ober AJ, Lamp K, Lind M, Setodji C, Osilla KC, Hunter SB, McCullough CM, Becker K, Iyiewuare PO, Diamant A, Heinzerling K, & Pincus HA (2017). Collaborative Care for Opioid and Alcohol Use Disorders in Primary Care: The SUMMIT Randomized Clinical Trial. JAMA Internal Medicine, 177(10), 1480–1488. 10.1001/jamainternmed.2017.3947 [PubMed: 28846769]

Williams EC, Achtmeyer CE, Thomas RM, Grossbard JR, Lapham GT, Chavez LJ, Ludman EJ, Berger D, & Bradley KA (2015). Factors Underlying Quality Problems with Alcohol Screening Prompted by a Clinical Reminder in Primary Care: A Multi-site Qualitative Study. Journal of General Internal Medicine, 30(8), 1125–1132. 10.1007/s11606-015-3248-z [PubMed: 25731916]

Williams EC, Gupta S, Rubinsky AD, Glass JE, Jones-Webb R, Bensley KM, & Harris AHS (2017). Variation in receipt of pharmacotherapy for alcohol use disorders across racial/ethnic groups: A national study in the U.S. Veterans Health Administration. Drug and Alcohol Dependence, 178, 527–533. 10.1016/j.drugalcdep.2017.06.011 [PubMed: 28728114]

Williams EC, Rubinsky AD, Lapham GT, Chavez LJ, Rittmueller SE, Hawkins EJ, Grossbard JR, Kivlahan DR, & Bradley KA (2014). Prevalence of clinically recognized alcohol and other substance use disorders among VA outpatients with unhealthy alcohol use identified by routine alcohol screening. Drug and Alcohol Dependence, 135, 95–103. 10.1016/j.drugalcdep.2013.11.016 [PubMed: 24360928]

Wise EA (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. Journal of Personality Assessment, 82(1), 50–59. [PubMed: 14979834]

Yeung K, Richards J, Goemer E, Lozano P, Lapham G, Williams E, Glass J, Lee A, Achtmeyer C, Caldeiro R, Parrish R, & Bradley K. (2020). Costs of using evidence-based implementation strategies for behavioral health integration in a large primary care system. Health Services Research, 55(6), 913–923. 10.1111/1475-6773.13592 [PubMed: 33258127]

Table 1

Descriptive Statistics for Full Sample and for Primary Care and Mental Health Subsamples

		Full sample (N = 454)		Primary care subsample (n = 271)		Mental health subsample (n = 79)		P-value of difference between subsamples
		n	(%)	n	(%)	n	(%)	
Age (y)	18–24	57	(12.6%)	23	(8.5%)	15	(19.0%)	**p = .01**
	25–44	238	(52.4%)	142	(52.4%)	45	(57.0%)	
	45–64	130	(28.6%)	86	(31.7%)	17	(21.5%)	
	65+	29	(6.4%)	20	(7.4%)	2	(2.5%)	
Sex	Female	197	(43.4%)	107	(39.5%)	42	(53.2%)	**p = .04**
	Male	257	(56.6%)	164	(60.5%)	37	(46.8%)	
Race	Asian or Asian American	16	(3.5%)	9	(3.3%)	4	(5.1%)	p = .82
	Black or African American	18	(4.0%)	10	(3.7%)	3	(3.8%)	
	Native American or Alaskan Native	8	(1.8%)	3	(1.1%)	1	(1.3%)	
	Native Hawaiian or Pacific Islander	0	(0.0%)	0	(0.0%)	0	(0.0%)	
	White	364	(80.2%)	225	(83.0%)	63	(79.7%)	
	More than one race	13	(2.9%)	6	(2.2%)	3	(3.8%)	
	Other race	13	(2.9%)	5	(1.8%)	0	(0.0%)	
	Unknown race	22	(4.8%)	13	(4.8%)	5	(6.3%)	
Ethnicity	Hispanic	27	(5.9%)	18	(6.6%)	2	(2.5%)	p = .36
	Not Hispanic	405	(89.2%)	239	(88.2%)	72	(91.1%)	
	Unknown ethnicity	22	(4.8%)	14	(5.2%)	5	(6.3%)	
AUD Severity (T1)	No AUD (0 or 1 symptoms)	54	(11.9%)	31	(11.4%)	13	(16.5%)	p = .46
	Mild AUD (2 or 3 symptoms)	61	(13.4%)	39	(14.4%)	12	(15.2%)	
	Moderate AUD (4 or 5 symptoms)	67	(14.8%)	39	(14.4%)	14	(17.7%)	
	Severe AUD (6+ symptoms)	272	(59.9%)	162	(59.8%)	40	(50.6%)	
AUD symptom counts (T1), M (SD)		6.25	(3.40)	6.24	(3.39)	5.53	(3.60)	p = .11
AUD diagnosed by healthcare provider in past two years		194	(42.7%)	104	(38.4%)	41	(51.9%)	**p = .04**

Note. P-values reflect differences between the primary care only and mental health only subgroups and were computed using chi-square tests. Patients in the primary care and mental health subsamples completed both Alcohol Symptom Checklists in a primary care or specialty mental health setting, respectively.

Table 2

Test-Retest Reliability Coefficients (and 95% CI's) for the Alcohol Symptom Checklist

	Test-retest reliability (95% CI)			
Full-scale measures	**Full sample (N = 454)**	**Primary care subsample (n = 271)**	**Mental health subsample (n = 79)**	**P-value of difference between subsamples**
Number of AUD criteria	0.79 (0.76, 0.82)	0.82 (0.77, 0.85)	0.74 (0.62, 0.83)	0.17
DSM-5 AUD severity	0.74 (0.69, 0.78)	0.75 (0.69, 0.80)	0.73 (0.61, 0.82)	0.74
Any AUD (mild, moderate, or severe vs. no AUD)	0.57 (0.46, 0.68)	0.58 (0.44, 0.72)	0.51 (0.29, 0.72)	0.55
AUD moderate or severe (vs. no AUD or mild AUD)	0.63 (0.54, 0.71)	0.62 (0.52, 0.73)	0.62 (0.44, 0.79)	0.97
Item-level responses				
1. Tolerance	0.55 (0.48, 0.63)	0.63 (0.54, 0.72)	0.46 (0.26, 0.65)	
2. Withdrawal	0.70 (0.64, 0.77)	0.76 (0.68, 0.84)	0.59 (0.41, 0.78)	
3. Larger/longer	0.55 (0.47, 0.63)	0.62 (0.52, 0.72)	0.47 (0.27, 0.66)	
4. Quit/control	0.58 (0.50, 0.66)	0.62 (0.52, 0.72)	0.52 (0.34, 0.70)	
5. Time spent	0.55 (0.47, 0.63)	0.62 (0.53, 0.72)	0.51 (0.31, 0.71)	
6. Physical/psychological problems	0.44 (0.35, 0.54)	0.42 (0.30, 0.55)	0.54 (0.35, 0.72)	
7. Neglect roles	0.54 (0.46, 0.62)	0.59 (0.50, 0.69)	0.45 (0.25, 0.65)	
8. Hazardous use	0.61 (0.54, 0.69)	0.63 (0.53, 0.73)	0.52 (0.31, 0.73)	
9. Social/interpersonal problems	0.57 (0.50, 0.65)	0.60 (0.50, 0.70)	0.57 (0.39, 0.75)	
10. Craving	0.61 (0.53, 0.68)	0.65 (0.56, 0.74)	0.60 (0.42, 0.77)	
11. Activities given up	0.60 (0.52, 0.67)	0.64 (0.55, 0.74)	0.36 (0.15, 0.57)	

Note. Test-retest reliability coefficients were estimated using a one-way single-measures agreement intraclass correlation coefficient (ICC) for number of AUD criteria, weighted kappa for DSM-5 AUD severity, and kappa for all binary indicators (any AUD, AUD moderate or severe, and item-level responses).

Table 3

Predictors of Discrepancies Between Alcohol Symptom Checklist Results at T1 and T2.

		Primary care subsample (n = 271)	Mental health subsample (n = 79)
Patient characteristics	**df**	**χ^2 (p-value)**	**χ^2 (p-value)**
Age	3	0.64 (0.89)	3.32 (0.35)
Sex	1	3.38 (0.07)	0.86 (0.35)
Medicaid	2	1.99 (0.37)	4.12 (0.13)
Medicare	2	1.59 (0.45)	5.58 (0.06)
Past two-year AUD diagnosis by healthcare provider	1	1.75 (0.19)	0.02 (0.88)

Note. For chi-square tests, discrepancies were coded as present if T1 checklist results were consistent with no or mild AUD (≤ 3 criteria) and T2 results were consistent with moderate or severe AUD (≥ 4 criteria) or vice versa, and were coded as absent if both T1 and T2 checklist results were consistent with no AUD or consistent with AUD.

Table 4

Changes in Criteria Counts Required for “Reliable Change” within Individual Patients

	Change in AUD criteria counts required to conclude reliable change within a single patient		
Sample	**80% confidence**	**90% confidence**	**95% confidence**
Full sample	± 3 criteria	± 4 criteria	± 5 criteria
Primary care subsample	± 3 criteria	± 4 criteria	± 4 criteria
Mental health subsample	± 4 criteria	± 5 criteria	± 6 criteria

Note. Values in the table indicate the amount of change in AUD criteria required to conclude that an individual patient has experienced “reliable change” between two measurement occasions – i.e., that a change in Alcohol Symptom Checklist scores between two time points was unlikely attributable to test-retest measurement error at the $p < .20$, $p < .10$, or $p < .05$ level.