Predicting Substance Use Behaviors: A Multi-Dataset, Mega-Analysis

Dahyeon Kang, Ph.D.[1], Catharine E., Fairbairn, Ph.D.[2], Lan Jiang., M.S.[3], Eddie Caumiant, M.S.[2], & Nigel Bosch, Ph.D.[3,4]

[1] University of Washington School of Medicine, Department of Psychiatry and Behavioral Sciences

[2] University of Illinois at Urbana-Champaign, Department of Psychology

[3] University of Illinois at Urbana-Champaign, School of Information Sciences

[4] University of Illinois at Urbana-Champaign, Department of Educational Psychology

Abstract

Predicting substance use behaviors is a critical public health challenge. While self-report survey data has been central to understanding substance use, its longitudinal predictive utility has yet to be quantified. We conducted a "mega"-analysis of 44,862 participants, integrating 197 harmonized self-report predictors across demographic, behavioral, familial, and psychosocial domains. Specifically, we employed both linear regression and machine learning approaches to evaluate the extent to which alcohol, tobacco, and cannabis use could be predicted. Models that handled high-dimensional, correlated predictor sets substantially outperformed unregularized linear regression, reflecting the known limitations of fitting unregularized regression to large, collinear predictor sets rather than a general advantage of machine learning. Random forest and regularized regression (Elastic Net) yielded comparable predictive performance ($R^2 \approx 14\%$ at baseline, $\approx 4\%$ for longitudinal change). In exploratory leave-one-dataset-out analyses, models did not generalize to independent held-out datasets ($R^2 \approx 0$), indicating that predictive associations were largely context-specific. Externalizing behaviors predicted a significant proportion of substance use variability, particularly at baseline, while the predictive power of demographic and familial factors became apparent in models predicting change over time. Combining large-scale harmonization with computational modeling clarifies both the potential and limits of forecasting substance use trajectories.

Keywords: mega-analysis, substance use disorder, individual differences, machine learning

General Scientific Summary

This study examined how well self-report survey data can predict substance use behaviors across seven large longitudinal datasets (N = 44,862). Behavioral, demographic, and family factors predicted within-dataset substance use at modest levels ($R^2 \approx 14\%$ cross-sectionally; $R^2 \approx 4\%$ longitudinally). However, predictive models did not generalize to independent held-out datasets ($R^2 \approx 0$ in leave-one-dataset-out analyses), indicating that predictive relationships were highly context-specific. Regularized linear regression and machine learning performed comparably. These findings clarify the current limits of predicting substance use from commonly available survey measures and highlight the substantial challenges that remain before such models could support individual-level risk identification.

INTRODUCTION

Substance use—particularly the heavy and frequent use of alcohol, tobacco, and cannabis—is a significant public health concern, contributing to substantial societal and economic burdens worldwide (Substance Abuse and Mental Health Services Administration, 2022; World Health Organization, 2024b). Sustained patterns of heavy substance use are associated with a range of negative outcomes, underscoring the importance of identifying early predictors of substance use behaviors that may persist or escalate over time. In recognition of this need, the U.S. National Institute on Drug Abuse (NIDA) has consistently prioritized "identifying the biological, environmental, behavioral, and social causes and consequences of drug use and addiction" (2000-2005 NIDA Strategic Plan; 2022-2026 NIDA Strategic Plan), while the World Health Organization (WHO) has called for efforts aimed at "tackling the determinants that drive the acceptability, availability, and affordability of alcohol consumption" (World Health Organization, 2024a). Yet, despite coordinated global efforts and decades of research, identifying consistent and generalizable predictors of substance use patterns remains challenging, due in part to the complex and multifactorial nature of substance use behaviors as well as, potentially, methodological factors that have heretofore restricted the use of available data.

Survey methods, particularly self-reported assessments, have long been the cornerstone of large-scale substance use research due to their accessibility, cost-effectiveness, and scalability (Coughlan et al., 2009). These tools have significantly advanced behavioral research by capturing a broad range of individual, social, and environmental factors across diverse populations (Armor et al., 1978; Schwarz, 1999; Delva et al., 2010; Tang et al., 2024). However, despite their utility, researchers have pointed to inconsistencies in results yielded by survey data in predicting substance use trajectories over time (Green et al., 2024; Tang et al., 2024). While

some studies report significant associations between baseline factors measured within survey assessments—such as gender, race, peer relationships, family history, and religiosity—and future substance use behaviors, others find minimal or no longitudinal relationships (Allen et al., 2012; Atherton et al., 2016; Chassin et al., 1999; Chen & Jacobson, 2012; Gau et al., 2007; Green et al., 2024; Morales et al., 2020; Orcutt & Schwabe, 2012). Researchers have attributed these inconsistencies to a variety of factors, including methodological issues such as sample homogeneity, limited dataset size or scope, and constraints of traditional analytic approaches—highlighting the potential value of more comprehensive modeling approaches capable of handling high-dimensional, correlated predictors in this area.

Survey Datasets and Substance Use Research

One major limitation in existing research stems from the widespread reliance on single datasets. While valuable, single-study analyses can introduce study-specific artifacts, systematic biases, and methodological inconsistencies that can limit the generalizability of findings (Day & Robles, 1989; Ferguson & Brannick, 2012; Higgins & Altman, 2008). As a result, predictors identified in one context often fail to replicate across diverse populations or settings. These concerns mirror broader challenges in psychosocial research, where variability in results has raised questions about the robustness and replicability of established predictors of substance use behaviors (Pearson et al., 2022). Further, another contributing factor may be the underutilization of available data. Many studies rely on datasets and/or analyses featuring a limited number of variables (Bidonde et al., 2023; Kraemer & Blasey, 2004; McLellan et al., 2000; Podsakoff et al., 2003), limiting their capacity to capture the multifaceted nature of substance use etiology spanning behavioral, psychological, environmental, and social domains. For example, efforts to explain heterogeneity in substance use trajectories often rely on a single predictor or a

circumscribed set of variables (e.g., 5–10)—a level of parsimony that aligns with mechanistic hypothesis testing but often overlooks the complex, multifactorial nature of substance use development and the substantial covariation among contributing factors.

Addressing these challenges requires not only expanding the range of predictors considered but also integrating data across multiple studies through cross-study harmonization efforts, enabling more generalizable and reliable conclusions. However, prior harmonization work has often focused on substance use *problem* indexes, which vary substantially in how "problems" are defined and assessed (Hasin et al., 2013; Saunders et al., 1993; Tang et al., 2024). This inconsistency poses particular limitations for studying adolescents and young adults, who may engage in elevated patterns of use without meeting clinical thresholds. To capture meaningful patterns of heightened use and improve the validity of cross-study analyses, it is critical to prioritize well-defined, consistently measured self-report indices of substance use (e.g., use frequency and quantity) in harmonization efforts.

Data Analysis and Substance Use Prediction

Traditional statistical approaches, such as linear regression, have been widely used in substance use research, offering valuable insights into predictors of use behaviors. However, most of these models assume linear relationships between variables and often struggle to account for the complex, non-linear and interactive nature of substance use determinants (Babyak, 2004; Cohen et al., 2003; Cox, 1972). Additionally, challenges such as multicollinearity, where highly correlated predictors obscure their unique contributions, can limit the reliability of these models (Daoud, 2017; J. H. Kim, 2019; P. Vatcheva & Lee, 2016). As such, more flexible and adaptive analytical techniques are likely to be useful in capturing the complexity of substance use behaviors and to improve predictive validity. Such models may have particular utility in

adolescent and young adult populations, where rapid developmental changes and shifting social contexts can both amplify risk and present unique windows for prevention (Chassin et al., 1999; Patrick & Terry-McElrath, 2017; Windle & Windle, 2012).

Importantly, unlike single-cause medical conditions, substance use emerges from a dynamic interplay of biological, psychological, social, and environmental factors that influence patterns of use and potential escalation to heavy or problematic consumption (e.g., Ray et al., 2010, Volkow et al., 2016, Epstein et al., 2010, Koob, 2015, Sayette et al., 2015, Badiani, 2013; Clapp & Shillington, 2001; Creswell, 2021; Fairbairn & Sayette, 2014). This multifactorial etiology underscores the heterogeneity of substance use behaviors, where no single determinant can fully account for their onset or progression (Litten et al., 2015; Witkiewitz et al., 2019, 2023). In fact, previous research has identified a wide range of individual and contextual factors associated with substance use risk (Bellis et al., 2007; Chassin et al., 1999; Clapp & Shillington, 2001; Gau et al., 2007; Morales et al., 2020). Demographic variables such as age, race/ethnicity, and gender have been consistently linked to substance use patterns, with distinct risk profiles emerging across groups—for instance, young adults and specific minority group dimensions may be at greater risk for heavy drinking or substance use (e.g., Bennett et al., 1999; Chen & Jacobson, 2012; Karriker-Jaffe et al., 2012; Orcutt & Schwabe, 2012; Patrick & Terry-McElrath, 2017). Some studies also suggest that interpersonal factors (e.g., relationship quality, perceived social support) play a critical role, as social connections and group dynamics can either amplify or mitigate substance use behaviors (Borsari et al., 2006; Caspers et al., 2005; Fairbairn et al., 2018; Gliksberg et al., 2022). Family history—e.g., marital status, divorce history, number of children—have also been found to influence an individual's likelihood of engaging in higher levels of substance use (Chassin et al., 1994, 1999; Vungkhanching et al., 2004). Further,

personality traits and behavioral tendencies, such as risk-taking behaviors or aggression (e.g., vandalism, risky sexual activity), have been linked with future substance use behaviors (Quinn et al., 2011; Quinn & Harden, 2012).

Notably, although a large body of research discussed above points to shared pathways of risk underlying alcohol, tobacco, and cannabis use (Gau et al., 2007; Sinha, 2008; Wills & Cleary, 1996), evidence also suggests that some mechanisms may differ meaningfully across substances. Peer influences and social motives, for instance, appear more predictive of alcohol and cannabis use among young adults (Blumenthal et al., 2010; Borsari et al., 2006; Gliksberg et al., 2022; Phillips et al., 2018), whereas stress-related coping and habitual tendencies are more uniquely associated with tobacco use (Chessick, 1964; Khorrami et al., 2021; Torres & O'Dell, 2016). Similarly, racial and ethnic differences appear substance-specific: Black individuals tend to report lower alcohol consumption but higher cannabis use (e.g., Pacek et al., 2012; Zapolski et al., 2014), and Asian populations report higher rates of cigarette smoking, particularly among men (e.g., S. Kim et al., 2007; Shih et al., 2015; Thai et al., 2010). Personality traits also show substance-specific patterns, with risk-taking and sensation-seeking tendencies more predictive of alcohol and cannabis use (e.g., Ansell et al., 2015; MacPherson et al., 2010; Quinn & Harden, 2012), while neuroticism and anxiety-related traits have been linked to tobacco use (Becoña et al., 1998; Malouff et al., 2006; Zvolensky et al., 2015).

Most importantly, while these abovementioned factors have individually been shown to predict substance use behaviors, their combined effects and the complex interplay between them have not been fully explored. To date, only one study (Green et al., 2024) has attempted to comprehensively examine a broad range of variables to predict substance use initiation among teenagers, utilizing penalized logistic linear regression on a sizable sample (N=6,829). This study

found that sociodemographic factors were the most robust predictors of substance use, alongside meaningful contributions from cultural, environmental, and physical health factors. While these findings offer promising insights, they also underscore the need for analytical approaches capable of handling high-dimensional, correlated predictor sets. Traditional unregularized linear models often perform poorly in such settings due to multicollinearity and overfitting. More flexible methods—including both regularized regression (e.g., Elastic Net) and machine learning approaches (e.g., random forests)—may better accommodate the large, interrelated predictor sets relevant to substance use research.

The Current Study: Bridging Predictive Tools and Theoretical Frameworks

In the present research we revisit with new tools the enduring question: *to what extent can substance use be predicted*? Specifically, using a large-scale cross-harmonized longitudinal dataset featuring hundreds of predictors paired with machine learning analytic models, we explored the extent to which it is possible to predict patterns of alcohol, cannabis, and tobacco usage using the standard (survey) measures available to substance use researchers. The current study focuses specifically on alcohol, tobacco, and cannabis use in adolescence and young adulthood—a period of particular vulnerability and dynamic change, during which substance use behaviors are most likely to develop and solidify (Johnston et al., 2023; McGee et al., 2000; Schulenberg & Maggs, 2002; Windle & Windle, 2012). Further, alcohol, tobacco, and cannabis represent the most commonly used substances among young adults globally (World Health Organization, 2024b) and account for the majority of early substance use initiation and sustained use patterns during this critical developmental period (Johnston et al., 2023). By combining data across diverse populations, contexts, and time points, this study moves beyond the limitations of

single-study analyses, offering a more robust evaluation of predictive patterns across diverse cohorts, while also allowing assessment of cross-dataset generalizability.

The current research employs both regression models and machine learning techniques to evaluate predictors of alcohol, tobacco, and cannabis use behaviors. Our analysis focuses on four critical objectives. First, we calculate the proportion of variance in substance use at baseline (T1), defined as the earliest available assessment within each dataset, as well as the proportion of variance in changes in substance use from baseline to follow-up (T2), defined as the latest available subsequent assessment. Second, we compare the performance of machine learning and regression-based models. Third, we identify which individual predictors most effectively explain substance use behaviors, both cross-sectionally and longitudinally. Fourth, given variability in participant age ranges, follow-up intervals, and study designs across datasets, we assess the generalizability of predictive findings across cohorts using a cross-dataset harmonization framework.

RESULTS

Description of Datasets

A total of seven datasets[1], comprising 44,862 individual participants, were included in the current analyses (see Figure 1 Note for dataset titles and references). Across these datasets, a total of 41,629 variables underwent initial review for inclusion, yielding a total harmonized sample of 197 baseline predictors that were identified as being assessed in at least two datasets and were therefore included in models. The majority of these predictors (66.50%) were assessed

[1] The seven datasets included in the current analyses were: National Education Longitudinal Study (ICPSR 9389), National Youth Smoking Cessation Survey (ICPSR 34275), National Longitudinal Study of Adolescent to Adult Health (ICPSR 21600), Flint Michigan Adolescent Study (ICPSR 34598), Drug Use Trajectories (ICPSR 30862), High School and Beyond (ICPSR 7896), and National Longitudinal Survey of Youth (ICPSR 34923).

in two datasets; 23.35% appeared in three datasets; 6.60% in four datasets; 2.03% in five datasets; 0.51% in six datasets; and 1.02% were assessed across all seven datasets. Predictors included in analyses spanned a broad range of domains relevant to substance use risk, including demographics (e.g., gender, age, race/ethnicity), family demographics and living situation (e.g., parental education, marital status, household structure), emotional and psychosocial functioning (e.g., self-esteem, emotional well-being), sexual behavior and health (e.g., age of sexual debut, number of sexual partners), and indicators of behavioral disinhibition (e.g., history of theft or physical aggression). Additional domains included civic engagement, religious involvement, leisure activities, and peer relationships. See Supplementary Table 1 for full list of variables included in analyses.

The average age of participants at sampling initiation was 18.11, with a range from 14.6 to 20.3 years. The average length of longitudinal follow-up was 6.43 years, ranging from 2 to 18 years. The average age of participants at follow-up was 24.92. Regarding outcomes, the following substance use outcome variables were examined: (a) quantity of alcohol use (k=5); (b) frequency of alcohol use (k=7); (c) frequency of binge drinking (k=6); (d) frequency of cigarette use (k=3); (e) quantity of cigarette use (k=4); (f) frequency of marijuana use (k=3). Analytic and demographic features of each dataset are reported in Table 1.

Percentage of Variance Explained by Predictive Models

Random Forest Models. To examine the extent to which self-report predictors could explain substance use behavior, we applied a random forest model to each dataset using all 197 harmonized baseline predictors. Across datasets, the 197 predictors accounted for 14% of the variance in substance use at baseline ($r = .376$, $SE = .066$, $t = 5.668$, $p < .001$, $R^2 = .141$). Results were generally consistent across substances: alcohol ($r = .380$, $SE = .077$, $t = 4.958$, $p < 0.001$),

cigarettes (r = .427, SE = .056, t = 7.620, p < .001), and cannabis (r = .424, SE = .021, t = 20.196, p = .002), $F(2, 23)$ = .474, p = .628 (see Table 2). The random forest model also performed well in predicting changes in substance use from T1 to follow-up (T2), r = .188, SE = .038, t = 5.005, p < .001, R^2 = .035, explaining approximately 4% of the variance in longitudinal change. Results were generally consistent across substance types, including alcohol (r = .201, SE = .052, t = 3.867, p = .002), cannabis (r = .189, SE = .028, t = 6.741, p = .021), and cigarettes (r = .181, SE = .037, t = 4.942, p = .003). We also examined results by length of longitudinal follow-up, categorized as short-term (0–2 years), mid-term (3–4 years), and long-term (5+ years). Some differences in effect sizes emerged across time spans for longitudinal assessment, although these differences were non-significant, $F(1, 23)$ = 2.315, p = .142: follow-up intervals of 0–2 years (r = .155, SE = .050, t = 3.107, p = .011), 3–4 years (r = .106, SE = .098, t = 1.087, p = .391), and 5+ years (r = .273, SE = .049, t = 5.558, p < .001). Finally, results of models in which each dataset is excluded in turn, examining the broad robustness of findings to dataset-level exclusions, are presented in Table 2.

Linear Regression Models. Regarding the performance of models employing more traditional analytic approaches, the linear regression model also explained a significant proportion (7%) of the variance in substance use at baseline (r = .262, SE = .070, t = 3.752, p < .001, R^2=.069). However, the initial model exhibited substantial multicollinearity, with variance inflation factor (VIF) values exceeding the commonly accepted threshold of 5 (Akinwande et al., 2015). To address this, variables with VIF values greater than 5 were removed, yielding a refined model with improved performance, r = .345, SE = .060, t = 5.773, p < .001, R^2=.119. Thus, although the random forest model accounted for a slightly larger proportion of variance (R^2=.141), the difference between the two models was reduced after

addressing multicollinearity. For changes in substance use over time, the initial linear regression model was less effective, explaining less than 1% (vs. 4% for random forest) of over-time progression in substance use, $r = .081$, $SE = .048$, $t = 1.676$, $p = .106$, $R^2 = .007$; however, the adjusted linear regression model demonstrated substantial improvement, with 3% of variance explained ($r = .170$, $SE = .034$, $t = 5.041$, $p < .001$, $R^2 = .029$) —albeit in the context of a model from which approximately 36% of predictors were removed to address multicollinearity.

Elastic Net Models. For Elastic Net models, results indicated that predictors accounted for 13% of the variance in substance use at baseline ($r = .366$, $SE = .070$, $t = 5.196$, $p < .001$, $R^2 = .134$), closely approximating the performance of the random forest model. For longitudinal change, Elastic Net models explained 3% of the variance in substance use ($r = .178$, $SE = .039$, $t = 4.568$, $p < .001$, $R^2 = .032$), again demonstrating modest but statistically significant predictive performance.

Key Individual Predictors of Substance Use

Given that random forest and Elastic Net models demonstrated comparable predictive performance, and given that random forest SHAP values provide a well-validated framework for quantifying individual predictor contributions, individual predictor analyses focus on the results of random forest models. To quantify the contribution of each predictor to model predictions, we utilized SHapley Additive exPlanations (SHAP) values, a widely used approach in machine learning that attributes importance scores to individual features based on their marginal contributions to the prediction (Lundberg & Lee, 2017). Higher SHAP values indicate greater influence of a predictor on the model's output. To summarize the importance of each variable across participants, we computed the mean absolute SHAP value for each predictor, allowing us to rank predictors based on their overall contribution to explaining variance in substance use

outcomes. Figure 2 visualizes the top 10 predictors of substance use behaviors, separately for alcohol, cigarette, and cannabis outcomes. For baseline (T1) substance use, across substances, involvement in a physical fight emerged as the most influential predictor (M_{SHAP} = .148), accounting for approximately 10% of the total predictor importance across the 197 variables included. Other highly influential predictors included number of sexual partners (M_{SHAP} = .058, 4.08%), mother's education level (M_{SHAP} = .058, 4.06%), and age (M_{SHAP} = .047, 3.31%). Substance-specific analyses showed that involvement in a physical fight (M_{SHAP} = .344, 27.32%), number of sexual partners (M_{SHAP} = .082, 6.47%), and age (M_{SHAP} = .054, 4.28%) were the most important predictors for alcohol use; mother's education level (M_{SHAP} = .151, 9.64%), involvement in a physical fight (M_{SHAP} = .069, 4.43%), and experience of sexual intercourse (M_{SHAP} = .062, 3.99%) for cigarette use; and age (M_{SHAP} = .068, 9.07%), gender (M_{SHAP} = .052, 6.86%), and experience of sexual intercourse (M_{SHAP} = .048, 6.42%) for cannabis use.

For changes in substance use from T1 to T2, random forest models identified age (M_{SHAP} = .097, 8.91%), involvement in a physical fight (M_{SHAP} = .075, 6.90%), gender (M_{SHAP} = .056, 5.09%), and number of sexual partners (M_{SHAP} = .051, 4.65%) as the highest contributors. Substance-specific analyses showed that involvement in a physical fight (M_{SHAP} = .216, 13.72%), age (M_{SHAP} = .084, 5.32%), and family relationship security (M_{SHAP} = .078, 4.94%) were key predictors for changes in alcohol use; age (M_{SHAP} = .153, 16.64%), gender (M_{SHAP} = .052, 5.66%), and number of sexual partners (M_{SHAP} = .037, 4.05%) for changes in cigarette use; and number of sexual partners (M_{SHAP} = .085, 14.12%), age (M_{SHAP} = .056, 9.22%), and gender (M_{SHAP} = .046, 7.57%) for changes in cannabis use.

Dataset-Level Moderators of Prediction Strength

Meta-regression analyses examined whether dataset-level characteristics accounted for variability in prediction strength across samples (see Table 3). Dataset significantly moderated prediction strength at baseline, $F(6,71) = 3.86$, $p = .002$, indicating variability across cohorts. Relative to the reference dataset (AddHealth), prediction performance was significantly lower in the DUT ($\beta = -.269$, $p = .014$), Flint ($\beta = -.454$, $p < .001$), and NYSCS ($\beta = -.271$, $p = .013$) datasets, whereas other datasets did not significantly differ. In contrast, dataset did not significantly moderate longitudinal prediction, $F(6,70) = 2.18$, $p = .055$, although the Flint dataset again showed lower prediction performance, $\beta = -.290$, $p = .003$. Among continuous moderators, follow-up length significantly moderated longitudinal prediction, $F(1,75) = 4.46$, $p = .038$, such that longer follow-up intervals were associated with modest increases in predictive performance, $\beta = .009$. Baseline sample size was also a significant moderator of baseline prediction, $F(1,76) = 6.05$, $p = .016$, with larger samples associated with slightly higher predictive performance ($\beta \approx .000$, $p = .016$). In contrast, mean sample age (baseline: $F(1,76) = .29$, p = .589; change: $F(1,75) = 1.07$, $p = .305$) and year of data collection (baseline: $F(1,76) = 1.05$, p = .309; change: $F(1,75) = .98$, $p = .325$) did not significantly moderate prediction strength. Similarly, sample size did not significantly moderate longitudinal prediction, $F(1,75) = .96$, $p = .330$.

In exploratory analyses examining cross-dataset generalizability, we implemented a leave-one-dataset-out (LODO) validation approach in which models were trained on six datasets and evaluated on the remaining held-out dataset. Across these analyses, predictive performance in the held-out datasets was consistently near chance, with R^2 values approximating zero (range: -.01 to .02), suggesting that predictive relationships identified within individual datasets did not generalize well to independent cohorts.

DISCUSSION

This study sought to address longstanding questions about the predictive utility of survey methods in substance use research by leveraging a "mega-analysis" of seven publicly available longitudinal datasets. Across data encompassing 44,862 participants and 197 variables, we found that approximately 14% of the variance in substance use at baseline and 4% of the variance in longitudinal change were explained by survey-based predictors. These findings contribute to an ongoing conversation about the extent to which self-report measures can predict substance use outcomes. While the cross-sectional prediction rate reflects a meaningful ability to capture current substance use behaviors, the substantially lower variance explained in models of change underscores the challenges of forecasting such a multifaceted behavior over time.

The decline in predictive power when moving from cross-sectional prediction at baseline to longitudinal prediction of change highlights the inherent complexity of substance use behavior, consistent with prior research suggesting that forecasting future patterns of use is particularly challenging across analytic methods (e.g., Janssen et al., 2015; Kulkarni et al., 2023; Shenoi et al., 2019; Yip et al., 2020). This discrepancy suggests that the limitation may not lie with self-report assessments themselves but rather with the broader difficulty of predicting change in behavior over time—a challenge compounded by both substantive and methodological factors. Although self-report measures can effectively capture static risk factors at a given moment, their predictive utility may weaken for trajectories of change, in part because behaviors evolve dynamically and in part because change scores introduce additional noise. Specifically, predicting change tends to be more difficult than predicting baseline levels, as stable individual differences (e.g., personality traits) often explain less variance in slopes than in intercepts. Moreover, because change scores are derived from two time points, they amplify measurement

error—especially in self-report data—where errors, even when partially correlated within-person, can substantially inflate overall variance (Brooks-Russell et al., 2015; Harris et al., 2008). Further, longitudinal prediction from surveys may be weakened by a reduction in common method biases—such as affirmation bias or mood-linked reporting—that artificially enhance cross-sectional associations but attenuate over time (Podsakoff et al., 2003, 2024).

Although survey-based predictors explained approximately 4% of variance in longitudinal substance use trajectories, it is important to contextualize this modest figure. Even with nearly 200 harmonized predictors across 44,000+ participants, the majority of variance in substance use change remains unexplained. Prior work on psychosocial predictors has typically yielded comparable or lower predictive power (Atherton et al., 2016; Brumback et al., 2021; Green et al., 2024; Hussong, 2003), particularly when using conventional linear models and single-sample designs. By contrast, our approach—combining large-scale cross-study harmonization with systematic predictive modeling—may offer a more robust estimate of this modest predictive signal. Future work could investigate whether more objective, behaviorally anchored predictors (e.g., performance-based measures, ecological assessments) might enhance longitudinal forecasting, given their potential to capture enduring risk factors less susceptible to self-report biases.

Although predictive modeling of substance use behavior presents clear challenges, our results illustrate some patterns worth noting. Linear regression models tended to yield low effect sizes, particularly when multiple predictors were included simultaneously. Random forest models produced effect sizes that generally fell in the small to moderate range. Additional analyses using Elastic Net indicated that predictive performance was comparable across regularized linear and tree-based approaches, suggesting that gains were attributable to the

ability to handle high-dimensional, correlated predictors rather than to a specific algorithm (Dwyer et al., 2018). These findings are consistent with prior work suggesting that, when appropriately specified, regularized regression models often perform similarly to more complex machine learning approaches.

Across both baseline and follow-up, we observed that demographic and relatively stable familial factors tended to emerge more prominently as predictors in longitudinal models compared to baseline models. One potential explanation is that these variables may represent more "enduring" markers of risk, less susceptible to change over time, in contrast to more dynamic behavioral indicators such as levels of sexual activity or externalizing behaviors. Another possible explanation is methodological: common methods bias, particularly for socially sensitive or stigmatized behaviors, may inflate associations in cross-sectional models where both predictors and outcomes are reported simultaneously. Over time, as measurement waves are separated, these common methods effects may diminish, allowing more stable, less bias-prone factors—such as demographics or family background—to account for a larger proportion of variance (Podsakoff et al., 2024). In this regard, some of the observed shifts in predictor prominence from baseline to follow-up may reflect not only changes in the salience of risk factors but also the influence of common methods variance.

Our findings highlight the enduring influence of both demographic and behavioral factors on substance use risk. Familial and socioeconomic variables—including marital status, father's employment, and mother's education—were meaningfully associated with substance use behaviors, consistent with ecological and developmental frameworks emphasizing the role of family structure and socioeconomic context (Hawkins, Catalano, & Miller, 1992; Zapolski et al., 2014). Married individuals, for instance, exhibited lower rates of substance use compared to

those who were single, divorced, or widowed, a protective effect often attributed to increased social support, emotional stability, and family responsibilities (Leonard & Eiden, 2007; Dawson et al., 2006), while marital disruption has been linked to elevated substance use risk through mechanisms such as heightened stress and weakened social networks (Kendler et al., 2016). Similarly, higher parental education was associated with lower substance use risk, which may reflect broader structural differences in access to economic, educational, and neighborhood resources, as well as differences in time and opportunity constraints faced by families (Ensminger et al., 1996; Patrick et al., 2012). Behavioral and psychosocial predictors also played a prominent role, particularly externalizing behaviors such as physical aggression, theft, gang-related violence, and risky sexual activity. These patterns are consistent with prior research suggesting that externalizing tendencies reflect underlying traits like impulsivity, sensation-seeking, and behavioral disinhibition that predispose individuals to substance use initiation and escalation (Zuckerman & Kuhlman, 2000; Dishion & Patterson, 2006). Importantly, the relative prominence of predictors shifted over time: dynamic externalizing behaviors were particularly influential in predicting baseline substance use, whereas more stable demographic and familial factors emerged as stronger predictors of longitudinal outcomes. This pattern underscores the complexity of substance use trajectories, suggesting that different classes of risk factors may dominate at different developmental stages, and highlights the need for temporally sensitive models of substance use risk.

An important consideration in interpreting the present findings is the role of psychopathology, particularly internalizing and externalizing dimensions, in predicting substance use. Although a range of items capturing emotional distress and behavioral dysregulation were included, formal diagnostic measures and harmonized multi-item psychopathology scales were

largely unavailable across datasets. As a result, internalizing symptoms were represented primarily through individual items (e.g., feeling fearful, difficulty concentrating, sleep disturbance), which may not fully capture the underlying latent constructs. Notably, internalizing indicators were less consistently identified among the strongest predictors, although some emerged as top predictors for specific outcomes (e.g., depressed mood and attentional difficulties for baseline cannabis use). In contrast, externalizing-related behaviors—such as aggression, delinquency, and risk-taking—were more consistently represented across datasets and more robustly predicted substance use outcomes. This pattern should not be interpreted as evidence that internalizing processes are less important; rather, it likely reflects differences in measurement consistency and construct representation across datasets. More broadly, many of the behavioral predictors identified here may reflect an underlying externalizing liability, consistent with transdiagnostic models of psychopathology (Dalgleish et al., 2020; Kotov et al., 2017). Future research using harmonized, multi-item and multi-informant assessments will be critical for more precisely disentangling the contributions of internalizing and externalizing dimensions to substance use trajectories.

Overall, these findings highlight the enduring influence of individual behaviors, social environments, and demographic characteristics on substance use risk. Consistent with previous literature and ecological and developmental theories (Bronfenbrenner, 1979; Masten et al., 2005), findings underscore the interplay between personal, familial, and societal factors in shaping substance use trajectories. These predictor patterns are consistent with prior theoretical models and may inform the design of prevention research, though it is critical to note that the modest proportion of variance explained—particularly longitudinally ($R^2 \approx 4\%$)—and the failure of models to generalize across datasets ($R^2 \approx 0$ in LODO analyses) means that the present results

do not establish a basis for individual-level risk screening or targeted prevention in new contexts. Rather, these findings identify which broad domains of risk replicate most consistently across cohorts, providing a foundation for future research with stronger measurement and prospective multi-site designs.

These findings also raise important questions about what may be missing from current approaches to predicting substance use. Many of the most predictive factors—such as externalizing behaviors, demographic characteristics, and lifestyle indicators—are consistent with prior research and therefore not unexpected. Rather than identifying entirely novel predictors, the present study advances the field by demonstrating, within a large-scale harmonized multi-dataset framework, which predictors replicate most consistently across cohorts. Notably, even with nearly 200 predictors, the overall proportion of variance explained—particularly longitudinally—remained modest. This suggests that improving prediction may require not simply expanding the number of baseline variables, but more precisely capturing underlying mechanisms, including proximal and dynamic processes (e.g., momentary affect, contextual cues, and social interactions), as well as the timing and within-person variability of these experiences. Such processes are not well captured by one-time self-report assessments and may be critical for understanding individual trajectories of substance use.

The leave-one-dataset-out (LODO) analyses provide an important contribution of the present study by directly testing whether prediction models developed in one set of large longitudinal datasets generalize to independent cohorts. A central promise of large-scale behavioral data science is that bigger samples, broader harmonization, and flexible computational models will yield more generalizable prediction (Joel et al., 2020; Salganik et al., 2020; Yarkoni & Westfall, 2017). Yet, even after integrating seven large longitudinal datasets,

harmonizing nearly 200 predictors, and applying cross-validated machine learning methods, predictive performance did not transport well across cohorts. Models that achieved modest prediction within datasets performed near chance when applied to independent held-out datasets. This finding suggests that the field's challenge is not simply one of scale. Larger samples are necessary, but they are not sufficient if the predictors being harmonized differ in wording, timing, developmental context, construct coverage, or proximity to the mechanisms driving substance use. In this sense, the poor cross-dataset transportability observed here is not merely a limitation of the present study; it is a substantive empirical result that reveals a deeper challenge for behavioral prediction research. The path toward generalizable prediction will likely require not only larger and more diverse samples, but also greater mechanistic precision in measurement, stronger alignment of constructs across studies, and prospective coordination of measures across cohorts. Thus, the present findings offer a cautionary but important conclusion: without greater measurement and theoretical precision, even very large harmonized datasets may produce models that predict within context but fail to generalize beyond it.

While our study offers several strengths, including the use of a large, diverse dataset and systematic evaluation of both regularized regression and machine learning approaches, it is important to acknowledge its limitations. A key challenge with random forest models is their limited interpretability; while they effectively identify key predictors, they do not readily provide insight into directionality or underlying causal mechanisms. Additionally, although our mega-analysis integrated seven large-scale longitudinal datasets, there was considerable variability in sample composition, measurement, and study design across studies, which could have contributed to inconsistencies in findings. Future studies should therefore prioritize collaboration

across research groups to promote consistency in measurement and sampling approaches, enabling more reliable cross-study comparisons and stronger cumulative knowledge in the field.

An additional limitation of the present study is the reliance on participant self-report survey measures. This focus was consistent with the preregistered aims of the study, which were to evaluate the predictability of later substance use from commonly used participant-report measures across longitudinal datasets. However, reliance on a single reporter raises the possibility of common-method or common-rater bias, particularly in models involving associations among risk-related behaviors reported by the same individual (Podsakoff et al., 2003, 2024). This concern is likely most pronounced for cross-sectional associations, whereas the primary analyses in the present study focused on longitudinal prediction across temporally separated assessment waves. Nonetheless, temporal separation does not fully eliminate common-method concerns. Informant reports, including parent and teacher reports, can provide valuable complementary perspectives and may demonstrate comparable or stronger validity for some constructs, particularly externalizing behavior. In the present study, however, informant-report measures were not consistently available or harmonizable across datasets after applying the same inclusion criteria used for participant-report variables. Future multi-site studies designed with harmonized multi-informant assessments would provide an important opportunity to evaluate whether predictive performance differs across participant, parent, teacher, and other informant reports.

Although the present study harmonized a large number of predictors (~200 variables) across multiple large-scale datasets, the set of available predictors was ultimately constrained by the measures that were consistently assessed across studies. Some theoretically important variables—such as family history of substance use or more comprehensive assessments of

psychiatric diagnoses—were not available in a harmonized form across datasets and therefore could not be included. As such, the present findings should be interpreted within the context of the available measures, and it is possible that inclusion of additional risk factors not captured here could improve predictive performance. Further, longitudinal change in substance use levels was operationalized using simple difference scores. Although this approach provided a transparent and harmonizable measure of change across datasets, difference scores can be vulnerable to measurement error more than latent change score or growth modeling approaches (Cronbach & Furby, 1970; Rogosa & Willett, 1983). Future research using more harmonized repeated-measures designs could apply latent change or growth-based approaches to better separate true intraindividual change from measurement error and baseline status.

Another limitation relates to the age[2] range of participants, as the datasets primarily focused on adolescents and young adults. While this focus allowed us to target a developmental period of heightened risk for substance use initiation and escalation, it limits the generalizability of our findings to older populations or other developmental stages. Further, although our data suggest that the present data are unlikely to have been meaningfully affected by attrition bias (see Supplementary Table 3), concerns about potential attrition should nevertheless be noted, as longitudinal studies may be subject to selective loss of participants over time. Finally, given the relatively small number of datasets included, we had limited statistical power to detect effects driven by dataset-level characteristics—such as the time span of longitudinal follow-up—which

[2] Sensitivity analysis excluding age-related predictors (e.g., current age, age at first sexual activity, age at first marriage) yielded results comparable to the primary analyses (baseline: $\beta = .370$, $SE = .067$, $t = 5.55$, $p < .001$, *95% CI* [0.233, 0.507]; change: $\beta = .169$, $SE = .042$, $t = 4.04$, $p < .001$).

could influence the magnitude of predictive relationships. Future work incorporating larger cross-dataset samples will be critical for unpacking such study-level sources of heterogeneity.

We also acknowledge the limitations associated with our literature search and dataset inclusion cutoff date (2020). This choice reflected the substantial labor required for large-scale data harmonization and variable coding across datasets—work that spanned multiple years and over 10,000 individual variables scanned and reviewed. In this sense, the project represents more of a long-term empirical effort than a traditional meta-analysis, with decisions made to ensure feasibility and methodological rigor. Future work could build upon this foundation by incorporating newly available datasets and extending these efforts across broader age ranges and more recent cohorts. Finally, the modest proportion of variance explained in our models suggests that survey-based predictors alone are insufficient to fully capture the complexity of substance use behaviors. Future research could enhance predictive accuracy by incorporating diverse data modalities, including neurobiological, genetic, and ecological momentary assessments, which may provide deeper insight into the mechanisms driving substance use (Volkow et al., 2016; Sinha, 2008).

METHOD

Data Selection Procedures

In order to search for relevant samples, datasets listed in the Alcohol Epidemiological Data Directory (AEDD) were scanned. Further, three additional databases, Inter-university Consortium for Political and Social Research (ICPSR), the database of Genotypes and Phenotypes (dbGaP), and Harvard Dataverse, were searched according to the following parameters; ICPSR: *“alcohol” or “alcohol abuse” or “addiction” or “alcohol consumption” or “cocaine” or “cigarettes” or “drug abuse” or “drug courts” or “drug dependence” or “drug*

law offenses" or "drug offenders" or "drug overdose" or "drug treatment" or "drug use" or "smoking" or "smoking cessation" or "substance abuse" or "substance abuse treatment" or "substance use" or "tobacco use". Additional filters: quantitative data type and longitudinal; DbGaP: *("alcohol" or "drug" or "substance" or "addiction") and ("longitudinal" or "prospective")*; Harvard Dataverse: *ti(("alcohol" or "drug" or "substance") and ("longitudinal" or "prospective")). Additional filters: social sciences, medical/health, life sciences.*

A total of 220 dataset abstracts were reviewed for potential inclusion in the current analyses (see Figure 1, PRISMA Flow Diagram). Datasets were included based on the following criteria: (1) The dataset was publicly available as of the date of data harmonization launch; (2) The dataset directly assessed substance use, including alcohol, nicotine/tobacco, or marijuana/cannabis. Specific measures included alcohol use frequency and quantity, binge drinking quantity, nicotine/tobacco use frequency and quantity, and marijuana/cannabis use frequency; (3) The dataset was longitudinal or prospective in nature, requiring at least one follow-up time point wherein the same substance use measure was collected as at baseline. For each dataset, baseline (T1) was defined as the earliest available assessment wave that included the relevant substance use measures, and follow-up (T2) was defined as the latest available subsequent wave in which the same measures were assessed. For example, in datasets with multiple waves (e.g., Waves 1-4), Wave 1 was treated as baseline and the final available wave (e.g., Wave 4) was used as the follow-up. This approach maximized the longitudinal interval for prediction while ensuring consistency in outcome measurement within each dataset. As a result, follow-up intervals varied across datasets due to differences in study design and assessment schedules. No age restrictions were applied to participants. Reasons for dataset exclusion at each stage of the selection process are detailed in Figure 1. The current research did not involve the

collection of direct information from human subjects—only the aggregation of de-identified material in publicly available datasets—and therefore was exempt from IRB approval. The overall design and analysis plan for the project was pre-registered (https://osf.io/g4w9z/).

Measures

Individual Variable Coding

Predictors of Substance Use. In total, 41,629 variables were reviewed and categorized for the purposes of cross-dataset harmonization. Predictor coding proceeded in two stages: 1-initial predictor identification; 2-predictor classification and harmonization. Regarding the former of these processes, to identify predictors of substance use, we manually reviewed all baseline variables for each dataset. Predictors were required to be assessed in at least two datasets in order to be included in the analyses for cross-dataset synthesis. Informant reports (e.g., parent or teacher assessments) were excluded due to inconsistent availability across datasets and to ensure a focus on predictors derived from self-report measures commonly used in large-scale survey research. Items assessing substance-related problems (e.g., SUD symptoms) were excluded because the assessment of substance-related problems varied substantially across datasets, thereby limiting the ability to synthesize the measures between samples.

Regarding the data harmonization stage, we developed a codebook to standardize retained variables across datasets. Variations in equivalent measures across datasets were standardized using a unified coding scheme. To provide just one example, when response options for a variable like "employment status" differed across datasets—such as (1) yes (employed), (2) no (unemployed), (3) prefer not to answer in one dataset and (1) currently working, (2) never worked, (3) not currently employed but was, (4) other in another—we harmonized these responses into a common coding scheme: (1) yes (employed), (2) no

(unemployed), (3) other. Original variables were then recoded to align with this unified framework. Initial predictor identification and harmonization procedures were led by the first (D.K.) and fourth (E.C.) authors, with coding and recoding carried out by E.C. and a team of undergraduate research assistants. Each variable was independently reviewed and categorized by at least two coders to ensure accuracy (Cohen's $\kappa = .91$ for initial categorization; $\kappa = .89$ for final harmonization). Discrepancies were reviewed and resolved through discussion among the coding team. No variables were excluded based on inter-rater reliability or quality control metrics. Variables were retained if they could be successfully harmonized across datasets using predefined coding procedures. The final unified codebook, along with detailed coding protocols, is provided in Supplementary Table 1.

Measures of Substance Use. The dependent variable in this study was substance use at both baseline and follow-up time points. For datasets that included multiple follow-up time points, the latest available follow-up that included substance use variables identical to those at baseline was selected for inclusion. To harmonize variable coding across datasets with differing response formats, the same unified codebook procedure described for predictor variables was also applied to substance use measures. For example, variations in how alcohol use frequency was measured across datasets (e.g., weekly vs. monthly frequency indexes; drink days vs. abstinent days) were standardized into a common coding scheme, allowing for consistent synthesis and comparison (see Supplementary Table 1). Recoding was conducted by E.C. and a team of undergraduate research assistants ($\kappa = .92$), with each variable recoded independently by at least two individuals to ensure accuracy and consistency. Discrepancies were resolved through discussion.

Dataset Characteristics

Dataset characteristics were coded by both the first and fourth authors (D.K. & E.C.). The following characteristics were coded: 1) sample size at baseline assessment; 2) sample size at follow-up assessment; 3) year of data collection; 4) overall racial composition of participants; 5) specific drug type examined (i.e., alcohol, nicotine, & cannabis); 6) pattern of substance use examined (e.g., frequency, quantity, binge use); 7) time lag between baseline and follow-up assessments, in months; 8) total number of baseline variables; and 9) number of baseline predictor variables included in the current analyses.

Data-Analytic Strategy

Our pre-registered data analytic plan and final syntax files for machine learning models are available on the Open Science Framework (OSF; https://osf.io/g4w9z/). The meta-analytic materials, including the final master codebook of predictors and effect size computation syntax, are also accessible via OSF.

Machine Learning. Each dataset was analyzed separately using Random Forest (RF), a non-parametric ensemble learning method well-suited for handling high-dimensional data and capable of capturing non-linear effects and complex interactions while minimizing overfitting (Breiman, 2001). Models were implemented using the *scikit-learn* Python package (Pedregosa et al., 2011) and were trained to predict both baseline substance use and longitudinal changes in substance use. Longitudinal change outcomes were operationalized as simple difference scores (T2 − T1), consistent with prior large-scale, multi-dataset work (Joel et al., 2020). Although difference scores can be more vulnerable to measurement error than latent change score approaches, this operationalization was selected because it could be implemented consistently across all harmonized datasets despite variation in design, follow-up timing, and measurement structure. Baseline substance use was not entered as a predictor in these models because it was

already embedded in the outcome definition, and its inclusion would have introduced mathematical coupling between predictor and outcome. Additional sensitivity analyses were conducted in which baseline substance use was included as a predictor in models predicting both follow-up substance use and difference-score outcomes. As expected, including baseline substance use improved model performance (see Supplementary Table 4).

For each dataset, we implemented 4-fold cross-validation to evaluate predictive performance. In each iteration, models were trained on three folds and evaluated on the held-out test fold. Predictive accuracy was assessed exclusively in the held-out test folds, and performance metrics (R and R^2) were averaged across folds to obtain a single cross-validated estimate of model performance for each dataset and outcome. These correlation-based performance estimates (R) were subsequently used as effect sizes in the meta-analysis; cross-validation folds were not treated as independent observations, thereby avoiding inflation of precision due to non-independence. We also calculated mean absolute error (MAE) for each random forest model as a case-level prediction-error metric. MAE reflects the average absolute difference between observed and predicted substance use values in the held-out test folds. Because outcomes were harmonized to common response scales across datasets, MAE was averaged across datasets within each timepoint and outcome type. Aggregated MAE values for random forest models are reported in Supplementary Table 5.

All preprocessing and model-estimation steps were conducted within the training data for each outer cross-validation fold and then applied without modification to the corresponding held-out fold. Thus, any imputation or preprocessing parameters, scaling parameters, feature availability decisions, model coefficients, and tuned hyperparameters were estimated using only the training data. No information from the held-out fold was used to select predictors, tune

hyperparameters, scale or impute predictors, or estimate model weights. Hyperparameter tuning was conducted using 4-fold cross-validation nested within the training data of each outer fold. Primary tuning focused on *min_samples_leaf*, which regularizes model complexity by requiring a minimum number of observations in each terminal node. Other random forest parameters were retained at scikit-learn default values. A sensitivity analysis additionally tuning max_features produced only marginal improvement in predictive performance relative to tuning min_samples_leaf alone (see Supplementary Table 2).

In addition, to further evaluate the generalizability of predictive models across datasets, we conducted exploratory leave-one-dataset-out analyses. In this approach, models were trained on pooled data from six datasets and then evaluated on the held-out seventh dataset, iterating such that each dataset served once as the independent test set. These analyses were conducted using the same modeling framework as the primary machine learning analyses, with all preprocessing, feature selection procedures, and model parameters determined exclusively within the training datasets and held constant when applied to the held-out dataset. Model performance was evaluated using the proportion of variance explained (R^2) in the held-out dataset.

Given our main analyses focus on the meta-analytic aggregation of predictor importance values across datasets—rather than direct model comparison based on these performance scores—SHapley Additive exPlanations (SHAP) values were computed for each trained random forest model to identify and quantify the contributions of individual predictors to model performance (Lundberg & Lee, 2017). SHAP values provide a consistent, model-agnostic framework for interpreting feature importance by estimating the marginal contribution of each predictor to the model's output. For each dataset and outcome, mean absolute SHAP values were calculated across all samples to summarize the average importance of each predictor.

Regression Models. We first estimated standard linear regression models to examine the extent to which baseline predictors explained variance in substance use outcomes. Given the large number of predictors and potential multicollinearity, we also estimated adjusted linear regression models in which predictors with high variance inflation factors (VIF > 5) were removed to improve model stability. To account for the high dimensionality and intercorrelation among predictors, we conducted additional analyses using Elastic Net regularization. Elastic Net combines L1 (lasso) and L2 (ridge) penalties, allowing for both variable selection and coefficient shrinkage, and is well-suited for predictive modeling in settings with many correlated predictors (Zhou & Hastie, 2005). Model performance was evaluated using cross-validated estimates of prediction accuracy, consistent with the approach used for the random forest models. For all regression-based models, coefficients were estimated exclusively within the training folds and then applied without modification to the held-out folds. Thus, coefficients varied across outer folds, as expected, because a new model was trained within each fold; however, no information from the held-out fold was used to estimate coefficients, select predictors, scale predictors, tune hyperparameters, or otherwise modify the fitted model.

Elastic Net hyperparameters were tuned using only the training portion of each outer cross-validation fold before evaluating prediction performance in the held-out test fold. Specifically, the models tuned both the overall regularization strength parameter and the L1/L2 mixing parameter. Across folds, selected values included regularization strength values of 0.001, 0.01, 0.1, 1, and 10, and L1/L2 mixing values of 0, 0.1, 0.3, 0.5, 0.7, 0.9, and 1.0. The selected regularization strength and L1/L2 mixing parameter were then fixed and applied to the held-out fold for performance evaluation.

Sensitivity power analyses indicated that the regression-based models were generally powered to detect small standardized effects. Across datasets, the minimum detectable standardized slopes ranged from 0.033 to 0.123. Minimum detectable slopes were 0.045 for AddHealth, 0.085 for DUT, 0.123 for Flint, 0.033 for HSB, 0.038 for NELS, 0.038 for NLSY97, and 0.071 for NYSCS. Thus, most datasets were powered to detect small effects, although Flint had the least sensitivity due to its smaller sample size.

Meta-Analysis. Results across the seven datasets were synthesized using multivariate random-effects meta-analysis models implemented via the *metafor* package in R (Viechtbauer, 2010). For each dataset and outcome, model performance was first quantified as the cross-validated correlation (*r*) between observed and predicted substance use. These effect sizes were then pooled across datasets. The models accounted for the hierarchical structure of the data by including two random intercepts: one capturing clustering of effect sizes within predictor/outcome lines and another capturing variability across study samples. Models were fitted using restricted maximum likelihood estimation (REML), and degrees of freedom were adjusted using a t-distribution to improve accuracy with smaller numbers of higher-level units. Separate meta-analyses were conducted for baseline substance use and for change in substance use over time. The reported R^2 values were obtained by squaring the pooled correlation coefficients.

To examine potential sources of heterogeneity across studies, we conducted a series of meta-regression analyses including dataset as a categorical moderator, as well as continuous dataset-level characteristics including follow-up length (time in months between baseline and follow-up assessments), mean sample age at baseline, baseline year of data collection, and

baseline sample size. Sensitivity analyses further stratified results by substance type (alcohol, cannabis, and cigarette) and longitudinal time interval (short-, mid-, and long-term follow-up).

Data Availability

The datasets analyzed in this study are publicly available and can be accessed through their respective public repositories. No new data were generated. Analysis code is available from the corresponding author upon reasonable request.

Author Note: This study relied exclusively on publicly available datasets and did not involve human subjects research requiring institutional review board approval. The datasets analyzed in this study are publicly available and can be accessed through their respective public repositories. No new data were generated. Analysis code is available from the corresponding author upon reasonable request. As part of prior dissemination, the corresponding author presented this work at the annual meeting of the Society for Research on Psychopathology (SRP) in September 2025.

REFERENCES

References marked with an asterisk indicate studies included in the meta-analysis.

Akinwande, M. O., Dikko, H. G., & Samson, A. (2015). Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. *Open Journal of Statistics*, *05*(07), 754–767. https://doi.org/10.4236/ojs.2015.57075

Allen, J. P., Chango, J., Szwedo, D., Schad, M., & Marston, E. (2012). Predictors of susceptibility to peer influence regarding substance use in adolescence. *Child Development*, *83*(1), 337–350.

Ansell, E. B., Laws, H. B., Roche, M. J., & Sinha, R. (2015). Effects of marijuana use on impulsivity and hostility in daily life. *Drug and Alcohol Dependence*, *148*, 136–142. https://doi.org/10.1016/j.drugalcdep.2014.12.029

Armor, D. J., Polich, J. M., & Stambul, H. B. (1978). Reliability and validity of self-reported drinking behavior. *Alcoholism and Treatment*, 173–211.

Atherton, O. E., Conger, R. D., Ferrer, E., & Robins, R. W. (2016). Risk and Protective Factors for Early Substance Use Initiation: A Longitudinal Study of Mexican-Origin Youth. *Journal of Research on Adolescence*, *26*(4), 864–879. https://doi.org/10.1111/jora.12235

Babyak, M. A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models. *Psychosomatic Medicine*, *66*(3), 411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Badiani, A. (2013). Substance-specific environmental influences on drug use and drug preference in animals and humans. *Current Opinion in Neurobiology*, *23*(4), 588–596. https://doi.org/10.1016/j.conb.2013.03.010

Baker, T. B., Piper, M. E., McCarthy, D. E., Majeskie, M. R., & Fiore, M. C. (2004). Addiction motivation reformulated: An affective processing model of negative reinforcement. *Psychological Review*, *111*(1), 33–51. https://doi.org/10.1037/0033-295X.111.1.33

Becoña, E., Vázquez, F. L., Fuentes, M. J., & del Carmen Lorenzo, M. (1998). Anxiety, affect, depression and cigarette consumption. *Personality and Individual Differences*, *26*(1), 113–119. https://doi.org/10.1016/S0191-8869(98)00129-9

Bellis, M. A., Hughes, K., Morleo, M., Tocque, K., Hughes, S., Allen, T., Harrison, D., & Fe-Rodriguez, E. (2007). Predictors of risky alcohol consumption in schoolchildren and their implications for preventing alcohol-related harm. *Substance Abuse Treatment, Prevention, and Policy*, *2–15*, 15.

Bennett, M. E., McCrady, B. S., Johnson, V., & Pandina, R. J. (1999). Problem drinking from young adulthood to adulthood: Patterns, predictors and outcomes. *Journal of Studies on Alcohol*, *60*(5), 605–614.

Bidonde, J., Meneses-Echavez, J. F., Hafstad, E., Brunborg, G. S., & Bang, L. (2023). Methods, strategies, and incentives to increase response to mental health surveys among adolescents: A systematic review. *BMC Medical Research Methodology*, *23*(1), 270. https://doi.org/10.1186/s12874-023-02096-z

Blumenthal, H., Leen-Feldner, E. W., Frala, J. L., Badour, C. L., & Ham, L. S. (2010). Social anxiety and motives for alcohol use among adolescents. *Psychology of Addictive Behaviors*, *24*(3), 529.

Borsari, B., Borsari, B., & Carey, K. B. (2006). How the quality of peer relationships influences college alcohol use. *Drug and Alcohol Review*, *25*(4), 361–370.

Brooks-Russell, A., Conway, K. P., Liu, D., Xie, Y., Vullo, G. C., Li, K., Iannotti, R. J., Compton, W., & Simons-Morton, B. (2015). Dynamic Patterns of Adolescent Substance Use: Results From a Nationally Representative Sample of High School Students. *Journal of Studies on Alcohol and Drugs*, *76*(6), 962–970. https://doi.org/10.15288/jsad.2015.76.962

Brumback, T., Thompson, W., Cummins, K., Brown, S., & Tapert, S. (2021). Psychosocial predictors of substance use in adolescents and young adults: Longitudinal risk and protective factors. *Addictive Behaviors*, *121*, 106985. https://doi.org/10.1016/j.addbeh.2021.106985

Caspers, K. M., Cadoret, R. J., Langbehn, D., Yucuis, R., & Troutman, B. (2005). Contributions of attachment style and perceived social support to lifetime use of illicit substances. *Addictive Behaviors*, *30*(5), 1007–1011.

Chassin, L., Pitts, S. C., DeLucia, C., & Todd, M. (1999). A longitudinal study of children of alcoholics: Predicting young adult substance use disorders, anxiety, and depression. *Journal of Abnormal Psychology*, *108*(1), 106–119. https://doi.org/10.1037/0021-843X.108.1.106

Chassin, L., Presson, C. C., Sherman, S. J., & Mulvenon, S. (1994). Family history of smoking and young adult smoking behavior. *Psychology of Addictive Behaviors*, *8*(2), 102–110. https://doi.org/10.1037/0893-164X.8.2.102

Chen, P., & Jacobson, K. C. (2012). Developmental Trajectories of Substance Use From Early Adolescence to Young Adulthood: Gender and Racial/Ethnic Differences. *Journal of Adolescent Health*, *50*(2), 154–163. https://doi.org/10.1016/j.jadohealth.2011.05.013

Chessick, R. D. (1964). The Problem of Tobacco Habituation. *JAMA*, *188*(10). https://doi.org/10.1001/jama.1964.03060360092021

Clapp, J. D., & Shillington, A. M. (2001). Environmental predictors of heavy episodic drinking. *The American Journal of Drug and Alcohol Abuse*, *27*(2), 301–313.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge Academic.

Coughlan, M., Cronin, P., & Ryan, F. (2009). Survey research: Process and limitations. *International Journal of Therapy and Rehabilitation*, *16*(1), 9–15. https://doi.org/10.12968/ijtr.2009.16.1.37935

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, *34*, 187–220.

Crabbe, J. C. (2002). Genetic contributions to addiction. *Annual Review of Psychology*, *53*(1), 435–462. https://doi.org/10.1146/annurev.psych.53.100901.135142

Creswell, K. G. (2021). Drinking together and drinking alone: A social-contextual framework for examining risk for alcohol use disorder. *Current Directions in Psychological Science*, *30*(1), 19–25. https://doi.org/10.1177/0963721420969406

Daoud, J. I. (2017). Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series*, *949*, 012009. https://doi.org/10.1088/1742-6596/949/1/012009

Day, N. L., & Robles, N. (1989). Methodological issues in the measurement of substance use. *Annals of the New York Academy of Sciences*, *562*(1), 8–13.

Delva, J., Allen-Meares, P., & Momper, S. L. (2010). *Cross-Cultural Research*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195382501.001.0001

Ducci, F., & Goldman, D. (2008). Genetic approaches to addiction: Genes and alcohol. *Addiction*, *103*(9), 1414–1428. https://doi.org/10.1111/j.1360-0443.2008.02203.x

Epstein, D. H., Willner-Reid, J., & Preston, K. L. (2010). Addiction and emotion: Theories, assessment techniques, and treatment implications. In J. D. Kassel (Ed.), *Substance abuse and emotion.* (pp. 259–280). American Psychological Association. https://doi.org/10.1037/12067-011

Fairbairn, C. E., Briley, D. A., Kang, D., Fraley, R. C., Hankin, B. L., & Ariss, T. (2018). A meta-analysis of attachment security and substance use. *Psychological Bulletin*, *144*(5), 532–555.

Fairbairn, C. E., & Sayette, M. A. (2014). A social-attributional analysis of alcohol response. *Psychological Bulletin*, *140*(5), 1361–1382. https://doi.org/10.1037/a0037563

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128.

Fromme, K., de Wit, H., Hutchison, K. E., Ray, L., Corbin, W. R., Cook, T. A. R., Wall, T. L., & Goldman, D. (2004). Biological and behavioral markers of alcohol sensitivity. *Alcoholism: Clinical and Experimental Research*, *28*, 247–256. https://doi.org/10.1097/01.ALC.0000113420.28472.25

Gau, S. S. F., Chong, M.-Y., Yang, P., Yen, C.-F., Liang, K.-Y., & Cheng, A. T. A. (2007). Psychiatric and psychosocial predictors of substance use disorders among adolescents: Longitudinal study. *British Journal of Psychiatry*, *190*(1), 42–48. https://doi.org/10.1192/bjp.bp.106.022871

* Giovino, G. A., & Barker, D. C. (2013). *National Youth Smoking Cessation Survey, 2003-2005: Version 2* (Version v2) [Dataset]. ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR34275.V2

Gliksberg, O., Livne, O., Lev-Ran, S., Rehm, J., Hasson-Ohayon, I., & Feingold, D. (2022). The Association Between Cannabis Use and Perceived Social Support: The Mediating Role of Decreased Social Network. *International Journal of Mental Health and Addiction*, *20*(5), 2799–2812. https://doi.org/10.1007/s11469-021-00549-4

Green, R., Wolf, B. J., Chen, A., Kirkland, A. E., Ferguson, P. L., Browning, B. D., Bryant, B. E., Tomko, R. L., Gray, K. M., Mewton, L., & Squeglia, L. M. (2024). Predictors of Substance Use Initiation by Early Adolescence. *American Journal of Psychiatry*, *181*(5), 423–433. https://doi.org/10.1176/appi.ajp.20230882

Harris, K. M., Griffin, B. A., McCaffrey, D. F., & Morral, A. R. (2008). Inconsistencies in self-reported drug use by adolescents in substance abuse treatment: Implications for outcome and performance measurements. *Journal of Substance Abuse Treatment*, *34*(3), 347–355. https://doi.org/10.1016/j.jsat.2007.05.004

Hasin, D. S., O'Brien, C. P., Auriacombe, M., Borges, G., Bucholz, K., Budney, A., Compton, W. M., Crowley, T., Ling, W., Petry, N. M., Schuckit, M., & Grant, B. F. (2013). DSM-5 Criteria for Substance Use Disorders: Recommendations and Rationale. *American Journal of Psychiatry*, *170*(8), 834–851. https://doi.org/10.1176/appi.ajp.2013.12060782

Higgins, J. P., & Altman, D. G. (2008). Assessing risk of bias in included studies. In J. P. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (pp. 187–241). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470712184.ch8

* Hummer, R. A., Aiello, A. E., Harris, K. M., & Udry, J. R. (2008). *National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2025 [Public Use]: Version 26* (Version v26) [Dataset]. ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR21600.V26

Hussong, A. M. (2003). Social influences in motivated drinking among college students. *Psychology of Addictive Behaviors*, *17*(2), 142–150. https://doi.org/10.1037/0893-164X.17.2.142

Janssen, T., Larsen, H., Peeters, M., Boendermaker, W. J., Vollebergh, W. A. M., & Wiers, R. W. (2015). Do online assessed self-report and behavioral measures of impulsivity-related constructs predict onset of substance use in adolescents? *Addictive Behaviors Reports*, *1*, 12–18. https://doi.org/10.1016/j.abrep.2015.01.002

Joel, S., Eastwick, P. W., Allison, C. J., Arriaga, X. B., Baker, Z. G., Bar-Kalifa, E., Bergeron, S., Birnbaum, G. E., Brock, R. L., Brumbaugh, C. C., Carmichael, C. L., Chen, S., Clarke, J., Cobb, R. J., Coolsen, M. K., Davis, J., De Jong, D. C., Debrot, A., DeHaas, E. C., … Wolf, S. (2020). Machine learning uncovers the most robust self-report predictors of relationship quality across 43 longitudinal couples studies. *Proceedings of the National Academy of Sciences*, *117*(32), 19061–19071. https://doi.org/10.1073/pnas.1917036117

Johnston, L. D., Miech, R. A., Patrick, M. E., O'Malley, P. M., Schulenberg, J. E., & Bachman, J. G. (2023). *Monitoring the Future National Survey Results on Drug Use, 1975-2022: Overview, Key Findings on Adolescent Drug Use*.

Karriker-Jaffe, K. J., Zemore, S. E., Mulia, N., Jones-Webb, R., Bond, J., & Greenfield, T. K. (2012). Neighborhood disadvantage and adult alcohol outcomes: Differential risk by race and gender. *Journal of Studies on Alcohol and Drugs*, *73*(6), 865–873.

Khorrami, Z., Zolala, F., Haghdoost, A., Sadatmoosavi, A., Ben Taleb, Z., Kondracki, A., Ward, K. D., Shahbaz, M., & Ebrahimi Kalan, M. (2021). Job-related stress and tobacco smoking: A systematic review. *Journal of Workplace Behavioral Health*, *36*(4), 259–277. https://doi.org/10.1080/15555240.2021.1960854

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, *72*(6), 558–569. https://doi.org/10.4097/kja.19087

Kim, S., Ziedonis, D., & Chen, K. (2007). Tobacco use and dependence in Asian Americans: A review of the literature. *Nicotine & Tobacco Research*, *9*(2), 169–184. https://doi.org/10.1080/14622200601080323

Koob, G. F. (2015). The dark side of emotion: The addiction perspective. *European Journal of Pharmacology*, *753*, 73–87. https://doi.org/10.1016/j.ejphar.2014.11.044

Koob, G. F., & Le Moal, M. (2008). Neurobiological mechanisms for opponent motivational processes in addiction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1507), 3113–3123. https://doi.org/10.1098/rstb.2008.0094

Kraemer, H. C., & Blasey, C. M. (2004). Centring in regression analyses: A strategy to prevent errors in statistical inference. *International Journal of Methods in Psychiatric Research*, *13*(3), 141–151. https://doi.org/10.1002/mpr.170

Kulkarni, K. R., Schafer, M., Berner, L. A., Fiore, V. G., Heflin, M., Hutchison, K., Calhoun, V., Filbey, F., Pandey, G., Schiller, D., & Gu, X. (2023). An Interpretable and Predictive Connectivity-Based Neural Signature for Chronic Cannabis Use. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *8*(3), 320–330. https://doi.org/10.1016/j.bpsc.2022.04.009

Kwako, L. E., Momenan, R., Litten, R. Z., Koob, G. F., & Goldman, D. (2016). Addictions Neuroclinical Assessment: A Neuroscience-Based Framework for Addictive Disorders. *Biological Psychiatry*, *80*(3), 179–189. https://doi.org/10.1016/j.biopsych.2015.10.024

Litten, R. Z., Ryan, M. L., Falk, D. E., Reilly, M., Fertig, J. B., & Koob, G. F. (2015). Heterogeneity of Alcohol Use Disorder: Understanding Mechanisms to Advance Personalized Treatment. *Alcoholism: Clinical and Experimental Research*, *39*(4), 579–584. https://doi.org/10.1111/acer.12669

Lundberg, S. M., & Lee, S. L. (2017). A unified approach to interpreting model predictions. In L. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774).

MacPherson, L., Magidson, J. F., Reynolds, E. K., Kahler, C. W., & Lejuez, C. W. (2010). Changes in Sensation Seeking and Risk-Taking Propensity Predict Increases in Alcohol Use Among Early Adolescents. *Alcoholism: Clinical and Experimental Research*, *34*(8), 1400–1408. https://doi.org/10.1111/j.1530-0277.2010.01223.x

Malouff, J. M., Thorsteinsson, E. B., & Schutte, N. S. (2006). The Five-Factor Model of Personality and Smoking: A Meta-Analysis. *Journal of Drug Education*, *36*(1), 47–58. https://doi.org/10.2190/9EP8-17P8-EKG7-66AD

McGee, R., Williams, S., Poulton, R., & Moffitt, T. (2000). A longitudinal study of cannabis use and mental health from adolescence to early adulthood. *Addiction*, *95*(4), 491–503.

McLellan, A. T., Lewis, D. C., O'Brien, C. P., & Kleber, H. D. (2000). Drug Dependence, a Chronic Medical Illness: Implications for Treatment, Insurance, and Outcomes Evaluation. *JAMA*, *284*(13), 1689. https://doi.org/10.1001/jama.284.13.1689

Morales, A. M., Jones, S. A., Kliamovich, D., Harman, G., & Nagel, B. J. (2020). Identifying Early Risk Factors for Addiction Later in Life: A Review of Prospective Longitudinal Studies. *Current Addiction Reports*, *7*(1), 89–98. https://doi.org/10.1007/s40429-019-00282-y

* *National Longitudinal Survey of Youth 1997*. (n.d.). [Dataset]. Bureau of Labor Statistics, U.S. Department of Labor. National Longitudinal Survey of Youth 1997 cohort, 1997-2021 (rounds 1-20).

Orcutt, J. D., & Schwabe, A. M. (2012). Gender, Race/Ethnicity, and Deviant Drinking: A Longitudinal Application of Social Structure and Social Learning Theory. *Sociological Spectrum*, *32*(1), 20–36. a9h.

P. Vatcheva, K., & Lee, M. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology: Open Access*, *06*(02). https://doi.org/10.4172/2161-1165.1000227

Pacek, L. R., Malcolm, R. J., & Martins, S. S. (2012). Race/Ethnicity Differences between Alcohol, Marijuana, and Co-occurring Alcohol and Marijuana Use Disorders and Their Association with Public Health and Social Problems Using a National Sample. *The American Journal on Addictions*, *21*(5), 435–444. https://doi.org/10.1111/j.1521-0391.2012.00249.x

Patrick, M. E., & Terry-McElrath, Y. M. (2017). High-intensity drinking by underage young adults in the United States: Underage high-intensity drinking. *Addiction*, *112*(1), 82–93. https://doi.org/10.1111/add.13556

Pearson, M. R., Schwebel, F. J., Richards, D. K., & Witkiewitz, K. (2022). Examining replicability in addictions research: How to assess and ways forward. *Psychology of Addictive Behaviors*, *36*(3), 260–270. https://doi.org/10.1037/adb0000730

Phillips, K. T., Phillips, M. M., Lalonde, T. L., & Prince, M. A. (2018). Does social context matter? An ecological momentary assessment study of marijuana use among college students. *Addictive Behaviors*, *83*, 154–159. https://doi.org/10.1016/j.addbeh.2018.01.004

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903.

Podsakoff, P. M., Podsakoff, N. P., Williams, L. J., Huang, C., & Yang, J. (2024). Common method bias: It's bad, it's complex, it's widespread, and it's not easy to fix. *Annual Review of Organizational Psychology and Organizational Behavior*, *11*(1), 17–61. https://doi.org/10.1146/annurev-orgpsych-110721-040030

Quinn, P. D., & Harden, K. P. (2012). Differential changes in impulsivity and sensation seeking and the escalation of substance use from adolescence to early adulthood. *Development and Psychopathology*, *1*, 1–17.

Quinn, P. D., Stappenbeck, C. A., & Fromme, K. (2011). Collegiate heavy drinking prospectively predicts change in sensation seeking and impulsivity. *Journal of Abnormal Psychology*, *120*, 543–556.

Ray, L. A., Mackillop, J., & Monti, P. M. (2010). Subjective responses to alcohol consumption as endophenotypes: Advancing behavioral genetics in etiological and treatment models of alcoholism. *Substance Use & Misuse*, *45*(11), 1742–1765. https://doi.org/10.3109/10826084.2010.482427

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., … McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Saunders, J. B., Aasland, O. G., Babor, T. F., De la Fuente, J. R., & Grant, M. (1993). Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction*, *88*(6), 791–804.

Sayette, M. A., Fairbairn, C. E., & Creswell, K. G. (2015). Alcohol and emotion. In C. A. Kopetz & C. W. Lejuez (Eds.), *Addictions: A Social Psychological Perspective* (1st ed., p. 98). Routledge.

Schulenberg, J. E., & Maggs, J. L. (2002). A developmental perspective on alcohol use and heavy drinking during adolescence and the transition to young adulthood. *Journal of Studies on Alcohol*, (14), 54–70.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), 93–105. https://doi.org/10.1037/0003-066X.54.2.93

Shenoi, R. P., Linakis, J. G., Bromberg, J. R., Casper, T. C., Richards, R., Mello, M. J., Chun, T. H., Spirito, A., & PEDIATRIC EMERGENCY CARE APPLIED RESEARCH NETWORK. (2019). Predictive Validity of the CRAFFT for Substance Use Disorder. *Pediatrics*, *144*(2), e20183415. https://doi.org/10.1542/peds.2018-3415

Shih, R. A., Tucker, J. S., Miles, J. N. V., Ewing, B. A., Pedersen, E. R., & D'Amico, E. J. (2015). Differences in substance use and substance use risk factors by Asian subgroups. *Asian American Journal of Psychology*, *6*(1), 38–46. https://doi.org/10.1037/a0036251

Sinha, R. (2008). Chronic Stress, Drug Use, and Vulnerability to Addiction. *Annals of the New York Academy of Sciences*, *1141*(1), 105–130. https://doi.org/10.1196/annals.1441.030

Substance Abuse and Mental Health Services Administration. (2022). *National Survey on Drug Use and Health 2022* (NSDUH-2022-DS0001). Center for Behavioral Health Statistics and Quality. https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2022-nsduh-2022-ds0001

Tang, Y., Caswell, E., Mohamed, R., Wilson, N., Osmanovic, E., Smith, G., Hartley, S. D., & Bhandari, R. (2024). A systematic review of validity of US survey measures for assessing substance use and substance use disorders. *Systematic Reviews*, *13*(1), 166. https://doi.org/10.1186/s13643-024-02536-x

Thai, N. D., Connell, C. M., & Tebes, J. K. (2010). Substance use among Asian American adolescents: Influence of race, ethnicity, and acculturation in the context of key risk and protective factors. *Asian American Journal of Psychology*, *1*(4), 261–274. https://doi.org/10.1037/a0021703

Torres, O. V., & O'Dell, L. E. (2016). Stress is a principal factor that promotes tobacco use in females. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *65*, 260–268. https://doi.org/10.1016/j.pnpbp.2015.04.005

* Turner, J. (2011). *Drug Use Trajectories: Ethnic/Racial Comparisons, 1998-2002 [United-States]: Version 1* (Version v1) [Dataset]. ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR30862.V1

* United States Department Of Education. Institute Of Education Sciences. National Center For Education Statistics. (1984). *High School and Beyond, 1980: A Longitudinal Survey of Students in the United States: Version 2* (Version v2) [Dataset]. ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR07896.V2

* United States Department of Education. National Center for Education Statistics. (1990). *National Education Longitudinal Study, 1988: Version 3* (Version v3) [Dataset]. ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR09389.V3

Volkow, N. D., Koob, G. F., & McLellan, A. T. (2016). Neurobiologic Advances from the Brain Disease Model of Addiction. *New England Journal of Medicine*, *374*(4), 363–371. https://doi.org/10.1056/NEJMra1511480

Vungkhanching, M., Sher, K. J., Jackson, K. M., & Parra, G. R. (2004). Relation of attachment style to family history of alcoholism and alcohol use disorders in early adulthood. *Drug and Alcohol Dependence*, *75*(1), 47–53. https://doi.org/https://doi.org/10.1016/j.drugalcdep.2004.01.013

Wills, T. A., & Cleary, S. D. (1996). How are social support effects mediated? A test with parental support and adolescent substance use. *Journal of Personality and Social Psychology*, *71*(5), 937–952. https://doi.org/10.1037/0022-3514.71.5.937

Windle, M., & Windle, R. C. (2012). Early onset problem behaviors and alcohol, tobacco, and other substance use disorders in young adulthood. *Drug and Alcohol Dependence*, *121*(1–2), 152–158. https://doi.org/10.1016/j.drugalcdep.2011.08.024

Witkiewitz, K., Kirouac, M., Baurley, J. W., & McMahan, C. S. (2023). Patterns of drinking behavior around a treatment episode for alcohol use disorder: Predictions from pre-

treatment measures. *Alcohol: Clinical and Experimental Research*, *47*(11), 2138–2148. https://doi.org/10.1111/acer.15183

Witkiewitz, K., Litten, R. Z., & Leggio, L. (2019). Advances in the science and treatment of alcohol use disorder. *Science Advances*, *5*(9), eaax4043. https://doi.org/10.1126/sciadv.aax4043

World Health Organization. (2024a). *Global alcohol action plan 2022-2030*.

World Health Organization. (2024b). *Global status report on alcohol and health and treatment of substance use disorders*.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yip, S. W., Kiluk, B., & Scheinost, D. (2020). Toward Addiction Prediction: An Overview of Cross-Validated Predictive Modeling Findings and Considerations for Future Neuroimaging Research. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(8), 748–758. https://doi.org/10.1016/j.bpsc.2019.11.001

Zapolski, T. C. B., Pedersen, S. L., McCarthy, D. M., & Smith, G. T. (2014). Less drinking, yet more problems: Understanding African American drinking and related problems. *Psychological Bulletin*, *140*(1), 188–223.

* Zimmerman, M. A. (2014). *Flint [Michigan] Adolescent Study (FAS): A Longitudinal Study of School Dropout and Substance Use, 1994-1997: Version 1* (Version v1) [Dataset]. ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR34598.V1

Zvolensky, M. J., Taha, F., Bono, A., & Goodwin, R. D. (2015). Big five personality factors and cigarette smoking: A 10-year study among US adults. *Journal of Psychiatric Research*, *63*, 91–96. https://doi.org/10.1016/j.jpsychires.2015.02.008

Table 1. Demographic features of each dataset.

	AddHealth	DUT	Flint	HSB	NELS	NLSY97	NYSCS
Survey year							
Baseline	1994-1995	1998-2000	1994	1982	1990	1997	2003
Follow-up	2008-2009	2000-2002	1997	1986	1992	2015	2005
Sample size							
Baseline	6504	1803	850	11995	12144	8984	2582
Follow-up	1742	524	741	10352	10948	7103	1431
Mean Age	16.16	20.01	14.55	20.24	15.87	19.01	20.30
Gender							
%Female	52.6%	47.0%	50.0%	52.7%	52.4%	48.8%	47.0%
Race/Ethnicity							
White	66.0%	69.7%	16.8%	48.0%	62.8%	51.9%	69.1%
Black	24.9%	27.0%	80.1%	26.6%	8.6%	26.0%	11.5%
Hispanic/Latino	11.4%	47.1%	N/A	18.9%	11.9%	N/A	13.2%
Other	7.8%	3.3%	4.1%	5.4%	10.5%	22.1%	6.2%
Outcomes							
Alcohol Frequency	Y	Y	Y	Y	Y	Y	Y
Alcohol Quantity	Y			Y		Y	
Binge Drinking	Y	Y	Y		Y	Y	Y
Cigarette Frequency	Y					Y	Y
Cigarette Quantity	Y	Y			Y		Y
Cannabis Frequency	Y	Y				Y	

Table 2. Meta-Regression Results of Models Predicting Substance Use Outcomes at Baseline Time Point and Changes from Baseline to Follow-Up.

1-a. Percentage of variance explained predicting substance use at T1

Random Forest	*r*	*SE*	*t*	*df*	*p*	*CI*		
Overall	0.376	0.066	5.668	25	<.001	[0.239,	0.512]	***
By Substance Type:								
Alcohol	0.380	0.077	4.958	15	0.000	[0.217,	0.544]	***
Cigarette	0.427	0.056	7.620	6	0.000	[0.290,	0.565]	***
Cannabis	0.424	0.021	20.196	2	0.002	[0.333,	0.514]	**
By Dataset:								
AddHealth	0.606	0.055	10.991	5	0.000	[0.464,	0.747]	***
DUT	0.369	0.052	7.154	3	0.006	[0.205,	0.533]	**
Flint	0.058	0.077	0.753	1	0.589	[-0.914,	1.030]	
HSB	0.492	0.008	65.544	1	0.010	[0.396,	0.587]	**
NELS	0.408	0.026	15.858	2	0.004	[0.297,	0.519]	**
NLSY97	0.439	0.032	13.880	4	0.000	[0.351,	0.526]	***
NYSCS	0.236	0.012	20.019	3	0.000	[0.198,	0.273]	***

1-b. Percentage of variance explained predicting changes in substance use from T1 to T2

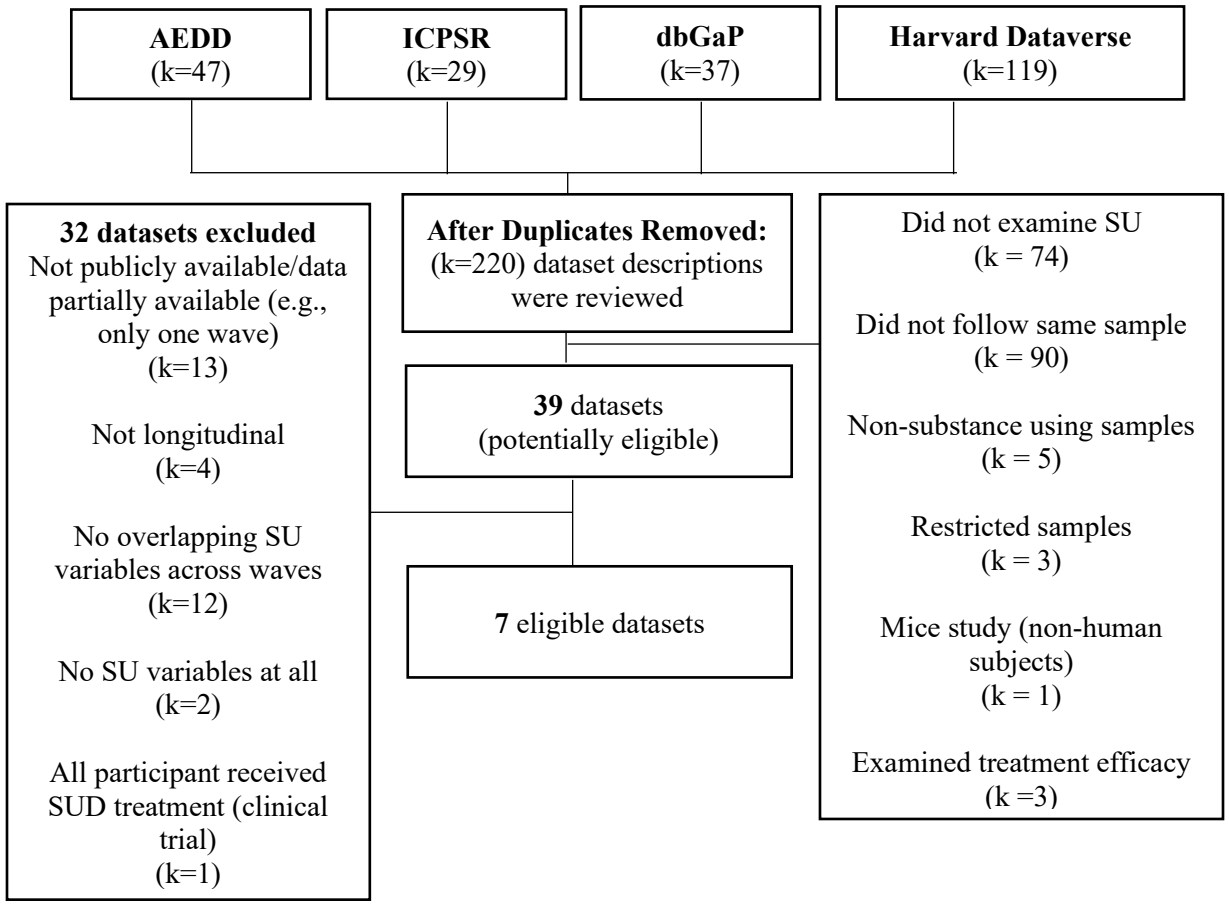
Random Forest	*r*	*SE*	*t*	*df*	*p*	*CI*		
Overall	0.188	0.038	5.005	24	<.001	[0.110,	0.265]	***
By Substance Type:								
Alcohol	0.201	0.052	3.868	14	0.002	[0.090,	0.312]	**
Cigarette	0.189	0.028	6.741	2	0.021	[0.068,	0.309]	*
Cannabis	0.181	0.037	4.942	6	0.003	[0.091,	0.270]	**
By Follow-Up Time Length:								
0-2 years	0.155	0.050	3.107	10	0.011	[0.044,	0.267]	*
3-4 years	0.106	0.098	1.087	2	0.391	[-0.314,	0.525]	
5+ years	0.273	0.049	5.558	10	0.000	[0.164,	0.382]	***
By Dataset:								
AddHealth	0.320	0.064	4.966	5	0.004	[0.154,	0.485]	**
DUT	0.255	0.046	5.518	3	0.012	[0.108,	0.402]	*
Flint	0.003	0.026	0.096	1	0.939	[-0.328,	0.333]	
HSB	0.200	0.007	28.578	1	0.022	[0.111,	0.289]	*
NELS	0.120	0.015	8.240	2	0.014	[0.057,	0.182]	*
NLSY97	0.222	0.023	9.778	4	0.001	[0.159,	0.285]	***
NYSCS	0.093	0.021	4.558	3	0.020	[0.028,	0.159]	*

Table 3. Meta-Regression Analyses Examining Moderators of Prediction Strength

Moderator		*β*	*SE*	*t*	*p*	*CI*		*F*	*p*
Dataset (Baseline)								3.864	.002**
	Intercept	0.506	0.074	6.861	<.001**	[0.359	0.653]		
	DUT	-0.269	0.107	-2.517	.014	[-0.481	-0.056]		
	Flint	-0.454	0.114	-3.984	<.001**	[-0.681	-0.227]		
	HSB	-0.038	0.113	-0.332	.741	[-0.263	0.188]		
	NELS	-0.160	0.109	-1.471	.146	[-0.377	0.057]		
	NLSY97	-0.080	0.105	-0.758	.451	[-0.289	0.130]		
	NYSCS	-0.271	0.107	-2.542	.013	[-0.484	-0.058]		
Dataset (Change)								2.184	.055
	Intercept	0.249	0.059	4.249	<.001**	[0.132	0.365]		
	DUT	-0.082	0.085	-0.958	.341	[-0.252	0.088]		
	Flint	-0.290	0.093	-3.109	.003*	[-0.476	-0.104]		
	HSB	-0.064	0.090	-0.715	.477	[-0.243	0.115]		
	NELS	-0.162	0.086	-1.880	.064	[-0.334	0.010]		
	NLSY97	-0.033	0.083	-0.392	.696	[-0.199	0.134]		
	NYSCS	-0.153	0.085	-1.805	.075	[-0.322	0.016]		
Follow-up Length		0.009	0.004	2.112	.038*	[0.001	0.018]	4.459	.038*
Mean Age (Baseline)		0.016	0.029	0.542	.589	[-0.042	0.073]	0.294	.589
Mean Age (Change)		0.017	0.016	1.033	.305	[-0.016	0.050]	1.067	.305
Year (Baseline)		-0.010	0.010	-1.023	.309	[-0.029	0.009]	1.047	.309
Year (Change)		0.004	0.004	0.990	.325	[-0.004	0.012]	0.980	.325
Sample Size (Baseline)		0.000	0.000	2.460	.016*	[0.000	0.000]	6.053	.016*
Sample Size (Change)		0.000	0.000	0.980	.330	[-0.000	0.000]	0.961	.330

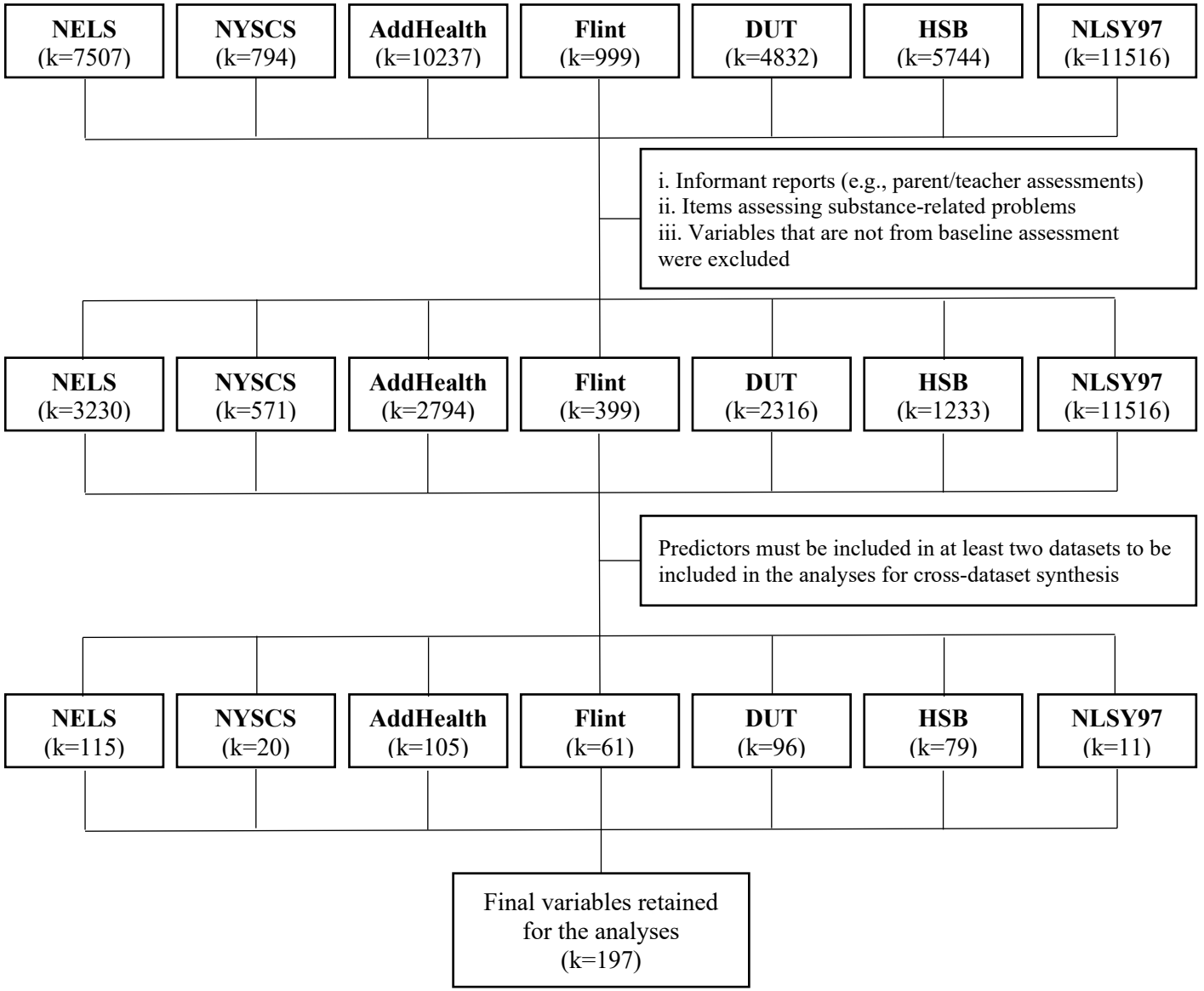
Note. AddHealth served as the reference category for dataset comparisons. Mean Age = the average participant age at baseline assessment. Year = the baseline data collection year for each dataset. Sample Size = the number of participants at baseline. Follow-up length represents the time interval (in months) between baseline and follow-up assessments. $p < .05$*, $p < .01$**.

Figure 1.a. PRISMA Flow Diagram for Identifying Eligible Datasets



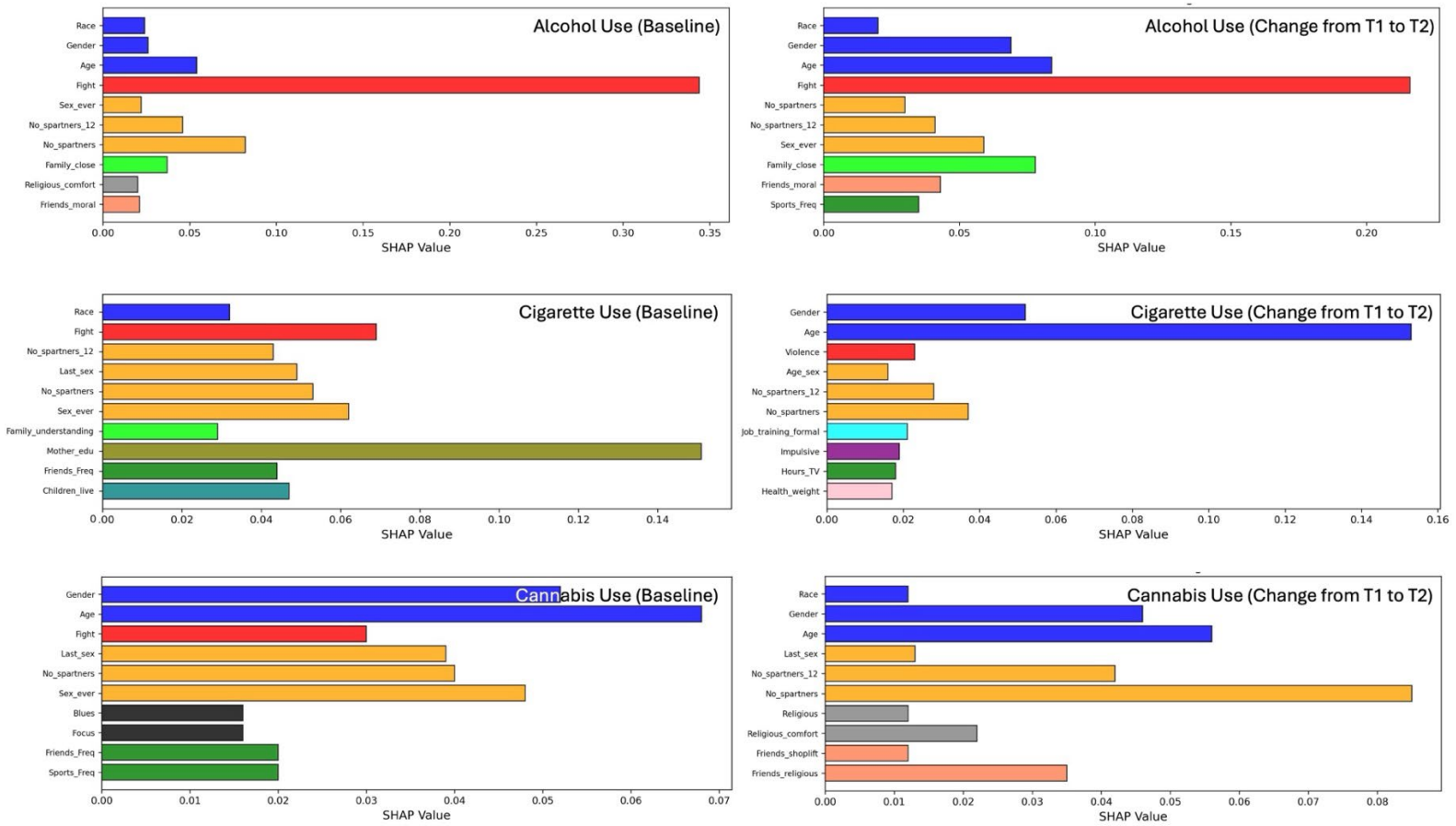
Note: AEDD = Alcohol Epidemiological Data Directory; ICPSR = Inter-university Consortium for Political and Social Research; dbGaP = the database of Genotypes and Phenotypes

Figure 1.b. PRISMA Flow Diagram for Identifying Eligible Predictor Variables



Note: NELS = National Education Longitudinal Study (United States Department of Education. National Center for Education Statistics, 1990); NYSCS = National Youth Smoking Cessation Survey (Giovino & Barker, 2013); AddHealth = National Longitudinal Study of Adolescent to Adult Health (Hummer et al., 2008); Flint = Flint Adolescent Study (Zimmerman, 2014); DUT = Drug Use Trajectory (Turner, 2011); HSB = High School and Beyond (United States Department Of Education. Institute Of Education Sciences. National Center For Education Statistics, 1984); NLSY97 = National Longitudinal Survey of Youth (*National Longitudinal Survey of Youth 1997*, n.d.). The study launch date for data harmonization was October 2018, and datasets included reflect those meeting eligibility criteria prior to this date. Harmonization and codebook development spanned over five years and involved multiple iterations to ensure high reliability and fidelity to the original sources. Mega-analytic efforts of this scale—encompassing thousands of variables—share more in common with large-scale empirical data collection than with traditional meta-analytic approaches. Accordingly, as with most large-scale empirical studies, findings from the present analysis can be interpreted as extending to the recent historical period, but not necessarily to the immediate present.

Figure 2. Bar Plots of SHAP values of Predictors of Substance Use Identified by Random Forest Analysis.



Note: See Supplemental Table 1 for full variable descriptions. Race = participant's race/ethnicity; Gender = participant's self-identified gender; Age = participant's age at baseline; Fight = history of involvement in physical fights; No_spartners = number of lifetime sexual partners; No_spartners_12 = number of sexual partners in the past 12 months; Sex_ever = ever engaged in sexual intercourse (yes/no); Last_sex = time elapsed since most recent sexual activity; Family_close = perceived closeness to family members; Friends_moral = perceived moral character of friends; Religious_comfort = comfort derived from religious faith or activities; Mother_edu = mother's highest level of educational attainment; Friends_Freq = frequency of spending time with friends; Children_live = currently living with child(ren); Family_understanding = perceived level of understanding and support from family; Sports_Freq = frequency of participation in sports activities; Focus = ability to maintain attention during tasks; Blues = frequency of experiencing depressed mood ("feeling blue"); Violence = history of perpetrating violent acts; Job_training_formal = participation in formal job training programs; Impulsive = tendency to act without forethought;

Hours_TV = average number of hours spent watching television per week; Health_weight = self-reported weight-related health concerns; Age_sex = age at first sexual intercourse; Friends_religious = number of friends involved in religious activities; Religious = strength of religious identification; Friends_shoplift = having friends who have engaged in shoplifting. Colors indicate predictor domains: blue = demographics; orange = sexual behaviors; red = violence; neon green = family support; olive green = family socioeconomic characteristics; gray = religion; coral = peer relationships; light blue = employment-related factors; dark green = leisure activities; purple = impulsivity; black = mood and cognitive symptoms; light pink = physical health.