TEDS Data

Nam-Anh Tran

1 Introduction

Substance use disorders (SUDs) represent a significant public health challenge in the United States, contributing substantially to increased morbidity, mortality, and societal costs. In 2022, an estimated 48.7 million Americans aged 12 or older were affected by SUDs. That same year, 107,941 overdose deaths were reported, while alcohol-related deaths totaled approximately 178,000 in 2021—an average of 488 deaths per day. In addition to these human costs, the economic burden is also severe: prescription opioid misuse incurs an estimated $78.5 billion annually in direct costs, and tobacco use is associated with over $300 billion in economic losses. These figures underscore the urgent need for improved prevention strategies, expanded treatment access, and informed policy interventions to mitigate the impact of substance use disorders [ref.]

Understanding substance use treatment patterns is essential for addressing the current public health crisis, as it helps identify barriers to care and supports the development of effective interventions to improve treatment outcomes and inform evidence-based policy. These patterns can be examined from multiple perspectives. For instance, disparities in treatment access, wait time to enter treatment, and length of stay often vary across regions, and such differences should be confirmed through statistical analysis. Treatment completion, a critical indicator of recovery success, may be influenced by various individual- and system-level factors. Moreover, poly-substance use is associated with elevated risks of overdose, co-occurring mental health disorders, and poorer treatment outcomes. As individuals with poly-substance use often require more complex and tailored treatment plans, understanding its prevalence is vital for designing comprehensive care strategies. Identifying and addressing these patterns enables more effective resource allocation, reduces disparities, enhances treatment retention, and ultimately lowers societal costs.

The Treatment Episode Data Set (TEDS) is a national repository of treatment data routinely collected by U.S. states to monitor their substance use treatment systems. It provides a comprehensive view of treatment admissions to publicly funded facilities across the United States [ref.]. Although TEDS presents certain limitations—such as non-random sampling and substantial missing data—it offers valuable insights into patterns of treatment access and outcomes.

This study aims to address three objectives using data from TEDS. First, we assess the spatial patterns of wait time to treatment (WTT) and length of stay (LOS) across states. Second, we examine the association between referral source and treatment completion, where evidence of association would suggest that referral source is a reliable predictor of treatment outcomes. Finally, we investigate the relationship between alcohol use and drug use, and whether this association remains consistent over the course of the treatment period.

We employ statistical models within a Bayesian framework to examine the associations between outcomes of interest and primary predictors. Bayesian methods naturally incorporate uncertainty through posterior distributions by combining prior information about parameters with observed data. This framework allows for ongoing refinement of parameter estimates as new data become available. Given that the TEDS data exhibit a hierarchical structure—where substance use observations at admission and discharge are nested within survey units, which are further nested within states—Bayesian hierarchical models are well-suited to account for this multilevel structure.

2 Data

TEDS comprises two primary components—the admissions dataset (TEDS-A) and the discharge dataset (TEDS-D)—which provide demographic, clinical, and substance use characteristics of individuals admitted to substance use treatment services. The unit of analysis is treatment admissions. TEDS represents a finite population, containing 1,351,768 records (i.e., admissions) collected across 48 U.S. states. As with many observational datasets, TEDS contains substantial missing data, largely due to variability in state reporting practices and resource or funding constraints in treatment facilities. To address these limitations, we redefine the target population for this study using three inclusion and exclusion criteria. These criteria are based on the outcomes and key predictors, and are designed to ensure that the estimands of interest can be estimated with a high degree of consistency and validity.

First, records with missing values for outcomes and key covariates are excluded to ensure the validity of the analysis. Most variables in TEDS contain missing data, which result from a combination of true missingness, non-collection, and invalid entries. While the first two types can be treated as standard missing data, invalid values indicate that the records do not meet the inclusion criteria defined by TEDS and should therefore be excluded. However, because TEDS does not distinguish between these types of missingness, we are unable to apply standard missing data techniques. As a result, we remove all such records to avoid violating TEDS's data quality standards and ensure a consistent analytic sample.

Second, we exclude states in which the categories of outcomes or primary predictors are not fully represented. Specifically, we first redefine categorical variables by collapsing categories with low proportions to avoid under- or over-representation. Then, we exclude states where the proportion of any category within the outcomes or primary predictors falls below 1%. This step mitigates the issue of zero- or near-zero cell counts, which can compromise model stability. As a result, the target population comprises 140,758 records across 23 states. Although this procedure excludes a substantial portion of the original TEDS data, it ensures that the analysis is not affected by inconsistencies arising from sparse or missing categories. However, final conclusions should be interpreted strictly within the context of this predefined target population.

We implement stratified sampling on the target population prior to analysis. Given the substantial size of the target population (140,758 records), fitting models directly—particularly Bayesian models that rely on iterative simulations—can be computationally intensive. Additionally, the use of the full dataset may produce overpowered evidence, where very small effect sizes (e.g., less than 0.05) appear statistically detectable but are not practically meaningful for decision-making. To address these issues, we reduce the dataset by selecting a representative subset through stratified sampling. This approach partitions the population into strata, from which random samples are drawn independently [ref.]. Stratified sampling ensures adequate representation within each stratum, enabling the detection of practically relevant differences while reducing computation time and avoiding misleading conclusions based on minimal effects.

Stratified sampling is conducted based on states to ensure that all 23 states remain represented in the analysis. This is essential for addressing research questions involving spatial patterns. The approach ensures a sufficient number of samples per state to produce reliable state-level estimates of the primary outcomes. The total sample size is determined using the estimated proportion of the outcome, a predefined margin of error of 0.03, and a Z-score of 1.96 corresponding to a 95% confidence level—values commonly used in practice [ref.]. This calculation is based on a frequentist framework. Since multiple outcomes are considered across the study's research objectives, we calculate the required sample size for each and adopt the largest as the final sample size to ensure adequate power for all analyses.

We employ a disproportional allocation strategy to ensure reliable state-level estimates. Specifically, each of the 23 states is assigned a minimum of 20 records, totaling 460 records (20×23). The remaining sample size is then distributed proportionally based on each state's share of the target population, and added to the minimum allocation. This approach guarantees adequate representation, even for smaller states. The final total sample size is 1,530. To address the imbalance introduced by disproportional allocation, sampling weights are incorporated into the analysis to improve the representativeness of the overall population. The sampling weight for each state is calculated as the ratio of the state's total number of records in the target population to the number of sampled records from that state.

To assess representativeness, we calculate the standardized proportion difference—a metric used to compare differences in proportions between two populations, scaled by a measure of variability. Smaller values indicate better representativeness of the sample relative to the target population [ref]. Figure 1 illustrates the reduction in sample size resulting from the data filtering and sampling process.

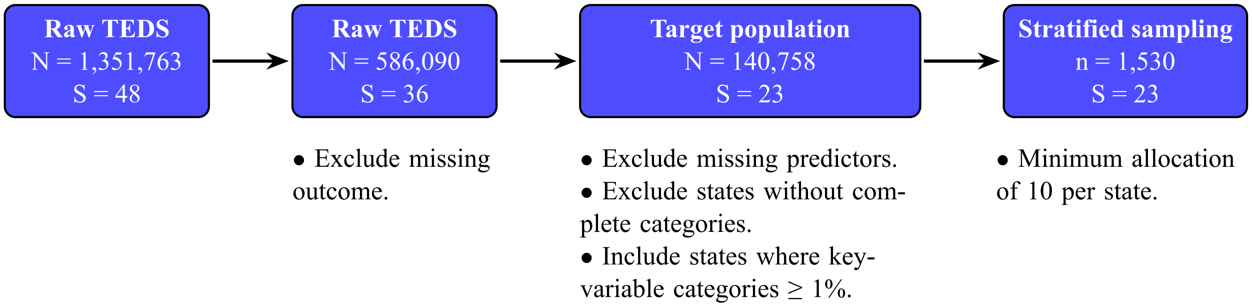
Raw TEDS
N = 1,351,763
S = 48
Raw TEDS
N = 586,090
S = 36
Target population
N = 140,758
S = 23
Stratified sampling
n = 1,530
S = 23
• Exclude missing outcome.
• Exclude missing predictors.
• Exclude states without complete categories.
• Include states where key-variable categories ≥ 1%.
• Minimum allocation of 10 per state.

Figure 1: Roadmap of data filtering and stratified sampling for TEDS analysis.

3 Analyses

Three main objectives are addressed based on the analysis of the stratified samples. As the survey units are admissions, multiple records could potentially originate from a single individual, highlighting the correlation present. However, TEDS maintains confidentiality by not providing subject IDs, which limits our ability to capture within-subject correlation. Nonetheless, TEDS supplies the state ID for each unit, allowing us to partially capture within-client correlation through within-state correlation. However, this is constrained to admissions occurring within the same state, as multiple admissions from an individual across different states are considered independent.

The variables relevant to the research questions include four primary outcomes: wait time to receive treatment, length of stay during the treatment episode, an indicator of treatment completion, and primary drug use at both admission and discharge follow-ups. To address the research objectives associated with these outcomes, we focus on the following key predictors: state indicator, referral source, and alcohol use at both admission and discharge. A summary of the key covariates—some of which are used to derive these predictors—is presented in Table 1.

3.1 Evaluation of regional pattern of WTT and LOS

We examine the regional patterns of wait time to treatment (WTT) by evaluating the distribution of its categories across U.S. states. Originally, WTT is a five-category variable. However, since only a small proportion of individuals in the three longest wait-time categories experienced a delay of eight days or more in accessing substance use treatment (approximately 2.2%), these categories are combined to form

Table 1: Summary of 10 primary covariates selected for their alignment with the research objectives.

	Overall
	(N=1530)
Age	
Mean (SD)	38.2 (12.1)
Median [Min, Max]	37.0 [13.0, 70.0]
Gender	
female	537 (35.1%)
male	993 (64.9%)
Race	
black	238 (15.6%)
others	190 (12.4%)
white	1102 (72.0%)
Alcohol reported at admission	
no	832 (54.4%)
yes	698 (45.6%)
Substance use at admission	
depressants	533 (34.8%)
opioids and stimulants	856 (55.9%)
others	141 (9.2%)
Frequency of use at admission	
daily use	629 (41.1%)
no use	497 (32.5%)
some use	404 (26.4%)
Mental and substance use disorders	
no	847 (55.4%)
yes	683 (44.6%)
Previous substance use treatment	
no	482 (31.5%)
yes	1048 (68.5%)
Referral source	
care provider	222 (14.5%)
community referral	110 (7.2%)
Court/criminal justice referral	485 (31.7%)
individual	713 (46.6%)
Education	
college/university	358 (23.4%)
highschool	1119 (73.1%)
lower than high school	53 (3.5%)

a three-category ordinal outcome: 0, 1–7, and 8 or more wait days. To analyze WTT, we employ an ordered logistic regression model within a Bayesian framework. Covariate selection for model adjustment is informed by four primary factors.

- Demographic factors (age, gender, race and education) reflect the population differences that affect the need for and access to treatment.
- Socio-economic factors (education, health insurance, employment, living arrangements at admission) reflect social and economic barriers that may affect WTT.
- Clinical characteristics (substance use at admission, frequency of use at admission, co-occurring mental and substance use disorders, previous substance use treatment episodes, and medication-assisted opioid therapy) account for the type, severity, and history of substance use, all of which can shape the urgency and availability of treatment.
- Treatment-related factors (referral source and type of treatment service/setting at admission) directly influence access pathways. Together, these covariates help control for variation.

We begin by exploring the variation in WTT and LOS across states as a preliminary evaluation, without adjusting for covariates. Specifically, we calculate the state-level proportions of WTT categories and the mean LOS. These descriptive measures form the basis for assessing geographic variation. As shown in Figure 2 (left), the proportion of WTT categories varies across the 23 included states. This observed variation suggests the need to account for between-state heterogeneity. Accordingly, we fit a model with a random intercept to capture state-level variability in WTT.

The model is formulated as follows.

$$\text{logit}(P(Y_{is} \leq j)) = \kappa_j - \eta_{is}, \text{ where } \eta_{is} = \beta_0 + \boldsymbol{x}_{is}'\boldsymbol{\beta}_1 + u_s,$$

$$Y_{is} = \begin{cases} 1, & \text{if } \eta_{is} \leq \kappa_1 \\ 2, & \text{if } \kappa_1 < \eta_{is} \leq \kappa_2 \\ 3, & \text{if } \eta_{is} > \kappa_2. \end{cases} \tag{1}$$

$$(j = 1, 2;\ i = 1, 2, \ldots, s_i;\ \ s = 1, 2, \ldots, 23)$$

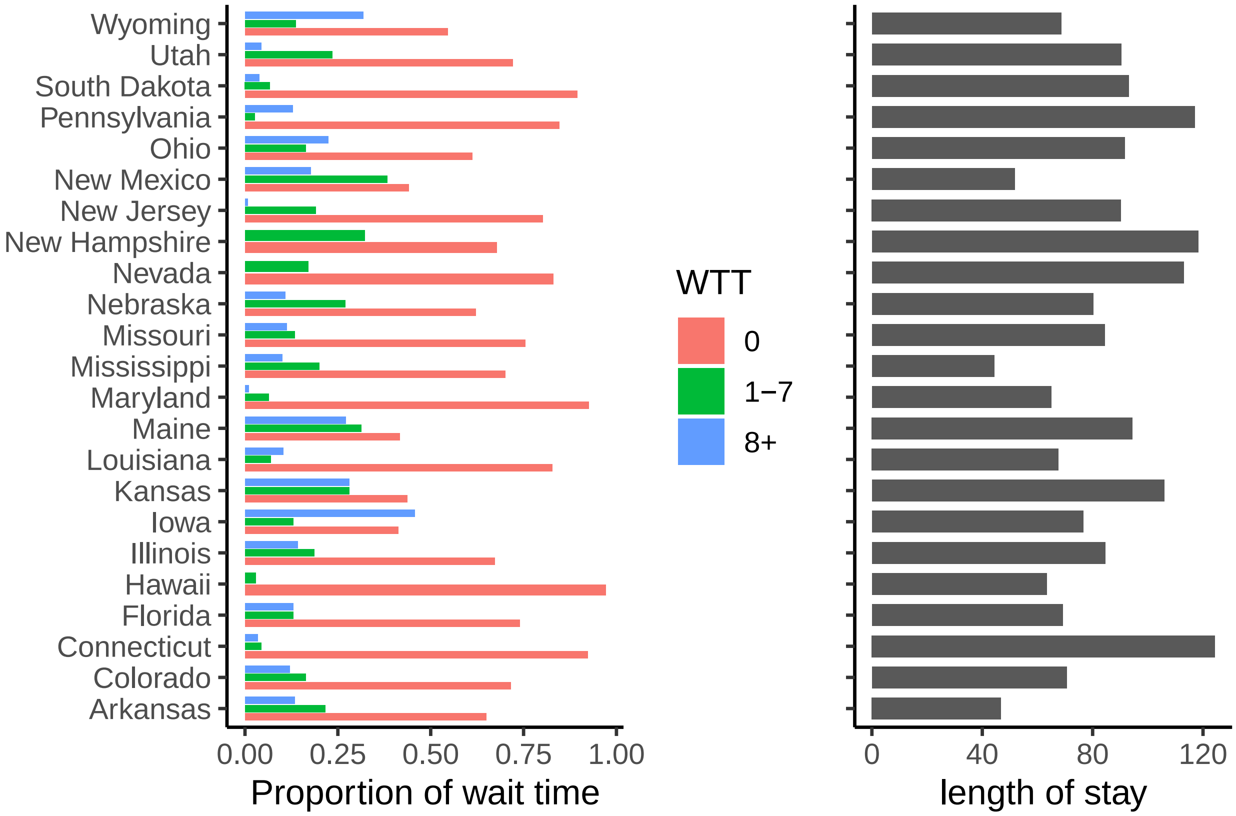

Figure 2: State-by-State proportion of wait time to treatment and the mean of treatment LOS.

where $P(Y_{is} \leq j)$ is the cumulative probability that admission i in state s falls into category j or below; κ_j is the threshold j for category j. $\boldsymbol{x}_{is}$ is a vector of covariates of admission i in state s. β_0 and $\boldsymbol{\beta}_1$ is fixed intercept and coefficients, corresponding to $\boldsymbol{x}_{ij}$. u_s is a random intercept of state s. The model is fitted based on the Bayesian framework with the prior distributions defined as follows.

$$\begin{aligned}
&\beta_0 \sim N(0, 5^2),\ \boldsymbol{\beta}_1 \sim N(\boldsymbol{0}, 5^2\boldsymbol{I}),\\
&u_j \sim N(0, \sigma_u^2),\ \sigma_u \sim N_+(0, 5^2),\\
&\kappa_j \sim N(0, 5) \text{ with constraint } \kappa_1 < \kappa_2.
\end{aligned} \tag{2}$$

The model assumes linearly additive fixed effects and a random intercept on the logit scale of the cumulative probability. A normal distribution is assigned to the threshold parameters κ_j, reflecting moderate uncertainty about the probability associated with each category. The term u_s represents a state-level random effect, capturing variation in WTT across states. As such, u_s serves as the key parameter for addressing the first research question. Specifically, we compute $\exp(u_j - u_{\text{reference}})$, which represents the odds ratio for the cumulative probability of WTT in state j relative to a predefined reference state, after adjusting for covariates.

The investigation of the regional pattern of LOS is conducted using a linear mixed-effects model. Although LOS is originally a categorical variable comprising 37 groups, it is treated as a continuous variable by assigning each observation the midpoint of its corresponding category. To account for regional variation, the model includes a state-level random intercept, motivated by the graphical evidence of between-state differences shown in Figure 2 (right). This random effect serves as the key parameter for assessing the spatial structure of LOS. The model also incorporates the same set of covariates used in the WTT model. However, because some of these covariates are measured at both admission and discharge, we use their updated values to reflect the status at the end of the treatment episode. The model is formulated as follows.

$$y_{is} = \beta_0 + \boldsymbol{x}_{is}'\boldsymbol{\beta}_1 + u_s + \epsilon_{is}, \quad (i = 1, 2, \ldots, s_i;\ s = 1, 2, \ldots, 23) \tag{3}$$

This model assumes linearly additive fixed effects, and the random effect captures baseline differences in LOS across states. We continue to use the same prior distributions specified in Equation 2. The regional pattern of LOS is assessed based on u_s, which represents the deviation in LOS for state s relative to the overall mean, after adjusting for covariates.

3.2 Evaluation of the predictor referral source of treatment completion

TEDS provides information on treatment completion through a seven-category variable indicating the reason for discharge. For the purpose of this analysis, we dichotomize this variable, focusing solely on treatment completion as the event of interest. Treatment completion is then regressed on referral source, originally a seven-category variable that identifies the person or agency referring the client to treatment. Due to the small sample sizes in some categories, we consolidate them to form a new four-category predictor, comprising: (1) individual (including self-referral), (2) care providers, (3) community referral, and (4) court/criminal justice referral. The set of covariates included for adjustment in the model is consistent with those used in the previous models.

The model is expressed below.

$$
\begin{aligned}
&Y_{is} \sim Bern(p_{is}), \\
&\text{logit}(p_{is}) = \beta_0 + \boldsymbol{x}'_{is}\boldsymbol{\beta}_1 + u_s. \\
&(i = 1, 2, \ldots, s_i;\ \ s = 1, 2, \ldots, 23)
\end{aligned}
\tag{4}
$$

The same prior distributions (Figure 2) are assigned to parameters in this model. The key parameter to address this research question is the coefficient corresponding to the referral source. We compare the effect on the treatment completion of each category to the reference group (individual). The significant difference in contribution of each group to the outcome implies the evidence of the association between the referral source and treatment completion. Thus, the referral source potentially predict treatment completion reliably.

3.3 Evaluation of the association between drug and alcohol use over admission and discharge times

The association between drug and alcohol use is evaluated using a logistic regression model. The outcome is a binary variable indicating drug use, and the primary predictor is a binary indicator of alcohol use. Both variables are measured at two follow-up time points, making them time-varying. The model also includes two additional time-varying categorical predictors: type of treatment service/setting and employment status. To account for potential confounding, the model adjusts for a set of time-invariant covariates, including age, gender, race, education, co-occurring mental health conditions, referral source, previous treatment episodes, and baseline employment status. These covariates have been commonly used in related previous studies and are likely to influence the relationship between drug and alcohol use.

The model also includes an indicator for time point and its interaction with alcohol use to capture potential temporal variation in the association. We do not include a random effect to account for within-individual correlation across observations, as only two time points are available per subject. With such limited longitudinal depth, modeling observation-level dependence may not be statistically reliable and could unnecessarily complicate model estimation. However, a random intercept is included to account for regional variation in drug use across states.

The model specification is analogous to that presented in Equation 4, and the same prior distributions are assigned to the model parameters. To address the third research question, we focus on the coefficients corresponding to alcohol use and its interaction with the time indicator. The coefficient for alcohol use, on the exponential scale, represents the odds ratio of drug use between individuals who did and did not use alcohol. The interaction term reflects how this association changes over time, specifically capturing the difference in the effect of alcohol use on drug use between the two follow-up points. Thus, while the main effect of alcohol use reflects its overall association with drug use, the interaction term captures changes in this relationship over the course of the treatment period.

Within the Bayesian framework, we address the research questions by evaluating the parameters of interest based on their posterior distributions. The models are estimated with sampling weights incorporated into the likelihood, ensuring that the posterior distributions reflect the target population rather than the sample alone. This approach enhances the accuracy of population-level inference and improves the generalizability of the findings.

All models are implemented in the R environment using `Stan`, which performs full Bayesian inference via Hamiltonian Monte Carlo. For each model, 4,000 posterior samples were drawn following 5,000 warm-up iterations. Convergence was assessed using the potential scale reduction statistic $\hat{R}$, with the value below 1.1 indicating satisfactory convergence.

4 Results

4.1 The spatial patterns of WTT and LOS

Arkansas is selected as the reference state, as its corresponding random intercept is approximately zero. We compare the odds ratios for the cumulative probability of WTT between each of the other states and this reference. Figure 3 (left) displays the posterior means and 95% highest density intervals (HDIs) of the state-level odds ratios. Non-overlapping or distinct HDIs suggest meaningful variation in the cumulative probability of WTT across states. These results provide evidence of regional heterogeneity in WTT.

An analogous result is observed for LOS. Figure 3 (right) presents the posterior means and 95% highest density intervals (HDIs) of LOS across states. While the HDIs of some states overlap, clear differences remain among others, providing sufficient evidence of regional heterogeneity in LOS.

4.2 Association between the referral source and treatment completion

Table 2 presents the posterior means and 95% HDIs of the odds ratios for treatment completion, comparing the three other referral source categories to individual referral, after adjusting for covariates. The results provide statistical evidence that referrals from care providers and court/criminal justice systems are associated with a higher likelihood of treatment interruption, while community referrals are positively associated with treatment completion. This is supported by the 95% HDIs, which do not include the null value of 1. These findings suggest that referral source is a meaningful and reliable predictor of treatment completion.

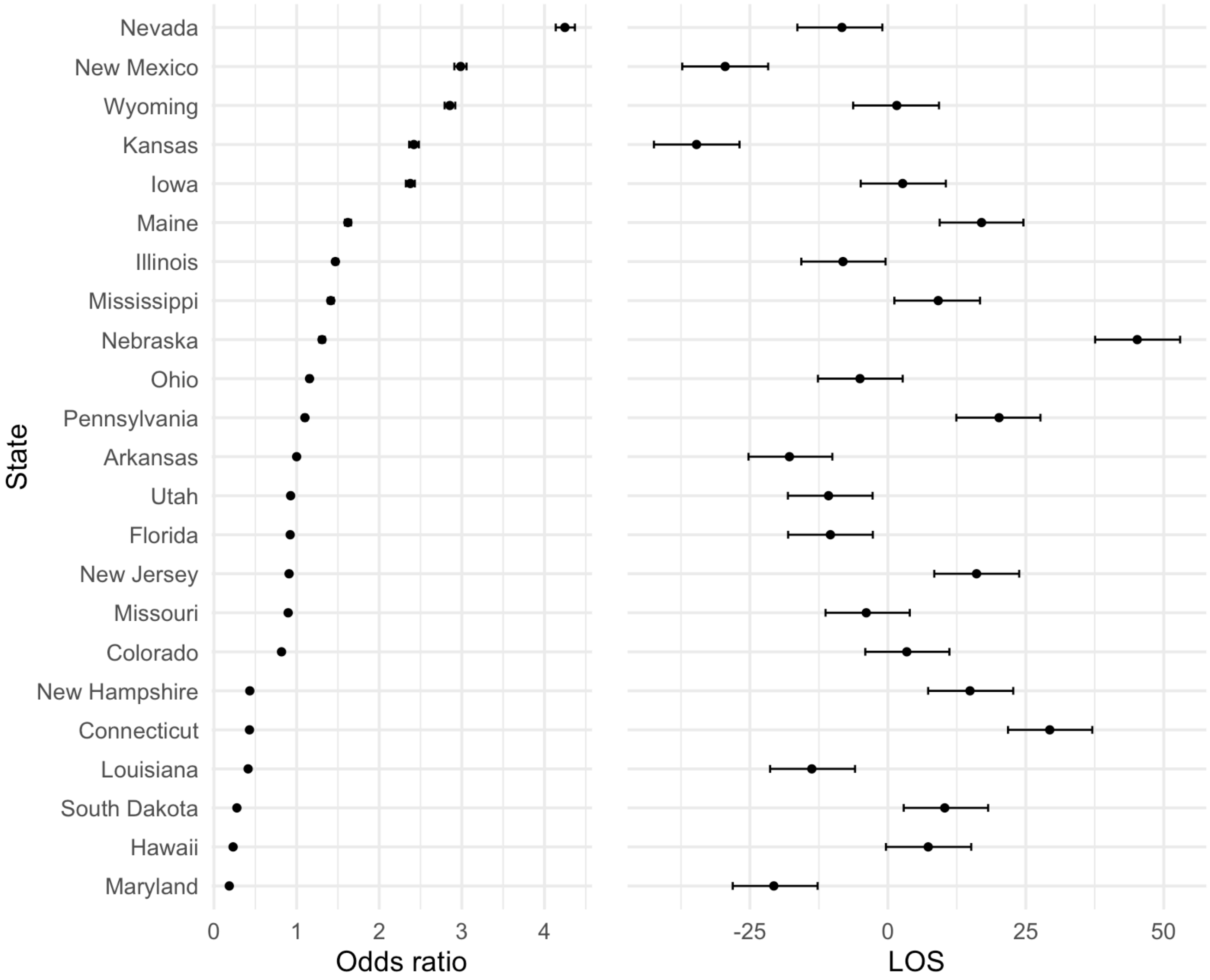

Figure 3: Mean and 95% HDI of the odds ratio associated with the cumulative probability of wait time between the states and Minnesota (the reference state) [left], and the random effects capturing the unique deviation in average LOS for each state from the overall mean, after accounting for fixed effects [right].

Table 2: Mean and 95% HDI of odds ratios associated with treatment completion between referral sources, of which the reference group is individual (includes self-referral).

variable	mean	lower	upper
Care provider	0.93	0.92	0.94
Community referral	1.21	1.20	1.22
Court/criminal justice referral	0.93	0.92	0.94

Table 3: Mean and 95% HDI of the coefficients corresponding to primary covariates.

variable	mean	lower	upper
alcohol use	<0.001	<0.001	<0.001
discharge	<0.001	<0.001	<0.001
alcohol use x discharge	0.35	<0.001	1.5

4.3 Association between drug use and alcohol use at admission and discharge time

The association between drug use and alcohol use is assessed using the odds ratio comparing drug use between individuals who consume alcohol and those who do not. The posterior mean and 95% highest density interval (HDI) of the odds ratio are close to zero (Figure 3), suggesting a strong negative association—that is, individuals who consume alcohol are substantially less likely to use other drugs. This association remains consistent at discharge. The interaction term between alcohol use and time has a 95% HDI of $[< 0.001, 1.5]$, which includes the null value of 1, indicating that there is no clear evidence of a change in this relationship over time. These results provide strong support for a persistent inverse association between alcohol and drug use throughout the treatment period.

5 Discussion

We addressed three research questions by fitting models that examine the associations of interest using TEDS data. First, we identified regional heterogeneity in both WTT and LOS, as reflected by differences in the odds ratios associated with their cumulative probabilities across states. Second, we found that referral source is an important predictor of treatment completion; specifically, referrals from care providers, community sources, and the court/criminal justice system were associated with meaningful differences in completion rates compared to individual referrals, based on the posterior distributions. Finally, we observed a persistent negative association between drug use and alcohol use throughout the treatment episode.

Nonetheless, the conclusions drawn from this study should be interpreted with caution due to several limitations. First, the definition of admission varies across states, which may compromise the validity of state-level comparisons based on admissions data. This limitation directly affects the first research objective, as the corresponding analysis used admissions as the unit of analysis. As a result, inconsistencies in how admissions are defined across states may have introduced ambiguity into the comparisons.

Second, the results should not be generalized to the national level. The target population is based on a reduced version of TEDS, derived through inclusion and exclusion criteria designed to ensure the validity of records and completeness of outcomes and key covariates. The resulting dataset includes only 23 of the 48 states originally reported in TEDS; therefore, the conclusions are applicable only to the included states. Within this target population, bias may arise due to the use of complete case analysis. To avoid incorporating invalid records, all cases with missing data were excluded. However, this approach also removed valid records with partial missingness. If the data are not missing completely at random (MCAR), complete case analysis may introduce selection bias due to a non-representative sample. Thus, the assumption of MCAR is necessary for the validity of the results. Although this is a strong assumption, it supports the use of a dataset comprising only complete and valid records.

As the analysis was conducted on sampled data, biases inherent to the sampling process introduce additional limitations. Specifically, we employed stratified sampling with minimum allocation, which results in a disproportionate sampling design. Although sampling weights were incorporated into model fitting to mitigate the effects of non-representativeness, the effectiveness of this adjustment depends on the properties of the weights themselves. We evaluated the distribution of the sampling weights using descriptive statistics and found them to be asymmetric and left-skewed. This pattern reflects over-sampling in smaller states and under-sampling in larger states—where a high weight corresponds to a low sampling fraction and thus indicates under-sampling. Consequently, small states are overrepresented due to the minimum allocation (resulting in lower weights), while larger states are underrepresented (resulting in higher weights). Additionally, the wide range of the weights indicates substantial variability in sampling fractions. This may reduce the influence of states with low weights in the analysis, potentially underestimating differences between subgroups of interest within those states.

Sampling bias could be mitigated by modifying the sampling design, such as increasing the sample size to reduce sampling error. However, a larger sample would require substantially more time to achieve model convergence, which was not feasible given the time constraints of the current study. Therefore, we recommend that the present findings be interpreted as exploratory or as part of a pilot analysis, rather than as definitive or conclusive. These results offer a foundation for future research and highlight the need for validation using a larger or more balanced sampling design. This preliminary analysis provides early insights that can inform future studies aimed at reducing sampling error and narrowing credible intervals through increased sample size.

References