Examining the Causal Impact of Recreational Marijuana Legalization on Psychotic Disorder-Related Hospitalizations:

Evidence from Logistic Regression and Propensity Score Matching

Alina Hordiienko

Minerva University, San Francisco, CA, USA

1. Introduction

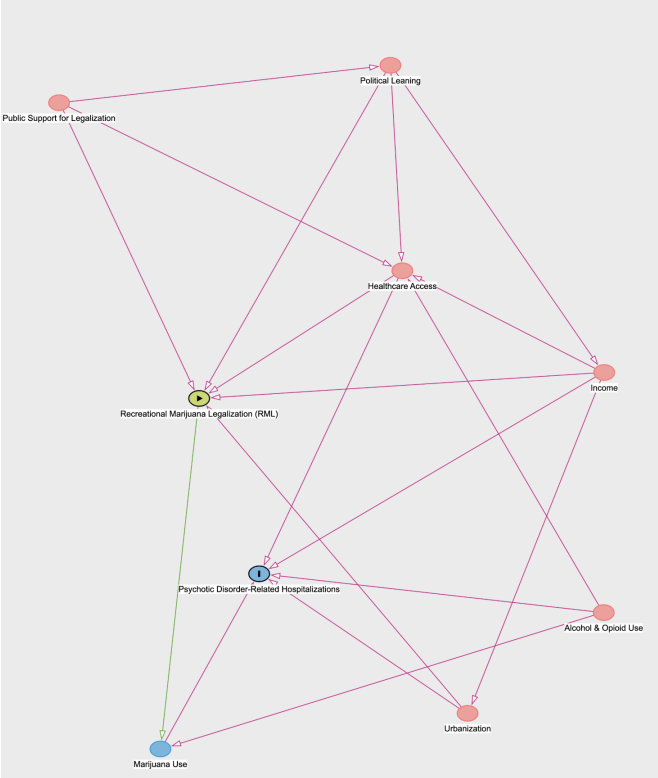
I investigate the causal effect of state-level recreational marijuana legalization on the incidence of psychotic disorder-related hospitalizations. In this analysis, I use data spanning from 2013 to 2017—a period during which several U.S. states, including Washington, Colorado, Oregon, California, Massachusetts, Nevada, and Maine, enacted recreational marijuana legalization policies. This time frame was selected because the staggered policy implementation during these years provides natural variation in legalization status across states. By integrating individual-level clinical data from the MH-CLD database with state-level socioeconomic and demographic indicators, I am able to control for various confounding factors that may influence hospitalization rates for psychotic disorders.

Addressing this question is important from a societal perspective as it provides evidence for evaluating the public health impact of recreational marijuana legalization. Empirical studies present mixed findings: for example, Athanassiou et al. (2023) report that legalization might lead to modest improvements in public health metrics through harm reduction, while Walker et al. (2023) highlight concerns that increased cannabis availability may be associated with a higher incidence of psychosis in vulnerable populations. Determining whether legalization is associated with changes in psychotic disorder-related hospitalizations is critical for policymakers, as it informs the balance between the potential benefits of reduced opioid use and other harm reduction effects against the risks of adverse mental health outcomes.

2. Causal Dynamics

I define the treatment variable as the state-level recreational marijuana legalization status. This variable is measured as a binary indicator, taking the value 1 when a state has implemented recreational marijuana legalization and 0 otherwise in a given year. The outcome variable is psychotic disorder-related hospitalizations, represented in the MH-CLD dataset by the indicator SCHIZOFLG. Although a continuous count of hospital admissions would ideally capture the prevalence of psychotic disorders, the available data record this outcome as a binary indicator on an individual level (1 if schizophrenia or other psychotic disorders were reported in the primary, secondary, or tertiary mental health diagnosis field in the given year, 0 if not). This measurement choice reflects data limitations and influences the selection of the logistic regression framework in my analysis.

Figure 1. *DAG*



In constructing the causal graph, I include a set of variables that may confound the relationship between recreational marijuana legalization (RML) and psychotic disorder-related hospitalizations. Demographic factors—age, gender, ethnicity, and education—can shape both a person's vulnerability to psychosis and the likelihood of appearing in MH-CLD records. At the state level, median

household income and urbanization serve as proxies for broader economic and infrastructural conditions that influence policy adoption and hospitalization rates. Additionally, I include baseline measures of mental illness (MentalIllnessYr) and marijuana use (MarijuanaUseYr) to capture the pre-legalization prevalence of psychiatric conditions and substance use, consistent with research highlighting their importance in shaping policy outcomes and mental health trends (Athanassiou et al., 2023; Walker et al., 2023).

Backdoor paths are critical to address in observational studies because they represent alternative, non-causal routes linking the treatment (RML) and the outcome (hospitalizations). If not blocked, these paths can introduce confounding bias. For example, state income may simultaneously affect the likelihood of adopting legalization and the quality of healthcare, thereby confounding the association between RML and hospitalizations. By identifying and adjusting for these variables, I aim to mitigate bias and isolate the effect of legalization.

Based on the causal graph, the key adjustment set includes demographic variables (age, gender, ethnicity, education), economic indicators (median household income), urbanization, and baseline measures of mental illness and marijuana use. Some constructs in the DAG, such as healthcare access, political leaning, and public support for legalization, are not directly observed in my dataset and therefore do not appear in the final regression. Nevertheless, by controlling for the available proxies in my models, I strive to block the major backdoor paths and produce a more reliable estimate of RML's effect on psychotic disorder-related hospitalizations.

3. Data

I use multiple data sources to construct the dataset. The primary source for the outcome variable is the MH-CLD dataset, which provides individual-level clinical data on psychotic disorder-related hospitalizations from state mental health agencies. For state-level socioeconomic and demographic covariates, I incorporate data from the American Community Survey (ACS) and the P2 Urban/Rural Census data. Information on substance use prevalence is drawn from the National Survey on Drug Use and Health (NSDUH), and details on recreational marijuana legalization status are obtained from the PDAPS dataset. Additional data from sources such as KFF complement the dataset by providing healthcare-related indicators. Together, these sources allow me to integrate clinical and contextual information spanning the years 2013 to 2017. A complete variable codebook is provided in Appendix A.

Despite the comprehensive nature of these datasets, several shortcomings exist. The MH-CLD data, for example, only include individuals already engaged with state mental health services, which limits the generalizability of findings to the broader population. Additionally, the outcome variable is recorded as a binary indicator, which may not fully capture the variation in the prevalence or severity of psychotic disorder-related hospitalizations. There are also concerns regarding missing data and potential mismeasurement; certain confounders, such as healthcare access and pre-legalization cannabis use, are only indirectly measured or represented by proxies. These issues highlight the need for caution when interpreting the estimated effects, as unobserved or poorly measured variables could bias the results.

4. Analysis

4.1. Regression Analysis

Regression Model

I estimate a logistic regression model to examine the effect of recreational marijuana legalization (Legalized) on psychotic disorder-related hospitalizations (SCHIZOFLG). The model includes the set of confounders identified in the previous section: AGE, GENDER, ETHNIC, EDUC, MentalIllnessYr, MarijuanaUseYr, MedianIncome, and UrbanPop. Formally:

$$\text{logit}(\Pr(\text{SCHIZOFLG}=1)) = \beta_0 + \beta_1\text{Legalized} + \beta_2\text{AGE} + \beta_3\text{GENDER} + \beta_4\text{ETHNIC} + \beta_5\text{EDUC} + \beta_6\text{MentalIllnessYr} + \beta_7\text{MarijuanaUseYr} + \beta_8\text{MedianIncome} + \beta_9\text{UrbanPop}.$$

Figure 1. Logistic

```
glm(formula = SCHIZOFLG ~ Legalized + AGE + GENDER + ETHNIC +
    EDUC + MentalIllnessYr + MarijuanaUseYr + MedianIncome +
    UrbanPop, family = binomial, data = final_df)

Coefficients:
                Estimate Std. Error  z value Pr(>|z|)
(Intercept)    -4.775e+00  2.136e-01  -22.359  < 2e-16 ***
Legalized      -7.952e-01  1.191e-02  -66.739  < 2e-16 ***
AGE5            7.775e-01  2.432e-02   31.970  < 2e-16 ***
AGE6            9.109e-01  2.281e-02   39.933  < 2e-16 ***
AGE7            1.008e+00  2.264e-02   44.531  < 2e-16 ***
AGE8            1.104e+00  2.284e-02   48.363  < 2e-16 ***
AGE9            1.143e+00  2.311e-02   49.471  < 2e-16 ***
AGE10           1.217e+00  2.290e-02   53.140  < 2e-16 ***
AGE11           1.351e+00  2.261e-02   59.772  < 2e-16 ***
AGE12           1.525e+00  2.291e-02   66.583  < 2e-16 ***
AGE13           1.646e+00  2.428e-02   67.796  < 2e-16 ***
AGE14           1.577e+00  2.479e-02   63.631  < 2e-16 ***
GENDER2        -9.898e-01  7.271e-03 -136.119  < 2e-16 ***
ETHNIC2         6.411e-02  4.578e-02    1.400   0.1614
ETHNIC3         1.613e-01  2.170e-02    7.431 1.08e-13 ***
ETHNIC4         1.340e-01  1.867e-02    7.177 7.11e-13 ***
EDUC2          -3.232e-01  1.449e-01   -2.231   0.0257 *
EDUC3          -9.074e-02  1.446e-01   -0.628   0.5302
EDUC4          -3.641e-01  1.444e-01   -2.521   0.0117 *
EDUC5          -7.950e-01  1.446e-01   -5.498 3.85e-08 ***
MentalIllnessYr -3.722e+00  1.061e-01  -35.082  < 2e-16 ***
MarijuanaUseYr   1.516e+01  2.716e-01   55.818  < 2e-16 ***
MedianIncome     2.099e-05  8.292e-07   25.318  < 2e-16 ***
UrbanPop        -1.723e-01  1.612e-01   -1.069   0.2851
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 556424  on 614231  degrees of freedom
Residual deviance: 513316  on 614208  degrees of freedom
  (3148155 observations deleted due to missingness)
AIC: 513364

Number of Fisher Scoring iterations: 5
```

Although logistic regression does not rely on the same assumptions as ordinary least squares (OLS), I perform certain diagnostic checks (residual plots, Q–Q plot, and VIF) because they are required by the assignment instructions. Below, I discuss these checks while acknowledging their limited applicability in a logistic framework.

In logistic regression, the assumption is that predictors have a linear relationship with the log-odds of the outcome, not with the raw outcome itself. Nevertheless, I plotted Pearson

residuals versus predicted values (y_hat) to look for any large systematic pattern that might indicate mis-specification. As shown in Figure 2, the residuals do not form a typical cloud around zero, but instead appear in distinct clusters above and below the dashed line. This pattern reflects the binary nature of the outcome rather than a violation akin to heteroskedasticity in OLS. There is no strong indication of a systematic trend in these clusters, suggesting that no major mis-specification is evident.

Figure 2. *Residuals vs Predicted Values*

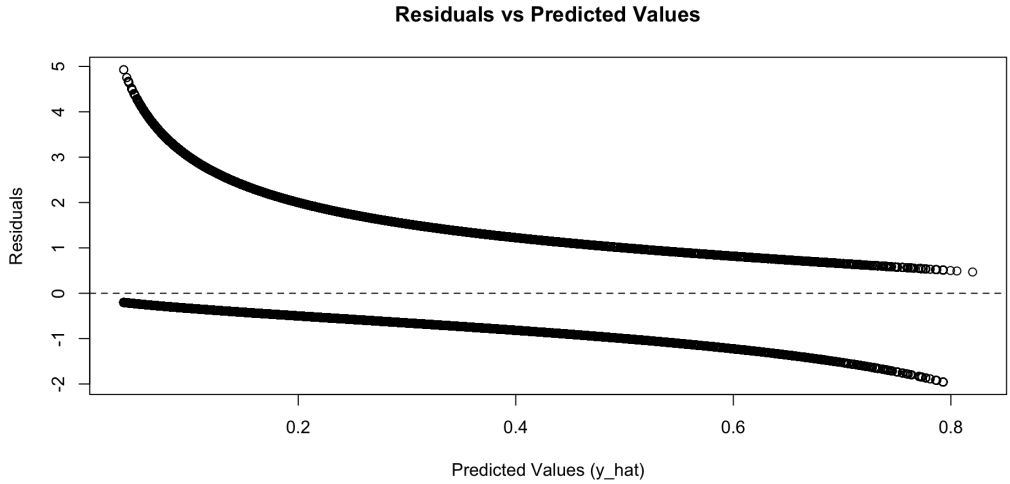
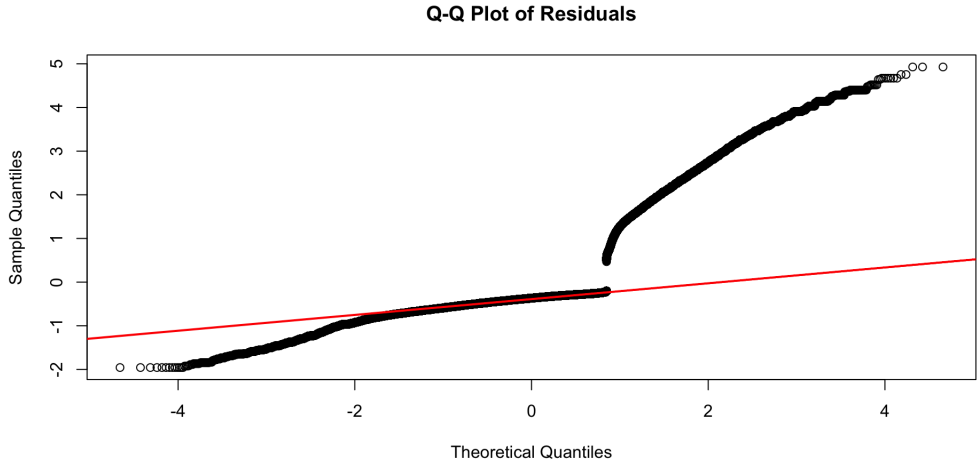


Figure 3. *Q–Q Plot of Residuals*



The Q–Q plot (Figure 3) shows that the residuals deviate substantially from the 45-degree line, especially in the tails. However, in logistic regression, residuals are not expected to follow a normal distribution. Consequently, large deviations from normality are typical and do not necessarily indicate a problem. Nonetheless, performing this check helps identify extreme observations or outliers that could influence the model fit.

To assess whether any predictors are highly correlated, I calculated variance inflation factors (VIF) . The results, shown below, indicate that most variables have VIF values comfortably below 10. These values suggest that severe multicollinearity is not present. While Legalized and MentalIllnessYr have slightly higher VIFs, they remain below the conventional threshold of 10, indicating that the model coefficients are relatively stable and interpretable.

Figure 4. *VIF Results*

	GVIF	Df	GVIF^(1/(2*Df))
Legalized	6.647978	1	2.578367
AGE	1.097745	10	1.004674
GENDER	1.014552	1	1.007250
ETHNIC	1.139881	3	1.022060
EDUC	1.178664	4	1.020760
MentalIllnessYr	5.663387	1	2.379787
MarijuanaUseYr	3.274961	1	1.809685
MedianIncome	2.053026	1	1.432839
UrbanPop	2.904424	1	1.704237

4.2. Matching Analysis

Because this is an observational study, I also implement a propensity score matching approach to further address potential confounding. Specifically, I estimate a propensity score for each individual based on the same covariates used in the logistic regression (AGE, GENDER, ETHNIC, EDUC, MentalIllnessYr, MarijuanaUseYr, MedianIncome, UrbanPop). I then apply nearest neighbor matching to pair each treated observation (Legalized = 1) with a comparable untreated observation (Legalized = 0).

Figure 5. *Matched Logistic Regression Model*

```
glm(formula = SCHIZOFLG ~ Legalized + AGE + GENDER + ETHNIC +
    EDUC + MentalIllnessYr + MarijuanaUseYr + MedianIncome +
    UrbanPop, family = binomial, data = matched_data)

Coefficients:
                  Estimate Std. Error  z value Pr(>|z|)
(Intercept)     -5.285e+00  2.777e-01  -19.030  < 2e-16 ***
Legalized       -9.924e-03  2.452e-02   -0.405  0.68571
AGE5             7.161e-01  3.200e-02   22.376  < 2e-16 ***
AGE6             8.464e-01  3.000e-02   28.210  < 2e-16 ***
AGE7             9.936e-01  2.963e-02   33.531  < 2e-16 ***
AGE8             1.093e+00  2.995e-02   36.495  < 2e-16 ***
AGE9             1.176e+00  3.015e-02   39.024  < 2e-16 ***
AGE10            1.254e+00  2.989e-02   41.971  < 2e-16 ***
AGE11            1.411e+00  2.948e-02   47.862  < 2e-16 ***
AGE12            1.638e+00  3.003e-02   54.540  < 2e-16 ***
AGE13            1.748e+00  3.193e-02   54.757  < 2e-16 ***
AGE14            1.658e+00  3.247e-02   51.082  < 2e-16 ***
GENDER2         -1.021e+00  9.656e-03 -105.766  < 2e-16 ***
ETHNIC2          1.490e-01  6.284e-02    2.371  0.01772 *
ETHNIC3          1.013e-01  3.257e-02    3.109  0.00188 **
ETHNIC4          1.300e-01  2.948e-02    4.408 1.04e-05 ***
EDUC2           -3.282e-01  1.616e-01   -2.031  0.04230 *
EDUC3           -1.044e-02  1.611e-01   -0.065  0.94830
EDUC4           -3.272e-01  1.608e-01   -2.035  0.04189 *
EDUC5           -7.463e-01  1.612e-01   -4.631 3.63e-06 ***
MentalIllnessYr -1.093e+01  2.335e-01  -46.815  < 2e-16 ***
MarijuanaUseYr   1.507e+01  3.432e-01   43.921  < 2e-16 ***
MedianIncome     3.373e-05  1.014e-06   33.248  < 2e-16 ***
UrbanPop         8.585e-01  1.897e-01    4.526 6.01e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 308626  on 312029  degrees of freedom
Residual deviance: 279976  on 312006  degrees of freedom
AIC: 280024

Number of Fisher Scoring iterations: 5
```

After matching, I re-run the logistic regression on the matched dataset. The matched analysis shows a smaller and statistically insignificant coefficient on Legalized compared to the full-sample regression. This indicates that any initial association observed in the unmatched sample may have been driven by differences in baseline characteristics between states that legalized marijuana and those that did not. By creating a more balanced sample, matching provides a clearer picture of the treatment effect and reduces the bias that can arise from confounding variables.

While the diagnostic tests are more relevant to OLS, they still offer a general sense of whether extreme values, collinearity, or gross model mis-specification might be influencing the logistic regression. Ultimately, the matching approach helps mitigate confounding, and the logistic model provides an appropriate framework for the binary outcome despite not meeting classical OLS assumptions for residual behavior.

5. Discussion

In the unmatched logistic regression, the coefficient on the treatment variable (Legalized) is negative and statistically significant at the 5% level. This implies that, in the full sample, states with recreational marijuana legalization have lower log-odds of psychotic disorder-related hospitalizations compared to non-legalized states. However, this significant association may be driven by pre-existing differences between states rather than a direct causal effect of legalization.

After applying nearest neighbor matching, the re-estimated model yields a near-zero and statistically insignificant coefficient for Legalized. This indicates that, once states are balanced on observed confounders (e.g., demographic, economic, and baseline mental health factors), the apparent negative association diminishes. In practical terms, the matched analysis suggests that recreational marijuana legalization does not have a meaningful impact on the likelihood of psychotic disorder-related hospitalizations.

6. Conclusion

In summary, my analysis examined the impact of recreational marijuana legalization on psychotic disorder-related hospitalizations using two analytical methods: logistic regression on the full sample and propensity score matching followed by logistic regression. The unmatched model indicated a statistically significant negative association between legalization and hospitalizations, suggesting that legalized states have lower odds of psychotic hospitalizations. However, after balancing the covariates through matching, the treatment effect became negligible and statistically insignificant. This contrast implies that

the initial association was likely driven by pre-existing differences in demographic, economic, and baseline mental health factors rather than a direct effect of legalization.

Several limitations constrain the interpretation of these findings. The MH-CLD data only capture individuals already engaged with state mental health services, which limits the generalizability of the results to the broader population. Additionally, the binary measurement of psychotic hospitalizations may oversimplify the underlying health outcomes. The integration of individual-level clinical data with state-level socioeconomic indicators further complicates the analysis. Future studies could benefit from using more representative datasets and exploring alternative modeling approaches, such as hierarchical or multilevel models, to better address the mixed data levels.

Overall, the matched analysis aligns with recent literature that highlights the importance of controlling for confounding factors when assessing the public health impacts of marijuana legalization (Athanassiou et al., 2023; Walker et al., 2023). My results suggest that, after accounting for relevant confounders, recreational marijuana legalization does not exert a significant effect on psychotic disorder-related hospitalizations, a finding that contributes to the ongoing debate in the literature.

References

Athanassiou, M., et al. (2023). *The Clouded Debate: A Systematic Review of Comparative Longitudinal Studies Examining the Impact of Recreational Cannabis Legalization on Key Public Health Outcomes*. Frontiers in Psychiatry. https://doi.org/10.3389/fpsyt.2022.1060656

Census.gov. *Index of /programs-surveys/popest/datasets/2010-2017/state/asrh*. (2017). https://www2.census.gov/programs-surveys/popest/datasets/2010-2017/state/asrh/

monQcle. (2017). *PDAPS - Recreational Marijuana Laws*. Pdaps.org. https://pdaps.org/datasets/recreational-marijuana-laws

SAMHSA. (2025a). *Data Files | CBHSQ Data SAMHSA*. Samhsa.gov. https://www.samhsa.gov/data/data-we-collect/mh-cld/datafiles

SAMHSA. (2025b). *National Survey on Drug Use and Health (NSDUH) State Data Releases*. Samhsa.gov. https://www.samhsa.gov/data/nsduh/state-reports

Status of State Action on the Medicaid Expansion Decision | KFF. (2023, May 8). KFF. https://www.kff.org/affordable-care-act/state-indicator/state-activity-around-expanding-medicaid-under-the-affordable-care-act/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

Walker, M., et al. (2023). *The Effect of Recreational Cannabis Legalization and Commercialization on Substance Use, Mental Health, and Injury: A Systematic Review*. Public Health, 221, 87–96. https://doi.org/10.1016/j.puhe.2023.06.012

10. Appendices

Appendix A: Variable Codebook

State
Description: U.S. state identifier (lowercase)
Data Type: Categorical
Notes: e.g., "california", "oregon"; serves as a geographic identifier

year
Description: Year of observation
Data Type: Numeric
Notes: Values range from 2013 to 2017

SCHIZOFLG
Description: Indicator for psychotic disorder-related hospitalization
Data Type: Binary
Notes: 1 = hospitalization recorded; 0 = no hospitalization

Legalized
Description: Recreational marijuana legalization status
Data Type: Binary
Notes: 1 = legalized; 0 = not legalized; defined by state policy and year

AGE
Description: Age group of the individual
Data Type: Factor
Notes: Recoded into categorical groups (e.g., "5" representing 21–24 years)

GENDER
Description: Gender of the individual
Data Type: Factor
Notes: Typically coded as 1 = male, 2 = female

ETHNIC
Description: Hispanic or Latino origin
Data Type: Factor
Notes: Categories include: Mexican, Puerto Rican, Other Hispanic, Not Hispanic

RACE
Description: Race of the individual
Data Type: Factor
Notes: Categorical variable; may include categories such as White, Black, Asian, etc. (Note: sometimes omitted due to collinearity with ETHNIC)

EDUC
Description: Educational attainment
Data Type: Factor
Notes: Categories: Special education, 0–8, 9–11, 12 (or GED), More than 12

MentalIllnessYr
Description: Estimated number of mental illness cases, adjusted as a proportion
Data Type: Numeric
Notes: Calculated as (cases × 1000) / pop_18plus, representing state-level rate

MarijuanaUseYr
Description: Estimated number of marijuana use cases, adjusted as a proportion
Data Type: Numeric
Notes: Calculated as (cases × 1000) / pop_18plus, reflecting state-level prevalence

MedianIncome
Description: Median household income
Data Type: Numeric
Notes: Reported in U.S. dollars; sourced from ACS data

MeanIncome
Description: Mean household income
Data Type: Numeric
Notes: Reported in U.S. dollars; sourced from ACS data

Insurance
Description: Proportion of individuals with health insurance coverage
Data Type: Numeric
Notes: Derived from ACS data

UrbanPop
Description: Proportion of the state population living in urban areas
Data Type: Numeric
Notes: Derived from P2 Urban/Rural Census data

Medicaid
Description: Medicaid adoption status
Data Type: Binary
Notes: 1 = adopted; 0 = not adopted; sourced from KFF data

pop_18plus
Description: Population aged 18 and older
Data Type: Numeric
Notes: State-level population estimate from census data

Appendix B: Backdoor Paths

- RML ← Public Support for Legalization → Psychotic Hospitalizations
- RML ← Political Leaning → Psychotic Hospitalizations
- RML ← Healthcare Access → Psychotic Hospitalizations
- RML ← Income → Psychotic Hospitalizations
- RML ← Alcohol & Opioid Use → Psychotic Hospitalizations
- RML ← Urbanization → Psychotic Hospitalizations
- RML ← Marijuana Use (pre-legalization) → Psychotic Hospitalizations
- RML ← Income → Urbanization → Psychotic Hospitalizations
- RML ← Income → Healthcare Access → Psychotic Hospitalizations
- RML ← Political Leaning → Income → Psychotic Hospitalizations
- RML ← Public Support for Legalization → Political Leaning → Psychotic Hospitalizations
- RML ← Healthcare Access → Income → Alcohol & Opioid Use → Psychotic Hospitalizations