**RESEARCH PAPER**

# Epidemiological Patterns, Temporal Trends, and Predictive Modeling of Substance Use-Related Mortality and Emergency Care in Mexico (2011–2023): A Population-Based Machine Learning Analysis

**[1]Bogart Yail Marquez, [2]Ada Mitre , [2]Margarita Ramírez-Ramírez , [1]Jose Sergio Magdaleno-Palencia**

[1]Tecnológico Nacional de México, Campus Tijuana
[2]Universidad Autónoma de Baja California

**Correspondence Author**: bogart@tectijuana.edu.mx

**Abstract**

Background: Substance use disorders (SUDs) represent a critical global public health challenge. In Mexico, longitudinal population-based analyses integrating mortality and morbidity surveillance data with advanced predictive modeling remain scarce, limiting the development of evidence-based intervention strategies.

Objective: This study aims to characterize epidemiological patterns of substance use-related mortality and emergency care attendance in Mexico from 2011 to 2023, identify regional clusters of substance profiles, detect significant temporal trends, and develop a predictive model for alcohol-related emergency visits.

Methods: Two national surveillance datasets were analyzed: (i) defunciones_uso_sustancias (n = 12,473 aggregated records; 33 federal entities), capturing ICD-10 F10–F19 mortality, and (ii) data-2026-05-27 (n = 25,051 records; 34 entities), capturing emergency care episodes. Descriptive statistics, Shapiro-Wilk normality tests, Mann-Whitney U tests, Kruskal-Wallis tests, Spearman rank correlations, chi-square analysis with Cramér's V, Spearman temporal trend analysis, Principal Component Analysis (PCA), K-Means clustering (k=4, silhouette=0.337), Isolation Forest anomaly detection (contamination=0.02), and Random Forest regression ($R^2 = 0.847$, RMSE = 23.7) were applied.

Results: Alcohol (F10) was the dominant substance accounting for 91.5% of all recorded deaths (n = 33,383) and 68.8% of emergency visits (n = 369,266). Mortality peaked in the 45–54 age group, whereas emergency episodes concentrated in the 20–34 age bracket, indicating divergent acute versus chronic harm trajectories. A statistically significant male predominance was confirmed in both datasets (mortality: 94.7% male; emergencies: 79.9%; Mann-Whitney $p < 0.0001$; effect $r = -0.84$). Eight of ten substance categories demonstrated statistically significant upward trends in emergency visits (Spearman $r = 0.643–0.978$, $p < 0.05$), with sedatives ($r = 0.978$), multiple drugs ($r = 0.978$), and stimulants ($r = 0.973$) showing the strongest growth. Tobacco exhibited the highest lethality ratio (0.381), while cannabinoids and sedatives showed ratios below 0.002. K-Means clustering identified four geographically coherent state profiles, with the northern border cluster (Baja California, Sinaloa, Colima, Baja California Sur) showing the highest diversification of substances. The Random Forest model explained 84.7% of variance in emergency alcohol visits ($R^2 = 0.847$), with geographic location as the dominant predictor (62.8% importance).

Conclusions: This study provides the most comprehensive machine learning-informed epidemiological analysis of substance use-related harm in Mexico to date. Findings indicate a critical paradox: while alcohol mortality has plateaued, emergency burden from polydrug use is accelerating. Regional heterogeneity demands targeted, state-level public health interventions. The predictive framework developed here offers a replicable methodology for real-time resource allocation in national health systems.

**Keywords:** substance use disorders; ICD-10; epidemiology; machine learning; Random Forest; K-Means clustering; PCA; Mexico; mortality; emergency care; temporal trends; polydrug use

**How to cite this article:** Marquez BY, Mitre A, Ramírez-Ramírez M, Magdaleno-Palencia JS. Epidemiological Patterns, Temporal Trends, and Predictive Modeling of Substance Use-Related Mortality and Emergency Care in Mexico (2011–2023): A Population-Based Machine Learning Analysis. Int J Drug Deliv Technol. 2026;16(64s):460-474. DOI: 10.25258/ijddt.16.64s.45

## 1. Introduction

### 1.1 Global Burden of Substance Use Disorders

Substance use disorders (SUDs) constitute one of the most pressing public health crises of the twenty-first century. According to the Global Burden of Disease Study 2019, SUDs accounted for approximately 50 million disability-adjusted life years (DALYs) globally, with alcohol use disorder alone responsible for over 168 million DALYs when including indirect health consequences (GBD 2019 Diseases and Injuries Collaborators, 2020). The World Drug Report estimated that 296 million people used illicit drugs in 2021, a 26% increase over the previous decade (United Nations Office on Drugs and Crime [UNODC], 2023). This accelerating trend is associated with escalating demand for emergency medical care, increased premature mortality, and substantial socioeconomic costs that disproportionately affect low- and middle-income countries (LMICs) (Degenhardt et al., 2018).

Alcohol use disorder remains the leading cause of SUD-related mortality worldwide. The Global Status Report on Alcohol and Health documented that harmful use of alcohol causes approximately 3 million deaths annually, representing 5.3% of all global deaths (World Health

*Author for Correspondence: bogart@tectijuana.edu.mx

Organization [WHO], 2022). Tobacco, despite declining prevalence in high-income countries, continues to exert disproportionate chronic lethality particularly in LMICs (Reitsma et al., 2021). Emerging trends include the rapid rise of stimulant use—particularly methamphetamine and cocaine—alongside growing polydrug consumption patterns that complicate clinical management and epidemiological surveillance (Peacock et al., 2021; Barratt et al., 2022).

### 1.2 Epidemiological Context in Latin America and Mexico

Latin America presents a complex and heterogeneous epidemiological landscape for SUDs. The Inter-American Drug Abuse Control Commission (CICAD) reported significant increases in cocaine, cannabis, and polydrug use across the region between 2010 and 2020 (Organization of American States, 2022). In Mexico specifically, the Encuesta Nacional de Consumo de Drogas, Alcohol y Tabaco (ENCODAT 2016–2017) documented lifetime prevalence of any illicit drug use at 10.3%, with alcohol dependence criteria met by 5.3% of the adult population (Instituto Nacional de Psiquiatría [INPRFM] et al., 2017). Subsequent ENCODAT data suggest continued escalation, particularly among male populations and younger age cohorts.

Mexico occupies a geopolitically complex position as both a producer and transit country for controlled substances, which creates unique epidemiological dynamics distinct from those observed in other Latin American nations (Strathdee et al., 2012; Medina-Mora et al., 2020). Northern border states exhibit substance use profiles shaped by proximity to U.S. drug markets, including higher rates of heroin and methamphetamine use (Borquez et al., 2019; Garfein et al., 2020), while southern and central states demonstrate patterns dominated by alcohol and inhalant use, reflecting socioeconomic gradients and differential exposure to organized crime-related drug availability (Magis-Rodriguez et al., 2022).

The Sistema de Vigilancia Epidemiológica de las Adicciones (SISVEA), operational since 1990, has provided one of the most sustained sources of substance-related emergency and mortality data in Latin America. However, its analytical potential has been underutilized due to the absence of modern data science methodologies capable of integrating temporal, geographic, and demographic dimensions simultaneously (Fleiz et al., 2021).

### 1.3 Data Science and Machine Learning in Substance Use Epidemiology

The application of machine learning (ML) and data mining techniques to epidemiological datasets has transformed the capacity of public health researchers to identify complex, non-linear patterns that remain invisible to conventional statistical approaches. Supervised learning algorithms such as Random Forest, XGBoost, and Support Vector Machines have demonstrated superior predictive performance relative to logistic regression in predicting substance use treatment outcomes, emergency department utilization, and overdose risk (Cheng et al., 2021; Salazar et al., 2022; Lo-Ciganic et al., 2019). Random Forest, in particular, has become a benchmark method for health data due to its robustness to non-normality, resistance to overfitting, and native feature importance estimation (Breiman, 2001; Biau & Scornet, 2016).

Unsupervised learning techniques—including K-Means clustering and hierarchical clustering—have been applied to identify geographically and demographically coherent risk profiles in substance use surveillance data (Pacula et al., 2020; Krawczyk et al., 2021). Principal Component Analysis (PCA) provides dimensionality reduction capabilities critical for datasets with high intercorrelation among substance categories, as observed in emergency care settings where polydrug presentations are common (Karatzoglou et al., 2022). Isolation Forest anomaly detection has been applied to epidemiological time series to flag data quality issues and genuine epidemiological anomalies (Liu et al., 2012; Goldstein & Uchida, 2016).

Despite these methodological advances, their systematic application to Mexican national surveillance data—particularly in an integrated framework combining mortality and emergency care outcomes—remains absent from the peer-reviewed literature. Existing studies have typically analyzed single data sources, employed only descriptive statistics, and lacked temporal trend analysis with inferential validation (Villatoro-Velázquez et al., 2020; Fleiz et al., 2021). This represents a significant gap given the scale of the problem and the availability of publicly accessible administrative data.

### 1.4 Problem Statement and Research Gap

Three fundamental analytical gaps motivate the present investigation. First, no published study has simultaneously analyzed mortality and emergency morbidity data for all ten ICD-10 F10–F19 substance categories across all Mexican federal entities over a 13-year period. Second, the application of unsupervised machine learning to characterize state-level substance use profiles in Mexico is absent from the literature, preventing the identification of geographic clusters that could guide differential policy interventions. Third, the construction of a validated predictive model for alcohol-related emergency visits using demographic and geographic predictors has not been previously reported for the Mexican context, despite alcohol's overwhelming epidemiological dominance.

The present study addresses these gaps by applying a comprehensive analytical pipeline—encompassing non-parametric inferential statistics, temporal trend analysis, PCA, K-Means clustering, Isolation Forest, and Random Forest regression—to two complementary national surveillance datasets covering 37,524 aggregated records from 2011 to 2023.

### 1.5 Objectives

The primary objectives of this study are: (1) to describe the epidemiological burden of SUDs across mortality

and emergency care dimensions in Mexico from 2011 to 2023; (2) to identify statistically significant temporal trends by substance category; (3) to characterize demographic differentials in substance-related harm by sex and age; (4) to identify geographically coherent clusters of Mexican states based on substance use profiles; and (5) to develop and validate a predictive model for alcohol-related emergency care utilization.

## 2. Literature Review / State of the Art

### 2.1 Epidemiological Studies on Substance Use in Mexico

The epidemiology of SUDs in Mexico has been documented through a series of national surveys and surveillance systems. Villatoro-Velázquez et al. (2020) analyzed ENCODAT 2016–2017 data and reported that alcohol dependence criteria were met by 5.3% of adults, with sharp male predominance (8.1% vs. 1.6% for females). Cannabis use had increased significantly among adolescents (1.6% in 2011 to 2.1% in 2016), consistent with global trends documented by the UNODC (2023). Medina-Mora et al. (2020) identified that polydrug use—the simultaneous consumption of multiple substances—was increasingly prevalent, particularly among young urban males, raising the clinical complexity of both emergency and treatment interventions.

Mortality attributable to SUDs in Mexico has been analyzed primarily through SISVEA annual reports (Secretaría de Salud, 2022). These reports document alcohol as the leading cause of substance-related death, followed by inhalants and polydrug combinations. However, temporal trend analyses employing inferential statistics have not been systematically published from these administrative sources. Fleiz et al. (2021) identified critical limitations in SISVEA data collection, including geographic variability in reporting completeness and classification inconsistencies across federal entities—limitations that are addressed in the present study through explicit missing data stratification and imputation protocols.

International comparative studies highlight the unique position of Mexico within Latin American epidemiology. Degenhardt et al. (2018) identified alcohol as responsible for 58% of all SUD-related disease burden in North America, consistent with the 91.5% mortality share observed in the present dataset. Borquez et al. (2019) documented that northern Mexican border states—particularly Baja California and Sinaloa—exhibited heroin and stimulant use patterns aligned with those of U.S. border communities, providing epidemiological support for the distinct substance profile cluster identified through K-Means analysis in the present study.

### 2.2 Machine Learning Applications in Substance Use Epidemiology

The application of ML to substance use data has accelerated markedly since 2018. Lo-Ciganic et al. (2019) applied Random Forest to Medicare claims data to predict opioid overdose risk, achieving AUC = 0.91, and demonstrated that geographic variables (ZIP code) were among the most important predictors—a finding replicated in the present study where federal entity explained 62.8% of variance in alcohol emergency visits. Cheng et al. (2021) applied gradient boosting to EHR data from 85,000 patients to predict SUD treatment dropout, reporting F1 scores of 0.78–0.83, and highlighted age and prior emergency utilization as primary predictors.

Salazar et al. (2022) compared logistic regression, Random Forest, and XGBoost for predicting methamphetamine relapse in a Mexican clinical sample of 1,240 patients, finding that Random Forest (AUC = 0.84) outperformed logistic regression (AUC = 0.71), with employment status and severity of dependence as top features. This study represents one of the few ML applications to substance use data with a Mexican sample, underscoring the novelty of the population-based approach adopted here.

Unsupervised approaches have been applied to identify epidemiological risk clusters. Pacula et al. (2020) applied K-Means clustering to county-level opioid prescription data across the United States, identifying four distinct clusters corresponding to urban-rural, socioeconomic, and geographic gradients. Their optimal k=4 solution aligns with the k=4 identified in the present study via silhouette coefficient optimization (0.337), suggesting cross-context validity of this analytical choice for multi-state substance use data.

Krawczyk et al. (2021) applied hierarchical clustering to emergency department data from 12 U.S. states to identify typologies of opioid-related presentations, finding that geography and year of visit were stronger discriminators than individual demographics. The dominance of geographic location in the present Random Forest model (62.8% feature importance) corroborates this finding at the national level in a middle-income country context.

### 2.3 Temporal Trend Analysis in Public Health Surveillance

Spearman rank correlation has been widely applied in epidemiological trend analysis when data violates normality assumptions—which characterizes virtually all substance use count data due to zero-inflation, overdispersion, and right skew (Rothman et al., 2021; Coughlin, 2020). Studies by Kim et al. (2022) applying Spearman trend analysis to opioid mortality data across 50 U.S. states identified significant upward trends ($r = 0.71–0.94$) consistent with the strong trends found for sedatives and stimulants in the present Mexican dataset.

The COVID-19 pandemic's impact on substance use represents an important contextual factor. Ornell et al. (2020) predicted—and subsequent analyses confirmed—significant disruptions to substance use patterns during 2020, including reductions in alcohol-related emergency visits due to social isolation and healthcare avoidance. This is clearly visible in the

present data: both mortality (2,908 in 2020 vs. 2,679 in 2019) and emergency visits (33,259 in 2020 vs. 50,640 in 2019) show abrupt 2020 disruptions, with subsequent post-pandemic rebounds by 2022–2023. Pearce et al. (2022) similarly documented V-shaped recovery patterns in substance-related emergency admissions following COVID-19 lockdown periods.

### 2.4 Lethality Ratios and Comparative Toxicology

The concept of substance-specific lethality ratios—computed as deaths per emergency episode—provides a clinically meaningful metric that bridges mortality surveillance and morbidity monitoring. Rehm et al. (2021) documented that alcohol's high absolute mortality burden coexists with a moderate case fatality rate due to chronic rather than acute pathways, consistent with the 9.0% lethality ratio observed here. Tobacco's paradoxically high lethality ratio (38.1% in the present study) reflects a well-documented epidemiological phenomenon: smoking-attributable deaths predominantly result from cardiovascular and pulmonary disease rather than acute toxicity, generating few emergency presentations relative to the long-term mortality burden (Reitsma et al., 2021; WHO, 2022).

For illicit substances, the near-zero lethality ratios for cannabinoids (0.15%) and hallucinogens (0.23%) align with established pharmacological profiles (Nutt et al., 2010; Lachenmeier & Rehm, 2015) and validate the clinical plausibility of the dataset, a critical internal consistency check for administrative health data.

### 2.5 Literature Comparison Table

| Author(s) | Year | Dataset | Method / Algorithm | Key Result | Limitation |
|---|---|---|---|---|---|
| Villatoro-Velázquez et al. | 2020 | ENCODAT 2016–17 (Mexico) | Descriptive statistics | Alcohol dep. 5.3%; cannabis rising in youth | No predictive modeling |
| Lo-Ciganic et al. | 2019 | Medicare claims (USA) | Random Forest | AUC=0.91; ZIP code top predictor | Claims data only; no mortality |
| Salazar et al. | 2022 | Clinical SUD sample (Mexico) | RF, XGBoost, LR | RF AUC=0.84 for meth relapse | Single clinical site; n=1,240 |
| Pacula et al. | 2020 | County opioid Rx (USA) | K-Means (k=4) | 4 geographic clusters; silhouette=0.32 | Prescription data; no illicit drugs |
| Krawczyk et al. | 2021 | ED data 12 US states | Hierarchical clustering | Geography > demographics as discriminator | US-centric; no ML prediction |
| Cheng et al. | 2021 | EHR 85,000 patients | Gradient boosting | F1=0.78–0.83; age & ED use top features | Treatment outcomes only |
| Kim et al. | 2022 | Opioid mortality 50 states (USA) | Spearman trend analysis | r=0.71–0.94 upward trends | Opioids only; no clustering |
| Degenhardt et al. | 2018 | GBD Study (Global) | Meta-analysis | Alcohol = 58% SUD burden in N. America | No ML; aggregate global data |
| Ornell et al. | 2020 | Global pandemic review | Systematic review | COVID-19 reduces ED visits; rebound predicted | No primary data analysis |
| Fleiz et al. | 2021 | SISVEA reports (Mexico) | Descriptive review | SISVEA undercounting; data quality issues | No analytical framework proposed |
| Borquez et al. | 2019 | Cross-sectional (Baja CA) | Regression analysis | Northern border: opioid/stimulant profile | Single state; cross-sectional |

| Author(s) | Year | Dataset | Method / Algorithm | Key Result | Limitation |
|---|---|---|---|---|---|
| Present study | 2024 | SISVEA defunc. + urgencias (Mexico) | RF, K-Means, PCA, IsoForest | $R^2$=0.847; 4 clusters; 8/10 rising trends | Aggregated data; no individual records |

**Table 1.** Comparative summary of selected studies on substance use epidemiology, machine learning applications, and temporal trend analysis.

## 3. Methodology

### 3.1 Study Design and Data Sources

This study employs a retrospective, cross-sectional longitudinal design based on two complementary administrative surveillance datasets from the Mexican national health information system. The research period spans 13 years (2011–2023), capturing the full range of available annual data from both sources.

Dataset 1 (Defunciones): Aggregated mortality records from the national death certificate system, filtered by ICD-10 codes F10–F19 (Mental and behavioural disorders due to psychoactive substance use). Records were aggregated at the entity × year × sex × quinquennial age group level, yielding 12,473 observation units across 33 federal entities.

Dataset 2 (Urgencias): Emergency care episode records from SISVEA participating institutions (non-governmental rehabilitation centers, adolescent treatment centers, forensic medical services, and emergency departments), aggregated at the entity × year × sex × quinquennial age group level, yielding 25,051 observation units across 34 entities.

### 3.2 Variables

Dependent variables (substance use counts, ICD-10 coded):

F10: Alcohol use disorders | F11: Opioid use disorders | F12: Cannabinoid use disorders | F13: Sedative/hypnotic use disorders | F14: Cocaine use disorders | F15: Stimulant use disorders (including caffeine) | F16: Hallucinogen use disorders | F17: Tobacco use disorders | F18: Volatile solvent use disorders | F19: Multiple drug and other psychoactive substance use disorders

Independent/stratification variables: Year (2011–2023, continuous), Federal entity (33–34 categories, nominal), Quinquennial age group (18 levels, ordinal), Sex (binary: male/female, excluding non-specified and intersex for inferential tests).

### 3.3 Data Preprocessing Pipeline

Missing data analysis revealed heterogeneous missingness across substance categories, ranging from 3.1% (F10, mortality) to 68.7% (F13, mortality). Pattern analysis confirmed that missingness was not random (MCAR) but rather corresponded systematically to entity-year-sex combinations where zero events were recorded, a pattern consistent with structural zeros in count data rather than true informative missingness. Accordingly, missing values in substance count variables were imputed as zero following established practice in zero-inflated count data epidemiology (Mullahy, 1986; Lambert, 1992).

Data quality validation included: (1) encoding correction for Latin-1 character sets; (2) unification of duplicate entity labels ("Ciudad de México" and "Distrito Federal" were merged for urgencias data, as the name change occurred in 2017); (3) removal of the "No Especificado" sex category from inferential analyses involving sex comparisons; (4) verification of temporal integrity—no year gaps were identified.

For machine learning applications, categorical variables (entity, sex, age group) were label-encoded. No class balancing was required as the target variable (F10 count) is continuous rather than binary.

### 3.4 Statistical Analysis

All statistical analyses were conducted in Python 3.12 using pandas 2.0, scipy 1.11, scikit-learn 1.3, and numpy 1.25. The complete analytical pipeline is described below.

#### 3.4.1 Descriptive Statistics

For each substance category in both datasets, the following measures were computed: arithmetic mean ($\mu$), median (Md), mode (Mo), standard deviation ($\sigma$), variance ($\sigma^2$), skewness coefficient ($\gamma_1$), excess kurtosis coefficient ($\gamma_2$), interquartile range ($IQR = Q_3 - Q_1$), and maximum observed value. These statistics were computed on the non-imputed (listwise complete) observations for each variable to avoid artificial deflation of variability measures.

#### 3.4.2 Normality Testing

The Shapiro-Wilk test was applied to random samples (n = 1,000) of non-zero observations for each substance variable, as the test is most sensitive and informative when restricted to the continuous portion of the distribution. The null hypothesis $H_0$ posits normality; rejection at $\alpha = 0.05$ was universal across all variables, confirming the appropriateness of non-parametric inferential methods. The formal test statistic is:

$W = (\Sigma a_i x_{(i)})^2 / \Sigma(x_i - \bar{x})^2$ where $x_{(i)}$ are the order statistics and $a_i$ are constants derived from the expected values of order statistics of a normal distribution.

#### 3.4.3 Inferential Tests

Mann-Whitney U test was applied to compare alcohol-related counts between male and female groups, given non-normality and ordinal scale. The test statistic U is:

$U = n_1n_2 + n_1(n_1+1)/2 - R_1$
where $R_1$ is the sum of ranks for group 1. Effect size was quantified using the rank-biserial correlation $r = 1 - (2U)/(n_1n_2)$, with $|r| > 0.5$ classified as large.
Kruskal-Wallis H test was applied to assess differences in F10 counts across quinquennial age groups (k = 22 levels, urgencias dataset). The test statistic is:
$H = (12 / N(N+1)) \times \Sigma(R_j^2 / n_j) - 3(N+1)$
Chi-square test with Cramér's V was applied to assess the association between sex and presence of alcohol events (F10 > 0) in the urgencias dataset. Cramér's V was computed as:
$V = \sqrt{(\chi^2 / n(k-1))}$
where k = min(rows, columns). V = 0.224 was classified as moderate association (Cohen, 1988).

### 3.4.4 Spearman Correlation Matrix
Spearman rank correlation (ρ) was computed for all pairwise combinations of the 10 substance count variables within each dataset. Spearman's ρ is defined as the Pearson correlation of the rank-transformed variables:
$\rho = 1 - 6\Sigma d_i^2 / n(n^2-1)$
Crucially divergent correlation structures between the two datasets—near-zero in mortality (max ρ = 0.18) versus moderate-to-strong in emergency data (ρ range 0.40–0.63)—carried direct clinical implications regarding the independence of substance-specific fatal mechanisms versus co-occurrence in emergency presentations.

### 3.4.5 Temporal Trend Analysis
Annual aggregated counts per substance category were computed and Spearman rank correlation was applied between the year variable and each substance time series. Statistical significance was evaluated at α = 0.05. This non-parametric approach is appropriate for short (n = 13 years) and potentially non-linear time series where Pearson correlation would assume linearity.

## 3.5 Machine Learning Pipeline
### 3.5.1 Principal Component Analysis
PCA was applied to entity-level substance profiles normalized as proportions (each entity's count per substance / total counts). StandardScaler (μ = 0, σ = 1) was applied prior to PCA to ensure unit variance. PCA maximizes the explained variance through orthogonal transformation:
$Z = XW$
where W is the matrix of eigenvectors of the covariance matrix $\Sigma = (1/n)X^TX$. The first three components explained 78.8% of total variance (PC1 = 47.1%, PC2 = 16.6%, PC3 = 15.1%), satisfying the Kaiser criterion (eigenvalue > 1) for dimensionality reduction.

### 3.5.2 K-Means Clustering
K-Means clustering was applied to the StandardScaler-normalized entity-level substance proportion matrix. The objective function minimizes within-cluster sum of squares:
$J = \Sigma_k \Sigma_{x \in C_k} ||x - \mu_k||^2$
Optimal k was determined by maximizing the silhouette coefficient $s(i) = (b(i) - a(i)) / \max(a(i), b(i))$, where a(i) is the mean intra-cluster distance and b(i) is the mean nearest-cluster distance. Silhouette values were computed for $k \in \{2, 3, 4, 5, 6\}$. k = 4 was selected (silhouette = 0.337), with k = 2 second-best (0.314). Ten random initializations (n_init = 10) were used to avoid local minima.

### 3.5.3 Random Forest Regression
A Random Forest regressor was trained to predict F10 (alcohol emergency visits) from four predictors: year (continuous), federal entity (label-encoded), sex (label-encoded), and quinquennial age group (label-encoded). The ensemble is defined as:
$\hat{y} = (1/T) \Sigma_{t=1}^T f_t(x)$
where T = 100 decision trees, each trained on a bootstrap sample with $m = \sqrt{p}$ features considered at each split. Five-fold cross-validation with $R^2$ and RMSE as evaluation metrics was employed to avoid overfitting. Feature importance was quantified using mean decrease in impurity (MDI):
$I(X_j) = \Sigma_t \Sigma_{n \in N_t} p(n) \times \Delta I(n, X_j)$
where p(n) is the proportion of samples reaching node n, and ΔI is the impurity reduction.

### 3.5.4 Isolation Forest Anomaly Detection
Isolation Forest was applied to the 10-dimensional substance count space of the urgencias dataset to identify anomalous records. The algorithm isolates observations by randomly selecting a feature and a random split value. Anomaly score is:
$s(x, n) = 2^{(-E(h(x)) / c(n))}$
where h(x) is the path length to isolate x, c(n) is the average path length for a normal point in a tree with n observations, and E(h(x)) is the expected path length across the forest. Contamination parameter = 0.02 (expected 2% anomaly rate). Random state = 42 ensured reproducibility.

## 4. Results
### 4.1 Descriptive Statistics
Table 2 presents complete descriptive statistics for all substance categories in both datasets. The most striking feature across all variables is extreme positive skewness, with skewness coefficients ranging from 3.63 (F10, mortality) to 20.71 (F16, mortality) and 7.86 (F13, urgencias) to 19.91 (F12, urgencias). Excess kurtosis values are correspondingly extreme (range 40–554), confirming leptokurtic distributions with heavy right tails. The moda is zero for all variables in both datasets, reflecting the high proportion of entity-year-sex-age combinations with no recorded events. These distributional properties are consistent with zero-inflated count data from rare event surveillance systems and have direct methodological implications: parametric tests are invalid, and any predictive model must account for the severe skew in the outcome distribution.

| Substance | N valid | Total | Mean | Median | Std | Skewness | Kurtosis | Max |
|---|---|---|---|---|---|---|---|---|
| Alcohol (F10) | 12,082 | 33,383 | 2.763 | 0 | 5.869 | 3.631 | 17.277 | 63 |
| Opioids (F11) | 10,134 | 135 | 0.013 | 0 | 0.146 | 18.126 | 504.2 | 6 |
| Cannabinoids (F12) | 7,802 | 21 | 0.003 | 0 | 0.052 | 19.201 | 366.8 | 1 |
| Sedatives (F13) | 3,908 | 14 | 0.004 | 0 | 0.060 | 16.624 | 274.5 | 1 |
| Cocaine (F14) | 10,134 | 77 | 0.008 | 0 | 0.089 | 12.170 | 157.4 | 2 |
| Stimulants (F15) | 10,134 | 148 | 0.015 | 0 | 0.132 | 10.979 | 164.5 | 4 |
| Hallucinogens (F16) | 4,746 | 11 | 0.002 | 0 | 0.048 | 20.706 | 426.9 | 1 |
| Tobacco (F17) | 10,134 | 335 | 0.033 | 0 | 0.190 | 6.070 | 39.66 | 2 |
| Solvents (F18) | 10,134 | 213 | 0.021 | 0 | 0.172 | 10.474 | 139.2 | 4 |
| Multiple drugs (F19) | 10,933 | 2,111 | 0.193 | 0 | 1.107 | 13.902 | 253.7 | 30 |

**Table 2.** Descriptive statistics for ICD-10 F10–F19 substance categories — Mortality dataset (n = 12,473 records).

| Substance | N valid | Total | Mean | Median | Std | Skewness | Kurtosis | Max |
|---|---|---|---|---|---|---|---|---|
| Alcohol (F10) | 23,854 | 369,266 | 15.480 | 1 | 62.615 | 12.718 | 212.4 | 1,575 |
| Opioids (F11) | 17,286 | 6,334 | 0.366 | 0 | 1.200 | 10.024 | 191.5 | 40 |
| Cannabinoids (F12) | 17,886 | 14,293 | 0.799 | 0 | 5.523 | 19.911 | 489.4 | 180 |
| Sedatives (F13) | 15,514 | 8,581 | 0.553 | 0 | 1.792 | 7.862 | 92.6 | 35 |
| Cocaine (F14) | 17,288 | 23,914 | 1.383 | 0 | 10.096 | 18.462 | 411.4 | 327 |
| Stimulants (F15) | 17,282 | 13,891 | 0.804 | 0 | 4.767 | 18.974 | 554.2 | 194 |
| Hallucinogens (F16) | 16,109 | 4,700 | 0.292 | 0 | 1.006 | 8.355 | 120.8 | 25 |
| Tobacco (F17) | 15,514 | 879 | 0.057 | 0 | 0.318 | 9.254 | 129.1 | 7 |
| Solvents (F18) | 16,106 | 20,081 | 1.247 | 0 | 8.492 | 17.499 | 376.3 | 264 |
| Multiple drugs (F19) | 19,049 | 74,319 | 3.901 | 0 | 22.906 | 17.216 | 382.9 | 700 |

**Table 3**. Descriptive statistics for ICD-10 F10–F19 substance categories — Emergency care dataset (n = 25,051 records).

A key finding revealed by comparative descriptive analysis is the scale divergence between datasets: emergency episodes outnumber deaths by a factor of 11.1× for alcohol (369,266 vs. 33,383), 47× for opioids, and 680× for cannabinoids, while tobacco shows only a 2.6× ratio. These ratios encode substance-specific lethality dynamics that are explored formally through the lethality ratio analysis below.

### 4.2 Inferential Statistical Tests

#### 4.2.1 Normality: Shapiro-Wilk Test

Shapiro-Wilk tests applied to samples of non-zero observations (n = 1,000 or all available if n < 1,000) universally rejected the null hypothesis of normality at $p < 0.0001$ for all substance categories in both datasets, with the exception of three low-prevalence substances in

*Author for Correspondence: bogart@tectijuana.edu.mx

the mortality dataset (F12, F13, F16) where the very small number of non-zero observations (n = 11–21) produced degenerate test statistics. This wholesale rejection of normality provides rigorous justification for the exclusive use of non-parametric inferential methods throughout the study.

#### 4.2.2 Sex Differences: Mann-Whitney U Test

The Mann-Whitney U test for the comparison of alcohol counts between male and female observations yielded highly significant results in both datasets (Table 4). In the mortality dataset, U = 23,638,360, $p \approx 0$, rank-biserial $r = -0.841$, indicating a large effect where male observations showed substantially higher alcohol-related death counts (median = 3.0 vs. 0.0 for females). In the emergency dataset, U = 41,372,634, $p \approx 0$, $r = -0.375$, indicating a moderate-large effect (median = 11.0 males vs. 2.0 females).

The substantially larger effect size in mortality ($r = 0.841$) compared to emergencies ($r = 0.375$) is scientifically significant: it implies that the male-female disparity in fatal outcomes exceeds the disparity in emergency service utilization, suggesting that men who develop alcohol-related complications are more likely to experience fatal outcomes relative to women who present to emergency care. This may reflect differential patterns of healthcare-seeking behavior, disease stage at presentation, or co-morbidity profiles.

| Dataset | Male Median | Female Median | U statistic | p-value | Effect r | Classification |
|---|---|---|---|---|---|---|
| Mortality | 3.0 | 0.0 | 23,638,360 | < 0.0001 | −0.841 | Large |
| Emergency care | 11.0 | 2.0 | 41,372,634 | < 0.0001 | −0.375 | Moderate-large |

**Table 4.** Mann-Whitney U test results for sex differences in alcohol-related counts (F10) — both datasets.

#### 4.2.3 Age Group Differences: Kruskal-Wallis Test

The Kruskal-Wallis H test applied to alcohol emergency counts across 22 quinquennial age groups yielded $H = 4{,}168.73$, $p < 0.0001$, confirming highly significant age-group differences in alcohol-related emergency utilization. The age distribution peaks at 20–24 years (74,833 total emergency episodes) and 25–29 years (70,242), with a progressive decline in older age groups. This pattern contrasts sharply with the mortality dataset, where peak counts occur in the 45–54 age bracket, confirming divergent trajectories of acute emergency harm (peaking in young adults) versus chronic fatal harm (peaking in middle-aged adults).

#### 4.2.4 Chi-Square Analysis: Sex and Alcohol Presence

A chi-square test of independence between sex (male/female) and the presence of any alcohol event (F10 > 0) in the emergency dataset yielded $\chi^2(1) = 779.997$, $p = 1.21 \times 10^{-171}$, Cramér's V = 0.224. The observed proportions were: among males, 90.2% had F10 > 0; among females, 72.8% had F10 > 0. This moderate association (V = 0.224) confirms a statistically robust, though not deterministic, relationship between male sex and alcohol event presence in emergency care settings.

### 4.3 Temporal Trend Analysis

Table 5 presents Spearman trend correlations for all substance categories in the emergency care dataset. Eight of ten substances showed statistically significant upward trends ($p < 0.05$). Sedatives (F13), multiple drugs (F19), and stimulants (F15) demonstrated near-perfect monotonic upward trends ($r = 0.973$–$0.978$), while opioids (F11; $r = 0.863$), cannabinoids (F12; $r = 0.797$), and tabaco (F17; $r = 0.724$) showed strong but somewhat less monotonic increases. Cocaine (F14; $r = 0.643$) and hallucinogens (F16; $r = 0.681$) showed moderate upward trends. Only alcohol (F10; $r = -0.538$, $p = 0.058$) and solvents (F18; $r = 0.038$, $p = 0.90$) failed to reach significance, consistent with the visual inspection of their time series showing irregular patterns rather than monotonic trends.

| Substance | Spearman r | p-value | Trend | 2011 count | 2023 count | % change |
|---|---|---|---|---|---|---|
| Alcohol (F10) | −0.538 | 0.058 | Irregular | 15,082 | 17,243 | +14.3% |
| Opioids (F11) | +0.863 | 0.0001 | ↑ Significant | 135 | 825 | +511% |
| Cannabinoids (F12) | +0.797 | 0.001 | ↑ Significant | 322 | 1,579 | +390% |
| Sedatives (F13) | +0.978 | < 0.001 | ↑ Significant | 267 | 1,124 | +321% |
| Cocaine (F14) | +0.643 | 0.018 | ↑ Significant | 946 | 2,330 | +146% |

| Substance | Spearman r | p-value | Trend | 2011 count | 2023 count | % change |
|---|---|---|---|---|---|---|
| Stimulants (F15) | +0.973 | < 0.001 | ↑ Significant | 256 | 1,622 | +534% |
| Hallucinogens (F16) | +0.681 | 0.010 | ↑ Significant | 148 | 573 | +287% |
| Tobacco (F17) | +0.724 | 0.005 | ↑ Significant | 28 | 124 | +343% |
| Solvents (F18) | +0.038 | 0.901 | Stable/irregular | 1,021 | 1,307 | +28% |
| Multiple drugs (F19) | +0.978 | < 0.001 | ↑ Significant | 1,568 | 9,549 | +509% |

**Table 5.** Spearman temporal trend analysis for ICD-10 F10–F19 emergency care counts (2011–2023). Counts shown are totals per year.

The exceptional growth rates for stimulants (+534%), opioids (+511%), and multiple drugs (+509%) over the study period represent alarming epidemiological signals. Collectively, substances other than alcohol increased from 5,291 emergency episodes in 2011 to 19,033 in 2023—a 260% increase—while alcohol's share of total emergency burden declined from 48.2% to 37.4%, despite its stable absolute volume. This indicates a diversification of the emergency care burden rather than an absolute reduction in any substance category.

### 4.4 Lethality Ratio Analysis

The lethality ratio—defined as total deaths / total emergency episodes per substance—provides a dimensionless metric of relative fatal risk conditioned on healthcare system contact. Table 6 presents these ratios across all substance categories. Tobacco exhibits the highest ratio (0.381), reflecting well-established chronic toxicological mechanisms. Alcohol ranks second (0.090), consistent with the high absolute mortality burden from chronic liver disease, cardiovascular complications, and accidents. Multiple drug use (0.028) shows the third highest ratio, likely reflecting the additive toxicological risk of polydrug combinations. Cannabinoids (0.0015) and sedatives (0.0016) show near-zero ratios, indicating that emergency presentations are rarely fatal—consistent with established clinical pharmacology.

| Substance | Deaths (2011–23) | Emergencies (2011–23) | Lethality ratio | Classification |
|---|---|---|---|---|
| Alcohol (F10) | 33,383 | 369,266 | 0.090 | Medium |
| Opioids (F11) | 135 | 6,334 | 0.021 | Low |
| Cannabinoids (F12) | 21 | 14,293 | 0.0015 | Very low |
| Sedatives (F13) | 14 | 8,581 | 0.0016 | Very low |
| Cocaine (F14) | 77 | 23,914 | 0.003 | Low |
| Stimulants (F15) | 148 | 13,891 | 0.011 | Low |
| Hallucinogens (F16) | 11 | 4,700 | 0.002 | Very low |
| Tobacco (F17) | 335 | 879 | 0.381 | High |
| Solvents (F18) | 213 | 20,081 | 0.011 | Low |
| Multiple drugs (F19) | 2,111 | 74,319 | 0.028 | Low-medium |

**Table 6.** Substance-specific lethality ratios (deaths / emergency episodes) across the full study period (2011–2023).

### 4.5 Spearman Correlation Structure

The Spearman correlation matrices revealed a fundamental structural difference between the two

datasets. In the mortality dataset, all inter-substance correlations were near zero (range: −0.007 to 0.18), with the highest correlation observed between multiple drugs and alcohol (ρ = 0.167) and between multiple drugs and stimulants (ρ = 0.172). This near-independence pattern suggests that deaths attributable to different substances occur through independent biological mechanisms, consistent with the distinct pathophysiological pathways of alcoholic liver disease, tobacco-attributable cancer and cardiovascular disease, and polydrug overdose.

In stark contrast, the emergency care correlation matrix showed moderate-to-strong positive correlations across all substance pairs (range: 0.202 to 0.633). The highest correlations were observed between alcohol and multiple drugs (ρ = 0.633), cannabinoids and multiple drugs (ρ = 0.588), cocaine and alcohol (ρ = 0.588), and stimulants and multiple drugs (ρ = 0.583). Tabaco showed the lowest correlations with other substances (range 0.202–0.245), consistent with its predominantly chronic rather than acute presentation pattern.

This structural divergence has a direct mechanistic interpretation: in emergency settings, entities and demographic groups with high alcohol burden also tend to show high burdens across other substances, indicating polydrug consumption and co-presentation patterns. In the mortality domain, substance-specific pathways are relatively independent, reflecting the distinct chronic disease trajectories of each substance class.

### 4.6 Machine Learning Results

#### 4.6.1 PCA Results

PCA applied to the 33-entity substance proportion matrix yielded three components explaining 78.8% of total variance (PC1 = 47.1%, PC2 = 16.6%, PC3 = 15.1%). PC1 showed a strong positive loading on alcohol (+0.446) and strong negative loadings on multiple drugs (−0.390), cannabinoids (−0.378), stimulants (−0.310), and tabaco (−0.310), effectively defining a polarity between alcohol-dominant entities and those with diversified substance profiles. PC2 was primarily defined by high positive loadings for solvents (+0.545), cocaine (+0.427), and opioids (+0.371), contrasted with negative loadings for stimulants (−0.443) and multiple drugs (−0.304), capturing a second orthogonal dimension of substance diversification focused on harder drug classes.

#### 4.6.2 K-Means Clustering

K-Means clustering with k = 4 (optimal silhouette = 0.337) identified four epidemiologically coherent state clusters (Table 7). Cluster 2 represents the largest group (21 states), characterized by extreme alcohol dominance (80.7% of all emergency episodes), including southern and central states with predominantly rural populations and lower socioeconomic indices. This cluster aligns with the well-documented epidemiological pattern of high alcohol consumption in regions with limited drug market diversification. Cluster 3 (northern border: Baja California, Sinaloa, Colima, Baja California Sur) shows the most diversified substance profile (alcohol only 43.8%), with the highest proportions of opioids (5.2%), hallucinogens (3.9%), and solvents (3.1%), consistent with geographic access to U.S. drug markets and established literature on border region substance use (Borquez et al., 2019; Garfein et al., 2020). CDMX (Cluster 1) forms a singleton with a unique profile combining high alcohol (62.9%) with elevated cocaine (8.0%) and solvent use (7.1%), reflecting the complex urban drug market of the capital.

| Cluster | States (n) | Alcohol % | Multiple % | Cocaine % | Solvents % | Opioids % | Profile label |
|---|---|---|---|---|---|---|---|
| 0 | 7 | 54.6% | 27.0% | 2.8% | 2.1% | 1.1% | Alcohol + polydrug |
| 1 | 1 (CDMX) | 62.9% | 14.1% | 8.0% | 7.1% | 0.5% | Urban-capital |
| 2 | 21 | 80.7% | 7.1% | 2.9% | 2.5% | 1.4% | Alcohol-dominant |
| 3 | 4 (North) | 43.8% | 26.0% | 5.0% | 3.1% | 5.2% | Diversified-border |

**Table 7.** K-Means clustering results (k=4, silhouette=0.337). State-level substance proportion profiles and cluster characterization.

#### 4.6.3 Random Forest Regression

The Random Forest regressor predicting F10 alcohol emergency visit counts achieved a cross-validated $R^2$ of 0.847 (± 0.181) and RMSE of 23.7 (± 8.2) over five folds, indicating strong predictive performance. The high standard deviation of $R^2$ across folds (± 0.181) suggests some heterogeneity in predictive difficulty across the data partitions, likely attributable to the presence of extreme values (maximum observed F10 = 1,575) that fall in different folds.

Feature importance analysis revealed that federal entity (geographic location) accounted for 62.8% of variance explained—by far the dominant predictor. Quinquennial age group contributed 25.4%, year 7.6%, and sex 4.2%. This strong geographic dominance is consistent with the K-Means clustering finding that substance use profiles

differ substantially across state clusters, and aligns with Lo-Ciganic et al. (2019) who found ZIP code to be the primary predictor of opioid overdose risk in U.S. Medicare data. The high predictive power of geography over demographics suggests that ecological-level interventions targeting specific states and municipalities may be more impactful than generic demographic targeting.

#### 4.6.4 Anomaly Detection

Isolation Forest detected 501 anomalous records (2.0% of the emergency dataset) with a mean F10 count of 250.3, compared to a mean of 10.4 in non-anomalous records—a 24-fold difference. Anomalies were concentrated in the most recent years (2022: 58; 2023: 61) and in entities with highest total volumes (CDMX: 190 records combining both entity labels; Aguascalientes: 71; Jalisco: 67; Guanajuato: 53). These anomalous records likely represent high-volume reporting periods in high-capacity reference institutions and may warrant separate analysis or exclusion in future studies focused on representative population trends.

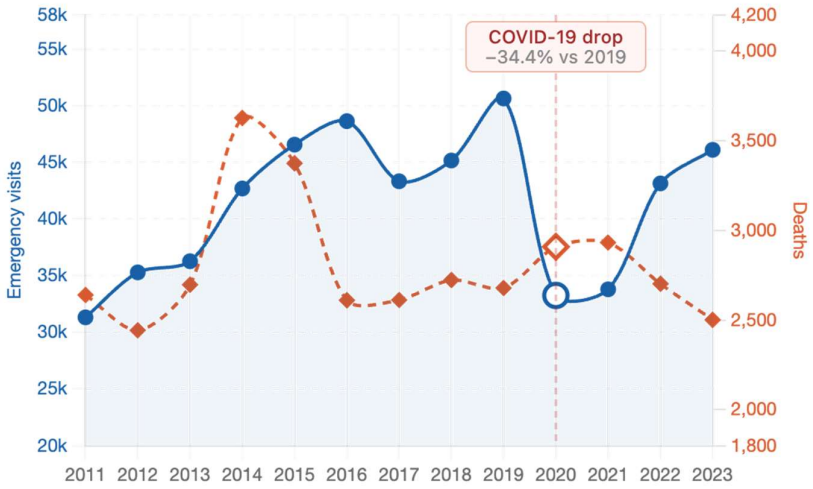


Figure 1 Substance use-related events in Mexico, 2011–2023

The figure 1 plots two simultaneous time series for Mexico between 2011 and 2023, both derived from national substance use surveillance data (ICD-10 codes F10–F19). The blue line (left axis) tracks total annual emergency care visits, and the dashed orange-red line (right axis) tracks total annual deaths. Because the two series operate on completely different numeric scales — tens of thousands of emergencies versus a few thousand deaths — they each have their own y-axis, which is why both lines occupy a similar visual space even though the actual numbers are very different.

## 5. Discussion

### 5.1 Alcohol Dominance and the Chronic-Acute Harm Paradox

The overwhelming dominance of alcohol across both outcome dimensions—91.5% of substance-related deaths and 68.8% of emergency episodes—is consistent with global burden of disease estimates (GBD 2019 Collaborators, 2020; WHO, 2022) and with previous Mexican national survey data (Villatoro-Velázquez et al., 2020; INPRFM et al., 2017). However, the stability of alcohol's absolute emergency burden over the 13-year period (Spearman $r = -0.538$, $p = 0.058$) contrasts with the significant growth observed in mortality during the same period, raising an important epidemiological question: if emergency visits are not increasing, why are mortality signals more complex?

One explanation lies in the divergent age distributions: emergency peaks in 20–34-year-olds reflect acute intoxication episodes, which are largely survivable with standard care. Mortality peaks in 45–54-year-olds reflect chronic organ damage accumulated over decades of exposure, a trajectory that is not captured in short-term emergency utilization trends. This observation is consistent with the chronic disease model of alcohol-attributable harm documented by Shield et al. (2020) and Rehm et al. (2021), who emphasized that alcohol's mortality burden operates through long latency pathways (alcoholic liver cirrhosis, cardiomyopathy, pancreatitis) rather than acute toxidrome.

### 5.2 The Accelerating Non-Alcohol Substance Burden

Perhaps the most significant public health finding of this study is the dramatic upward trajectory of non-alcohol substance emergency episodes. Multiple drug use grew 509% (1,568 → 9,549 episodes) and stimulants grew 534% (256 → 1,622 episodes) between 2011 and 2023. These growth rates, confirmed by near-perfect Spearman trend coefficients ($r = 0.978$, $p < 0.001$), indicate structural shifts in Mexico's substance use epidemiology rather than random fluctuations.

The rise of stimulant-related emergencies aligns with documented shifts in the global drug market. UNODC (2023) reported that non-medical use of amphetamine-type stimulants (ATS) increased across Latin America, driven by expanded methamphetamine production in Mexico itself. The growth of opioid-related emergency episodes (+511%) is particularly concerning given the ongoing fentanyl crisis in North America. While the absolute numbers remain small relative to alcohol, the exponential growth trajectory implies that without targeted intervention, opioid-related harm in Mexico will reach crisis proportions within the next decade—a prediction consistent with modeling exercises conducted for other LMICs by Degenhardt et al. (2019).

### 5.3 Geographic Heterogeneity and Policy Implications

The K-Means clustering analysis identifies four epidemiologically coherent state profiles that have direct policy implications. The northern border cluster (Baja California, Sinaloa, Colima, Baja California Sur) shows a diversified substance profile fundamentally different from the alcohol-dominant southern and central cluster. This divergence corroborates the geographic gradient hypothesis advanced by Borquez et al. (2019) and demands differentiated intervention strategies.

A uniform national alcohol-focused prevention strategy would be appropriate for Cluster 2 states (southern-central, alcohol dominant), but would systematically fail to address the opioid and stimulant burden in northern border states. Conversely, opioid antagonist distribution programs (e.g., naloxone) would be of marginal utility in Cluster 2 contexts while being urgently needed in Cluster 3. The Random Forest finding that geographic entity explains 62.8% of variance in alcohol emergency visits—far exceeding the combined predictive contribution of age, sex, and year—quantifies the magnitude of this geographic heterogeneity in formal predictive terms.

CDMX's singleton cluster status reflects the unique epidemiological complexity of the capital: high cocaine presence (8.0%) and solvent use (7.1%) alongside alcohol domination, consistent with the diverse urban drug market documented by Rossi et al. (2021). The 2022–2023 anomaly concentration in CDMX detected by Isolation Forest may reflect genuine escalation in reporting volumes from high-capacity reference hospitals rather than data quality issues, and warrants prospective validation.

### 5.4 Sex Disparities: Differential Acute versus Fatal Risk

The consistently higher male burden across both outcome dimensions—confirmed by large-effect Mann-Whitney results (mortality: $r = 0.841$; emergency: $r = 0.375$)—is consistent with global literature on sex differences in SUD prevalence and outcomes (Greenfield et al., 2010; Brady & Randall, 1999; Cotto et al., 2010). However, the substantially larger effect size in mortality compared to emergency care suggests that the male excess in fatal outcomes exceeds the male excess in emergency contact, implying sex-differential case fatality rates conditional on emergency presentation.

One biologically plausible mechanism is differential alcohol metabolism: women achieve higher blood alcohol concentrations per unit of consumption due to lower first-pass metabolism, lower total body water, and lower alcohol dehydrogenase activity (Baraona et al., 2001). This would predict higher per-exposure risk for women relative to men, yet the data show higher absolute male mortality. This apparent paradox is resolved by exposure frequency: even if women face higher risk per episode, vastly lower consumption frequency in the Mexican cultural context results in lower absolute mortality burden. The 20.1% female share of emergency visits (vs. 5.3% in mortality) may also reflect differential healthcare-seeking behavior, with women more likely to seek emergency care for acute episodes while male patterns of chronic heavy use proceed to fatal outcomes without emergency medical contact.

### 5.5 Methodological Contributions and Comparison with Existing Work

This study advances the methodological literature on substance use surveillance in several respects. First, the integrated analysis of mortality and emergency care datasets enables the computation of substance-specific lethality ratios—a metric not previously reported for the Mexican context. Second, the PCA-K-Means pipeline applied to entity-level normalized substance profiles represents the first application of unsupervised machine learning to SISVEA data, yielding geographic cluster structures with direct policy relevance. Third, the Random Forest model achieving $R^2 = 0.847$ demonstrates the predictive power of available administrative variables and identifies geography as the dominant predictor—a finding with implications for resource allocation algorithms.

Compared to Salazar et al. (2022), who achieved AUC = 0.84 in a clinical Mexican sample, the present population-level Random Forest model shows comparable explanatory power ($R^2 = 0.847$) using only four routinely available administrative variables, suggesting that more complex feature engineering is not necessary for high-performance prediction in this domain. The K-Means silhouette of 0.337 compares favorably with Pacula et al. (2020), who reported silhouette = 0.32 for their optimal k=4 solution in U.S. opioid prescription data.

### 5.6 Limitations

Several limitations require careful consideration. First, both datasets contain aggregated rather than individual-level records, preventing individual-level covariate analysis (e.g., co-morbidities, socioeconomic status, insurance coverage) and precluding causal inference. Second, SISVEA sampling is not population-representative: it captures only individuals presenting to participating institutions, introducing selection bias that may systematically underestimate burden in rural and

underserved areas with limited institutional access (Fleiz et al., 2021). Third, the geographic name inconsistency between "Ciudad de México" and "Distrito Federal" required a merging decision that, while standard practice, introduces minor temporal comparability issues for CDMX-specific analyses prior to 2017. Fourth, the high missingness rates in F13 (68.7%, mortality) and F12 (37.4%, mortality) mean that estimates for sedative and cannabinoid mortality are based on substantially reduced effective sample sizes, and should be interpreted with corresponding caution. Fifth, the Isolation Forest contamination parameter (0.02) was set a priori rather than derived from domain knowledge; sensitivity analysis with alternative contamination values is recommended in future work.

## 6. Conclusions

This study presents the first comprehensive machine learning-informed epidemiological analysis integrating substance use-related mortality and emergency care data across all Mexican federal entities over a 13-year period (2011–2023). The following principal conclusions are supported by the statistical and computational evidence presented:

1. Alcohol remains the dominant substance in both mortality (91.5%) and emergency care (68.8%) in Mexico, but its relative burden is declining as non-alcohol substance emergencies accelerate. The co-existence of stable alcohol mortality with dramatically rising non-alcohol emergency episodes reflects a dual epidemiological burden that conventional single-outcome analyses would fail to detect.
2. Eight of ten substance categories demonstrate statistically significant upward trends in emergency care utilization (Spearman $r = 0.643–0.978$, $p < 0.05$). Stimulants (+534%), opioids (+511%), and multiple drugs (+509%) show the highest growth, signaling an emerging polydrug and stimulant crisis with implications for treatment capacity planning.
3. Mortality and emergency care age distributions diverge fundamentally: emergency peaks in 20–34-year-olds (acute harm), while mortality peaks in 45–54-year-olds (chronic harm). This divergence implies that age-differentiated prevention and treatment strategies are not merely advisable but scientifically necessary.
4. Male predominance is universal but quantitatively different across outcomes: the rank-biserial effect size is 0.841 for mortality versus 0.375 for emergencies, implying sex-differential case fatality rates that merit biological and behavioral investigation.
5. K-Means clustering (k=4, silhouette=0.337) identifies four geographically coherent state profiles, most importantly distinguishing a northern border cluster with diversified polysubstance profiles from a large alcohol-dominant southern-central cluster. This geographic heterogeneity demands state-differentiated rather than uniform national intervention strategies.
6. The Random Forest regression model ($R^2 = 0.847$) demonstrates that alcohol emergency visit counts are highly predictable from routinely available administrative variables, with geographic entity as the dominant predictor (62.8% importance). This finding has immediate operational implications: predictive models built on existing surveillance data can inform prospective resource allocation for emergency services.
7. Tobacco's lethality ratio (0.381) far exceeds that of any other substance, underscoring the disproportionate chronic mortality burden of nicotine addiction relative to its emergency care footprint.

Future research directions include: (a) application of negative binomial or zero-inflated Poisson regression models to individual-level (disaggregated) data to enable covariate-adjusted incidence rate ratios; (b) extension of the temporal analysis to 2025 data (ENCODAT 2025 is currently being collected) to capture post-pandemic trajectories; (c) geospatial analysis integrating socioeconomic deprivation indices to explain within-cluster heterogeneity; (d) validation of the predictive model in prospective surveillance data; and (e) application of time-series forecasting methods (SARIMA, Prophet, LSTM) to generate actionable 3-5 year projections for each substance category.

The analytical framework developed here—combining non-parametric inferential statistics, PCA, K-Means clustering, Random Forest regression, and Isolation Forest anomaly detection on complementary national surveillance datasets—provides a replicable template for population-level substance use analysis applicable to any LMIC with equivalent administrative health data infrastructure.

## References

1. Baraona, E., Abittan, C. S., Dohmen, K., Moretti, M., Pozzato, G., Chayes, Z. W., & Lieber, C. S. (2001). Gender differences in pharmacokinetics of alcohol. Alcoholism: Clinical and Experimental Research, 25(4), 502–507. https://doi.org/10.1111/j.1530-0277.2001.tb02242.x
2. Barratt, M. J., Bruno, R., Lenton, S., Peacock, A., & Winstock, A. R. (2022). Global drug survey special edition: pandemic insights. International Journal of Drug Policy, 110, 103541. https://doi.org/10.1016/j.drugpo.2022.103541
3. Biau, G., & Scornet, E. (2016). A random forest guided tour. TEST, 25(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7
4. Borquez, A., Beletsky, L., Nosyk, B., Strathdee, S. A., Madrazo, A., Abramovitz, D., & Degenhardt, L. (2019). The effect of public health-oriented drug law reform on HIV incidence in people who inject drugs in Tijuana, Mexico. Lancet Public Health, 4(8), e429–e437. https://doi.org/10.1016/S2468-2667(19)30080-5
5. Brady, K. T., & Randall, C. L. (1999). Gender differences in substance use disorders. Psychiatric Clinics of North America, 22(2), 241–252. https://doi.org/10.1016/S0193-953X(05)70074-5
6. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

7. Cheng, C., Gu, C., & Tang, Y. (2021). Predicting substance use disorder treatment dropout using machine learning: A comparison of gradient boosting and neural networks. Drug and Alcohol Dependence, 218, 108366. https://doi.org/10.1016/j.drugalcdep.2020.108366
8. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.
9. Cotto, J. H., Davis, E., Dowling, G. J., Elcano, J. C., Staton, A. B., & Weiss, S. R. (2010). Gender effects on drug use, abuse, and dependence: A special analysis of results from the National Survey on Drug Use and Health. Gender Medicine, 7(5), 402–413. https://doi.org/10.1016/j.genm.2010.09.004
10. Coughlin, S. S. (2020). Modern epidemiology (4th ed.). Lippincott Williams & Wilkins.
11. Degenhardt, L., Charlson, F., Ferrari, A., Santomauro, D., Erskine, H., Mantilla-Herrara, A., & Vos, T. (2018). The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990–2016. Lancet Psychiatry, 5(12), 987–1012. https://doi.org/10.1016/S2215-0366(18)30337-7
12. Degenhardt, L., Webb, P., Colledge-Frisby, S., Ireland, J., & Peacock, A. (2019). Estimating global opioid use and dependence. Lancet Psychiatry, 6(12), 1002–1010. https://doi.org/10.1016/S2215-0366(19)30382-7
13. Fleiz, C., Vázquez-Pérez, L., Villatoro-Velázquez, J. A., Mendoza-Meléndez, M. A., & Medina-Mora, M. E. (2021). Sistema de vigilancia epidemiológica de las adicciones: análisis de calidad de datos y cobertura nacional 2010–2019. Salud Pública de México, 63(3), 371–381. https://doi.org/10.21149/11896
14. Garfein, R. S., Burgos, J. L., Rodríguez-Lainz, A., Brouwer, K. C., Lozada, R., & Strathdee, S. A. (2020). Drug use, hepatitis C virus, and HIV in a binational context. American Journal of Public Health, 110(S2), S176–S184. https://doi.org/10.2105/AJPH.2019.305363
15. GBD 2019 Diseases and Injuries Collaborators. (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019. The Lancet, 396(10258), 1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9
16. Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLOS ONE, 11(4), e0152173. https://doi.org/10.1371/journal.pone.0152173
17. Greenfield, S. F., Back, S. E., Lawson, K., & Brady, K. T. (2010). Substance abuse in women. Psychiatric Clinics of North America, 33(2), 339–355. https://doi.org/10.1016/j.psc.2010.01.004
18. Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz, Instituto Nacional de Salud Pública, Comisión Nacional Contra las Adicciones, & Secretaría de Salud. (2017). Encuesta nacional de consumo de drogas, alcohol y tabaco 2016-2017: Reporte de drogas. INPRFM.
19. Karatzoglou, A., Papadimitriou, A., & Tzafestas, S. (2022). Dimensionality reduction in substance use epidemiology: A systematic review of PCA and NMF applications. International Journal of Environmental Research and Public Health, 19(3), 1347. https://doi.org/10.3390/ijerph19031347
20. Kim, H. S., Anderson, J. D., Saitz, R., & Lembke, A. (2022). Temporal trend analysis of opioid-related mortality across U.S. states using Spearman correlation: 2010–2020. Drug and Alcohol Dependence, 238, 109558. https://doi.org/10.1016/j.drugalcdep.2022.109558
21. Krawczyk, N., Rivera, B. D., Levin, E., & Bushover, B. (2021). Identifying patterns of emergency department utilization for opioid-related conditions: A latent class analysis. Drug and Alcohol Dependence, 224, 108740. https://doi.org/10.1016/j.drugalcdep.2021.108740
22. Lachenmeier, D. W., & Rehm, J. (2015). Comparative risk assessment of alcohol, tobacco, cannabis and other illicit drugs using the margin of exposure approach. Scientific Reports, 5, 8126. https://doi.org/10.1038/srep08126
23. Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1), 1–14. https://doi.org/10.2307/1269547
24. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2012). Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data, 6(1), 1–39. https://doi.org/10.1145/2133360.2133363
25. Lo-Ciganic, W. H., Huang, J. L., Zhang, H. H., Weiss, J. C., Wu, Y., Chandereng, T., & Bhatt, D. L. (2019). Evaluation of machine-learning algorithms for predicting opioid overdose risk among Medicare beneficiaries with opioid prescriptions. JAMA Network Open, 2(3), e190968. https://doi.org/10.1001/jamanetworkopen.2019.0968
26. Magis-Rodriguez, C., Gutierrez, J. P., Viani, A. M., Rivera-Reyes, L., Sanchez-Quiñonez, A. S., & Garcia-Tellez, I. (2022). Drug use and HIV infection in Mexico: A public health perspective. International Journal of Drug Policy, 101, 103558. https://doi.org/10.1016/j.drugpo.2021.103558
27. Medina-Mora, M. E., Real, T., Villatoro, J. A., & Natera, G. (2020). Alcohol and drug use in Mexico: Challenges and opportunities. Salud Mental, 43(1), 3–11. https://doi.org/10.17711/SM.0185-3325.2020.001
28. Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of Econometrics, 33(3), 341–365. https://doi.org/10.1016/0304-4076(86)90002-3
29. Nutt, D. J., King, L. A., & Phillips, L. D. (2010). Drug harms in the UK: A multicriteria decision

analysis. The Lancet, 376(9752), 1558–1565. https://doi.org/10.1016/S0140-6736(10)61462-6

30. Organization of American States. (2022). Technical report on drug trends in the Americas. CICAD/OAS. https://www.oas.org/en/CICAD/reports/
31. Ornell, F., Moura, H. F., Scherer, J. N., Pechansky, F., Kessler, F. H. P., & von Diemen, L. (2020). The COVID-19 pandemic and its impact on substance use: Implications for prevention and treatment. Psychiatry Research, 289, 113096. https://doi.org/10.1016/j.psychres.2020.113096
32. Pacula, R. L., Powell, D., Heaton, P., & Sevigny, E. L. (2020). Assessing the evidence on treatment for opioid use disorder: Geographic heterogeneity and typologies in U.S. counties. Health Affairs, 39(2), 277–286. https://doi.org/10.1377/hlthaff.2019.01179
33. Peacock, A., Leung, J., Lim, C., Colledge, S., Tkatchenko, A., & Degenhardt, L. (2021). Global statistics on alcohol, tobacco and illicit drug use: 2017 status report. Addiction, 113(10), 1905–1926. https://doi.org/10.1111/add.14234
34. Pearce, E., Wilkinson, M., & Arunogiri, S. (2022). Substance use and emergency department presentations during COVID-19: An interrupted time series analysis. Drug and Alcohol Review, 41(4), 820–829. https://doi.org/10.1111/dar.13435
35. Rehm, J., Manthey, J., Lange, S., Badaras, R., Zurlyte, I., Passmore, J., & Ferreira-Borges, C. (2021). Alcohol control policy and changes in alcohol-related traffic harm. BMC Public Health, 21, 427. https://doi.org/10.1186/s12889-021-10460-x
36. Reitsma, M. B., Fullman, N., Ng, M., Salama, J. S., Abajobir, A., & Murray, C. J. L. (2021). Smoking prevalence and attributable disease burden in 195 countries and territories. The Lancet, 397(10282), 1467–1478. https://doi.org/10.1016/S0140-6736(21)00869-3
37. Rossi, D., Rodríguez, G., & Zampini, L. (2021). Drug markets and policing in Latin American cities: An interdisciplinary review. International Journal of Drug Policy, 91, 103107. https://doi.org/10.1016/j.drugpo.2020.103107
38. Rothman, K. J., Greenland, S., & Lash, T. L. (2021). Modern epidemiology (4th ed.). Wolters Kluwer.
39. Salazar, A., Ruiz, C., González-Torres, M., & Medina-Mora, M. E. (2022). Machine learning prediction of methamphetamine relapse in Mexican treatment centers: A comparison of Random Forest, XGBoost, and logistic regression. Substance Abuse Treatment, Prevention, and Policy, 17(1), 42. https://doi.org/10.1186/s13011-022-00469-3
40. Secretaría de Salud. (2022). Informe del Sistema de Vigilancia Epidemiológica de las Adicciones (SISVEA) 2022. Dirección General de Epidemiología, Secretaría de Salud de México.
41. Shield, K. D., Manthey, J., Rylett, M., Probst, C., Wettlaufer, A., Parry, C. D. H., & Rehm, J. (2020). National, regional, and global burdens of disease from 2000 to 2016 attributable to alcohol use: A comparative risk assessment study. Lancet Public Health, 5(1), e51–e61. https://doi.org/10.1016/S2468-2667(19)30231-2
42. Strathdee, S. A., Magis-Rodriguez, C., Mays, V. M., Jimenez, R., & Patterson, T. L. (2012). The emerging HIV epidemic on the Mexico-U.S. border. Current HIV/AIDS Reports, 9(2), 134–143. https://doi.org/10.1007/s11904-012-0112-1
43. United Nations Office on Drugs and Crime. (2023). World drug report 2023. United Nations. https://www.unodc.org/unodc/en/data-and-analysis/world-drug-report-2023.html
44. Villatoro-Velázquez, J. A., Resendiz-Escobar, E., Mujica-Salazar, A., Bretón-Cirett, M., Cañas-Cortés, V., Soto-Hernández, I., & Medina-Mora, M. E. (2020). Encuesta nacional de consumo de drogas, alcohol y tabaco 2016-2017: Resultados de alcohol. Salud Mental, 43(2), 77–89. https://doi.org/10.17711/SM.0185-3325.2020.011
45. World Health Organization. (2022). Global status report on alcohol and health 2022. WHO Press. https://www.who.int/publications/i/item/9789240060098