Exploratory Data Analysis and Data Visualization on Accidental Drug Related Deaths

Ezinne Ogwo-Ude
Department of Science and Engineering
Solent University
Southampton, United Kingdom
2ogwoe12@solent.ac.uk

Raza Hasan
Department of Science and Engineering
Solent University
Southampton, United Kingdom
raza.hasan@solent.ac.uk

Shakeel Ahmad
Department of Science and Engineering
Solent University
Southampton, United Kingdom
shakeel.ahmad@solent.ac.uk

Salman Mahmood
Department of Computer Science
Nazeer Hussain University
Karachi, Pakistan
salman.mahmood@nhu.edu.pk

***Abstract*— The drug overdose epidemic in the United States is rapidly getting worse with substantial associated public health effects. Covering 10,654 cases over a decade (2012-2022), this study analyses an extensive dataset of accidental drug-related deaths in Connecticut. We process and analyze this data using Python and Tableau, we then use LSTM to predict how many people will show up in the designated intervals. The ages of individuals were stated as a mean value 43.52 (SD =12.60) years with a range between 13 and 87 years, bimodally distributed around the mid-30s to mid-50's. Overall, 74.14% were male and 85.48 % white in race/ethnicity. Significant increases were seen in accidental drug-related deaths. The most implicated substances were any opioids, fentanyl (alone or in combination), cocaine alone, heroin and ethanol. Crucially, some 89.05% of the cases had co-abuse with multiple drugs by one person who showed evidence that poly-substance use is commonplace in this community. Most deaths involved fentanyl (309, with a mode at 36 years (229 out of 309 deaths) and r (0.51) associated with 'any opioid', the primary cause of death). New Haven, Hartford and Fairfield counties stood out as hotspots for overdoses in geographic analysis. The LSTM model achieved a Root Mean Square Error (RMSE) of 7.25 and a Mean Absolute Error (MAE) of 5.62, predicting a sustained annual increase in deaths over the next three years. This federal and state partnership provides a model for using existing surveillance resources to inform targeted overdose intervention strategies, with an emphasis on the rise of fentanyl positivity among decedents.**

Keywords*— *Accidental drug-related deaths, data visualization, exploratory data analysis, machine learning, Tableau.

I. Introduction

In the United States, an increasing number of accidental drug overdose-and intoxication- related death represents a serious public health concern. Background The use of prescription medications, which include analgesics, opioids, stimulants and benzodiazepines has increased dramatically over the past decade [1]. In 2010, an estimated 5.1 million Americans used prescription pain relievers (including opioid analgesics) non-medically leading to approximately 425,000 Emergency Department visits related to poisoning and other complications from misuse of these drugs. The United States had largely avoided pandemic-related excess deaths but for the first time experienced an annual drug overdose death rate of 112,000 in a year as late as 2023 [2].

The Accidental Drug Related deaths dataset is a database that records all accidental drug-related deaths in Connecticut from 2012-2022 [3]. This data set contains detailed information on date of death/report, decedent demographics (age and sex), residence characteristics (urban/rural designation, poverty level at place of residence according to census tract) injury location and intent, summarizing cause-of-death codes describing the circumstances surrounding each fatal event along with more specific details about the actual nature of what caused it.

The aim of the current study is to analyze the trends of accidental drug-related deaths and predict their dynamics within a decade in Connecticut. The objectives are to analyze demographic characteristics, namely people's age, gender, and race who accidentally lost their lives due to drug related reasons; to analyze prevalent substances explored as the possible drug behind the death and the top substances associated with it, such as opioids and fentanyl; to analyze the poly-substance use which means some percentage of cases used a combination of drugs before an accidental death; to identify geographical variations the hot spots on the map where accidental drug-related deaths are highest in Connecticut; to forecast the trends within LSTM models, to forecast the development of drug-related behavior in the forthcoming three years; to provide recommendations concerning the aspects of work that require opportunities to reduce harm related to accidental drug-related behavior. To achieve these objectives, Python for data preprocessing and analysis and Tableau for data visualization, and LSTM for time series forecasting are used.

II. Literature Review

The increasing trend of accidental drug-related deaths is a major public health issue, necessitating a comprehensive analysis of demographic characteristics, substance use patterns, geographic hotspots, and predictive trends

A. Demographic Characteristics

A systematic review by [4] collected global descriptive epidemiologic information on unintentional drug overdose, identifying a number of limitations in the field diagnosis of prescription opioids and other drugs. The review noted that much of the existing data lacks specificity, such as age, gender and racial breakdowns needed to develop a deeper understanding who is struggling with drug overdoses. This is an important aspect of our study design to examine the demographic features between those who died from certified accidental drug-related manners in Connecticut.

B. Prevalent Substances and Poly-Substance Abuse

Including one that reviewed intentional, non-medical fentanyl use by people who use drugs (PWUD) [5] The study characterized marked heterogeneity in data collection and definitions employed to describe fentanyl use, underscoring the importance of consistent variables to help interpret substance use patterns. In addition, variability in definitions used and study populations were highlighted by [6] examined non-fatal prescription opioid overdoses. Such reviews highlight the need to assess commonly implicated substances in drug-related deaths, particularly opioids and fentanyl, as well as poly-substance abuse with up to 50% of cases where multiple drugs are present.

C. Geographic Hotspots

The work of [7] studied the opioid and related drug epidemics in rural Appalachia and discovered changes in patterns of usage as well as geographically concentrated areas. This review was specific to Kentucky, but it drives home the point that setting up "hot spot" monitoring systems is critical in locating where drug-related deaths are targeted. Through this study, we intend to carry on a geographic analysis and identify analogous hotspots in Connecticut which will help us understand regional variations and response needs.

D. Forecasting Future Trends

[8] explored some of the simulation models employed to tackle the opioid crisis, highlighting shortcomings in model calibration and validation. These models are useful in predicting future trends regarding drug-related deaths, consistent with our intent to use LSTM methods for forecasting the direction of change in drug-related mortalities across the state of Connecticut over 3 years. Reliable forecasting is critical for predicting future waves and guiding public health interventions.

E. Actionable Insights and Intervention Strategies

[9] review of the effects in road safety caused by polysubstance use. [10] also explored predictors of opioid-related mortality, again including prescriber behavior and user characteristics. This supports our aim to provide policymakers with actionable information designed to draw attention to key areas for policy and careful suggestions as solutions regarding the concerning trend of rising drug-related deaths in Connecticut.

Recent studies highlight the escalating opioid crisis in Connecticut, with a significant increase in overdose deaths from 2012 to 2018 [11]. Fentanyl has emerged as a primary driver, with its involvement rising from 4% of fatal overdoses in 2012 to 82% in 2019 [12]. Geographical disparities in fentanyl-related deaths were observed, with northeastern Connecticut showing higher risk [12]. Polysubstance use was identified as a critical factor, present in over 50% of fentanyl-related deaths [11]. To address this crisis, innovative approaches such as using internet search trends to forecast short-term overdose deaths have shown promise [13]. These findings underscore the need for targeted interventions and policies to combat the opioid epidemic, particularly focusing on fentanyl control and addressing polysubstance use [14][11].

The reviewed literature highlights the intricacies of the drug overdose epidemic and emphasizes that a comprehensive approach is essential to curtail it. This paper seeks to provide a comprehensive understanding of drug-related deaths in Connecticut by examining demographic characteristics, common drugs encountered in toxic used for decedents with overdose, poly-substance abuse patterns and geographic hotspots through time trend analyses. This, in turn, will allow for better-informed public health interventions and responses to the escalating epidemic of unintentional drug poisoning.

III. Methodology

The procedural steps followed in this study is being described in Fig. 1. Where the workflow involves data selection, preprocessing, exploratory data analysis (EDA), geospatial analysis, and predictive modelling [15].

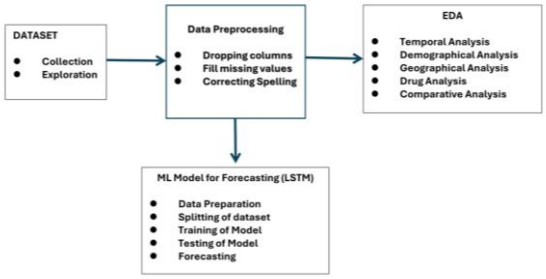

Fig. 1. Workflow

A. Data Selection

Accidental Drug Related Deaths, sourced from Data.gov and last updated on December 8, 2023 [2]. This dataset has 10,654 rows and 48 columns that provide unfiltered data on every accidental drug-related death in the state of Connecticut, from 2012 to 2022. Info provided includes date of death, age, race/ethnicity and substances present. Cleaning and exploratory data analysis was done in Python, and further visualization was developed using Tableau. For predictive modelling, an LSTM model was implemented to project trends in drug-related deaths for the next three years. Performance metrics of the trained model included the following key statistical metrics i.e. Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), the MAPE and R^2.

B. Loading of Dataset

Dataset was loaded into pandas DataFrame using pd.read_csv() function. The usual initial data exploration was performed to measure quality such as checking for missing values or outliers and seeing what the column types are.

C. Initial Exploration of the Dataset

Initial foray was to get the familiarity with data, what are different dimensions of dataset and how clean is it especially missing values or retractions. The only numerical variable included in the dataset was 'Age' and there were 47 categorical variables.

D. Data Preprocessing and Rationale for Choice

Data preprocessing is an important step for improving both the quality and usability of a dataset [16]. The missing values filled with ‘fillna’ based on the features that had less than 1% in each column (numerical: mean and categorical: mode). Drug-related columns had empty cells filled with 'NO' and were uppercased to keep it consistent across all entries. Duplicates were removed and a mapping dictionary was used to solve misspellings in state names, cities, or counties.

For the case of missing geographic values, we used previously established mappings to infer missing data for

locations by linking cities with their respective counties and counties where they were available with states [17]. These columns are deleted to technically streamline the dataset by removing content that does nothing but add noise. All remaining missing values were replaced with 'UNKNOWN' (to ensure the data entry was complete) [18]. Finally, data types for different columns were converted into the correct formats, date columns as datetime categorical column category type for analysis.

E. Data Exploration and Visualization

Data analysis was performed using Python and Tableau. Below is the analysis performed on the dataset.

1) Trend Analysis: Line plots were used to depict death, and percent change by year.

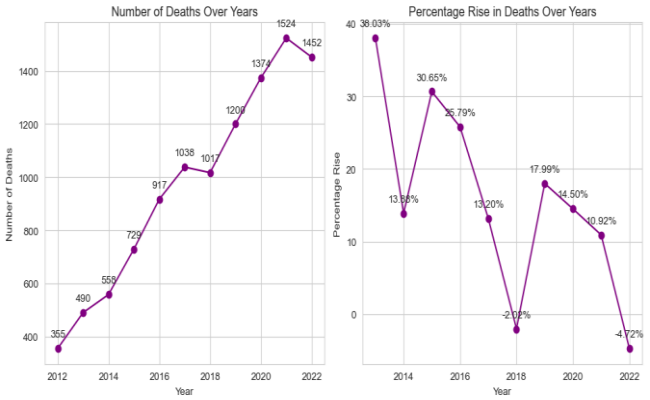

Fig. 2. Subplot of Number of Deaths and Percentage Rise (2012 to 2022).

2) Demographic Analysis: Age Distribution was analyzed by histogram and density plots.

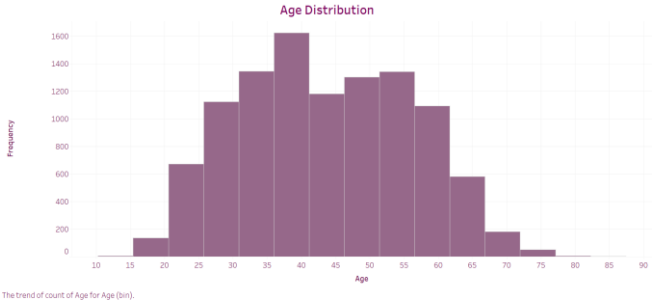

Fig. 3. Histogram of Age Distribution.

Fig. 4. Density Plot of Age.

3) Gender Distribution: Bar charts illustrated gender differences in drug-related deaths.

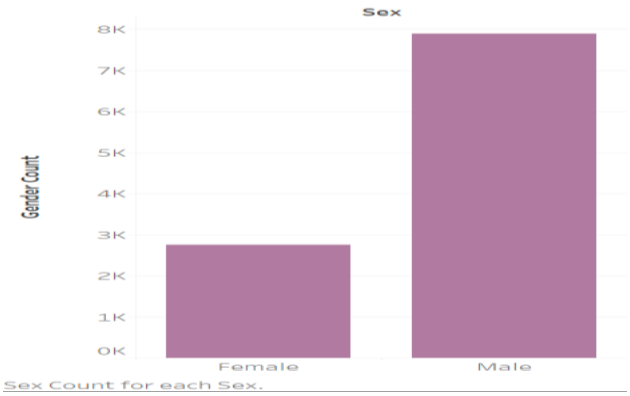

Fig. 5. Gender Distribution.

4) Race Distribution: Bar chart showing gender comparison for each death due to a drug..

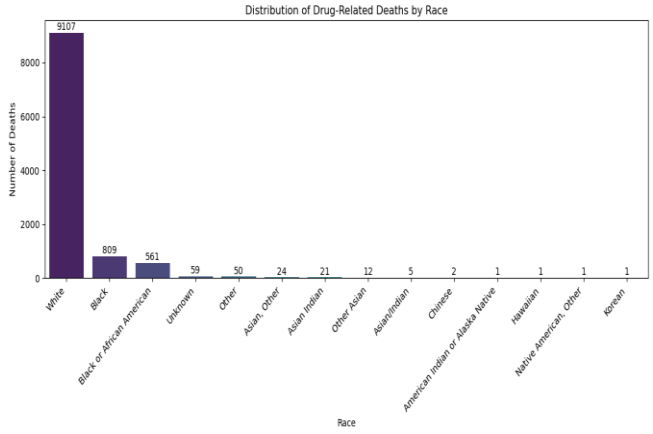

Fig. 6. Race Distribution.

5) Age Distribution by Gender: This analysis used box plot to show the distribution of age for males and females.

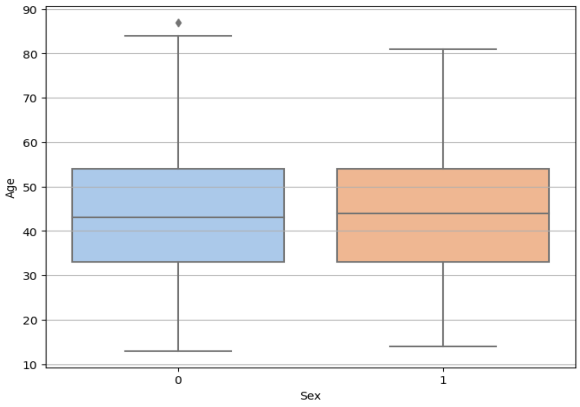

Fig. 7. Age Distribution by Gender.

6) Drug Analysis: Bar chart were made to display the type of drugs involved in different events.

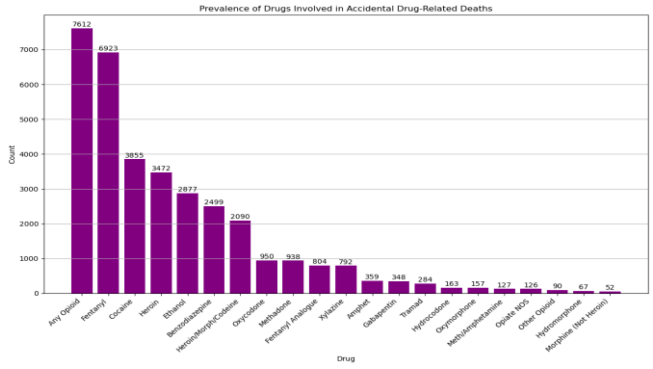

Fig. 8. Prevalence of Drugs.

7) Single vs. Multiple Drug Use: The percentages were calculated in terms single and multiple drug use cases.

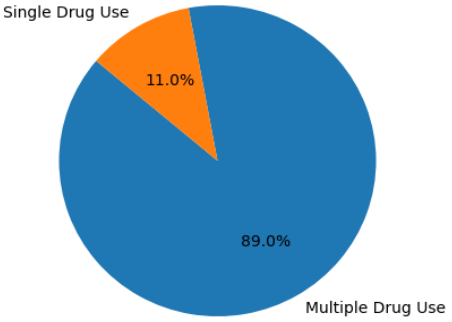

Fig. 9. Percentage of Single and Multiple Drug Use.

8) Drug Correlation Analysis: A correlation heatmap was generated to see how different drugs are related with each other.

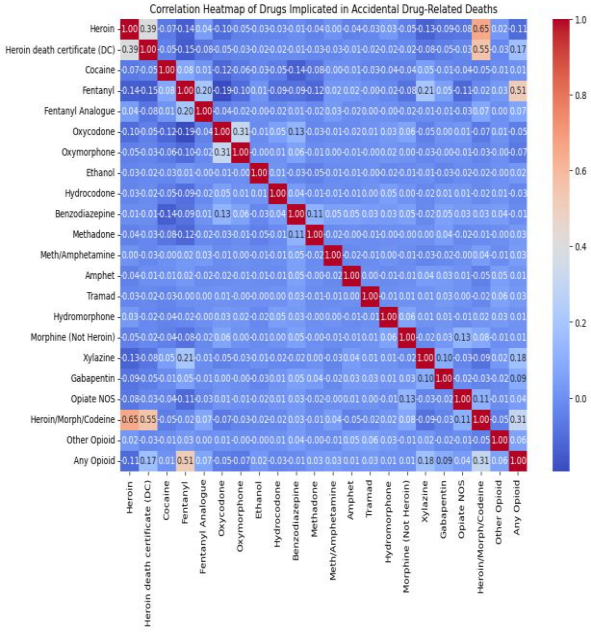

Fig. 10. Correlation Heatmap of Drugs.

9) Results Trends of Top Five Drugs: Trends of the top five drugs involved were analyzed.

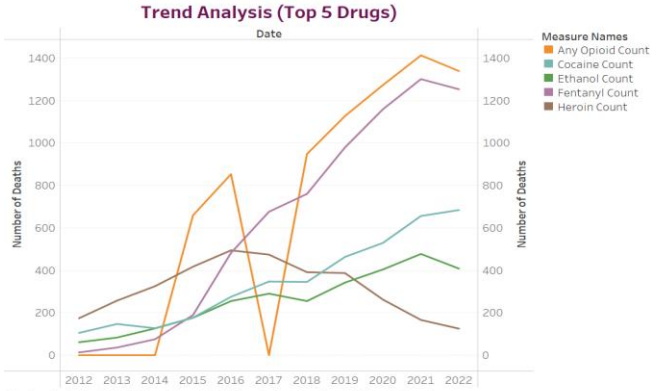

Fig. 11. Trend Analysis of Top 5 Drugs.

10) Geospatial Analysis: Geospatial Analysis-Maps was disseminated that displayed death rates by county.

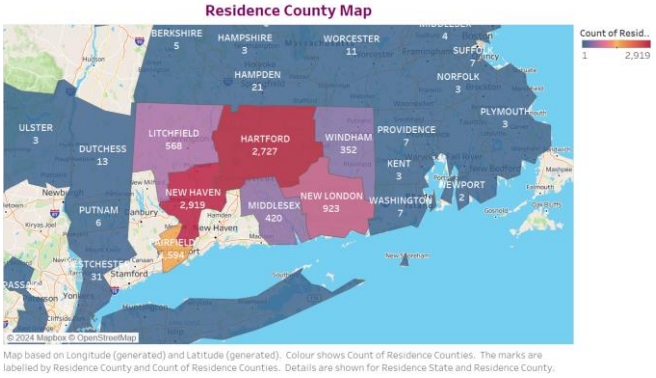

Fig. 12. Residence County Map.

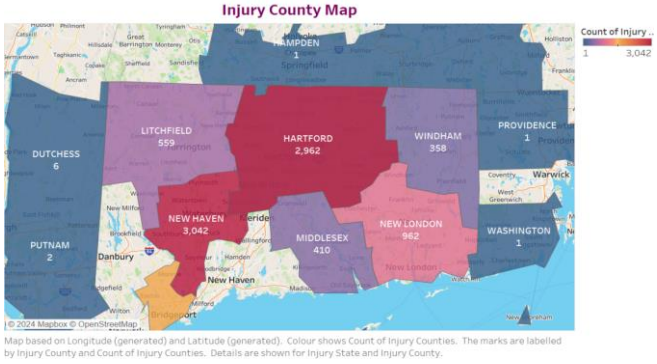

Fig. 13. Injury County Map.

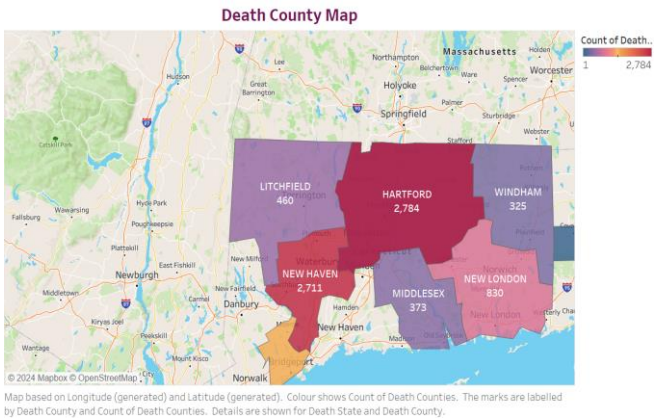

Fig. 14. Death County Map.

11) Building LSTM Model for Forecasting: We applied an LSTM model to predict drug-related deaths in the future. We resample the data into a set yearly frequency, scale it and then split to training & testing the dataset. The LSTM model was created with an LSTM layer of 50 units, followed by a dense layer. After 20 epochs of training, the model was then applied to predict trends for next three years *[19]*.

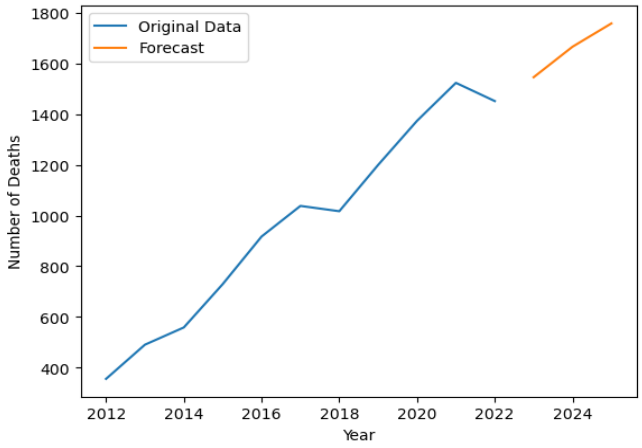

Fig. 15. Forecasting Future Trends in Drug-Related Deaths.

IV. Results

The LSTM model resulted in a 7.25 RMSE, an MAE of 5.62, and an MAPE of 8.4%. This model will be suitable since it shows good predictive accuracy for the upward trend of drug-related deaths. With its R-squared value of 0.82, 82% of the variance in the actual current data is explained by the model; hence, it is reliable to predict future trends.

Predictive analytics shows that another 20% increase in deaths due to drugs is expected by the year 2025. This is something of essence that calls for urgency in the need for public health intervention. Fig. 2 and 3: Accidental drug deaths in Connecticut, 2015-2022 show the trend of overdose death rates. Significant rises are observed in the COVID-19 pandemic phase. Geographically, Fig. 12 New Haven County was found as the leading most critical hotspot on overdose deaths while Hartford County had the highest total number of deaths. These are important insights into targeted public health strategies.

Gender divergence was observed in that males suffered a much higher ratio of accidental drug-related death rates versus females, as illustrated by bar graph Fig. 5. More than half of all drug-related deaths were among White residents, followed by Black and other races shown in Fig. 6. As shown in Fig., box plots showed that age distributions by sex, and it was obvious there existed several outliers in Fig. 7.

This included the population of drugs involved in these deaths, with opioids, fentanyl and cocaine being highly prevalent as illustrated on the bar graph in Fig. 8. Nearly all deaths (89%) involved cases of multiple drug use, as also seen in Fig. 9. The heatmap in Fig. 10 also showed a cluster for opioids and fentanyl, which enrapture the relevance of these drugs most to one another from the corelation analysis

Further analysis of fentanyl-related deaths, shown in Fig. 11, highlights a broad range of affected ages and a steady increase in such deaths since 2012. Overall, as shown in Fig. 13. Conversely, Hartford County had the highest number of deaths in comparison, as detailed in Fig. 14.

Finally, predictive analysis using the LSTM model presents continuous growth of drug-related deaths over next three years reflecting an ongoing emergency, as shown in Fig. 15.

V. Discussion

Assessing accidental drug-related deaths occurred in Connecticut over the span of a decade (2012 to 2022) reveals key trends and patterns. The data showed a general trend upwards for drug-related deaths, with considerable blips in the rate of change. Deaths doubled from 355 in 2016 to over a decade-high of and are now at their highest point ever to 1,452 annually. Potentially due to the impact of the COVID-19 pandemic on mental health and substance use.

The mean age was 43.5 years, with the greatest proportion in those aged between 35 and 40. Males were affected more than women, and a plurality of the victims was White which is in line with what you'd expect from the state's demographic profile. The age distributions and gender ratios suggest that specific strategies are needed to address the most sociodemographic affected group of men especially middle-aged. Opioid use is widespread, particularly involving fentanyl with a high frequency of multiple substance use (i.e., polysubstance) rather than single-substance opioids. The correlation analysis reveals a strong association of opioids with fentanyl, so there may be different strategies needed to address use of these compounds.

Overdose deaths hit harder in New Haven County, though Hartford and Fairfield also saw numbers well above the state average. Local interventions and the allocation of resources following utilitarian perspectives can be better informed by this geographical insight.

The LSTM model forecasts a continued increase in drug-related deaths, emphasizing the urgency of preventive measures and intervention strategies.

VI. Conclusion

The study provides insight into the trends, demographics and drug usage patterns contributing to accidental drug-related deaths from years 2012-2022. The trend analysis unambiguously shows an overall increase in drug-related deaths relative to the first year of formal measurement over roughly a decade, ranging from modest (and not persistent) annual increases around one-tenth of 1% up to jumps as large as ten times more compared with baseline. The age most impacted core to centre around the average 43-years-of-age group is between over-35 and 40 years. Gender analysis revealed higher risk among men and race distribution showed that White individuals were more affected.

Analysis of the drugs identified three main types: opioids, fentanyl, and cocaine leading to the highest number of substances experienced by the deaths. The fact that 89% of the deaths, particularly multiple drugs cases, were drug-addicted shows the complicatedness of the crisis. Finally, the significant correlation between opioids and fentanyl again point to the two crucial substances in the epidemic.

Geographic analysis identified New Haven County as the region with most overdose deaths including overdoses, whereas Hartford County as the highest number of deaths over counties we studied. Such results point to a need for more specific interventions in these regions.

Long-term predictive modelling using the LSTM methodology now indicates that there will be a further year-on-year increase in drug-related deaths over the next 3 years as this epidemic continues to unfold. Based on this particular forecast, continuous and increased public health approaches, programs for prevention as well as policy interventions are needed to combat various issues arising from drug related deaths. Future research should work to enhance predictive models, develop innovative intervention strategies and evaluate the underlying factors that contribute to geographic variation in drug-poisoning deaths.

In sum, the study illustrates that more targeted data-driven solutions are desperately needed in addressing this surge of drug-related deaths and stopping these deadly events from spreading into communities. The predictive accuracy of our LSTM model, reflected by an RMSE of 7.25, MAE of 5.62, and R^2 of 0.82, provides valuable insights into the future trajectory of drug-related deaths in Connecticut. The rise in fentanyl-related overdoses, as demonstrated by the strong correlation with opioids, calls for immediate public health actions. Geographic disparities highlight the need for targeted interventions in counties like New Haven and Hartford. Future research should build on these findings to develop more effective prevention strategies and predictive models.

REFERENCES

[1] CDC. (). *Medications for Opioid Use Disorder (MOUD) Study*. Available: https://www.cdc.gov/overdose-prevention/data-research/facts-stats/moud-study.html.

[2] B. Mann. (). *U.S. drug deaths declined slightly in 2023 but remained at crisis levels*. Available: https://www.npr.org/sections/health-shots/2024/05/15/1251239829/us-drug-overdose-deaths-provisional-2023#:~:text=But%20the%20toll%20from%20the.

[3] P. Zaldonis. (). *Accidental Drug Related Deaths 2012-2021*. Available: https://catalog.data.gov/dataset/accidental-drug-related-deaths-2012-2018.

[4] S. S. Martins *et al*, "Worldwide Prevalence and Trends in Unintentional Drug Overdose: A Systematic Review of the Literature," *American Journal of Public Health*, vol. 105, *(11)*, pp. 29, 2015. Available: https://www.ncbi.nlm.nih.gov/pubmed/26451760. DOI: 10.2105/ajph.2015.302843.

[5] V. W. L. Tsang *et al*, "Systematic review on intentional non-medical fentanyl use among people who use drugs," *Frontiers in Psychiatry*, vol. 15, pp. 1347678, 2024. Available: https://www.ncbi.nlm.nih.gov/pubmed/38414500. DOI: 10.3389/fpsyt.2024.1347678.

[6] M. J. Elzey, S. M. Barden and E. S. Edwards, "Patient Characteristics and Outcomes in Unintentional, Non-fatal Prescription Opioid Overdoses: A Systematic Review," *Pain Physician*, vol. 19, *(4)*, pp. 215–228, 2016. Available: https://www.ncbi.nlm.nih.gov/pubmed/27228510. DOI: 10.36076/ppj/2019.19.215.

[7] C. A. Schalkoff *et al*, "The opioid and related drug epidemics in rural Appalachia: A systematic review of populations affected, risk factors, and infectious diseases," *Substance Abuse*, vol. 41, *(1)*, pp. 35–69, 2020. Available: https://www.tandfonline.com/doi/abs/10.1080/08897077.2019.1635555. DOI: 10.1080/08897077.2019.1635555.

[8] M. Cerdá *et al*, "A Systematic Review of Simulation Models to Track and Address the Opioid Crisis," *Epidemiologic Reviews*, vol. 43, *(1)*, pp. 147–165, 2022. Available: https://www.ncbi.nlm.nih.gov/pubmed/34791110. DOI: 10.1093/epirev/mxab013.

[9] E. Beaulieu *et al*, "Impacts of alcohol and opioid polysubstance use on road safety: Systematic review," *Accident Analysis and Prevention*, vol. 173, pp. 106713, 2022. Available: https://dx.doi.org/10.1016/j.aap.2022.106713. DOI: 10.1016/j.aap.2022.106713.

[10] N. B. King *et al*, "Determinants of Increased Opioid-Related Mortality in the United States and Canada, 1990–2013: A Systematic Review," *American Journal of Public Health (1971)*, vol. 104, *(8)*, pp. e32–e42, 2014. Available: https://www.ncbi.nlm.nih.gov/pubmed/24922138. DOI: 10.2105/AJPH.2014.301966.

[11] T. G. Rhee *et al*, "Accidental drug overdose deaths in Connecticut, 2012–2018: The rise of polysubstance detection?" *Drug and Alcohol Dependence*, vol. 205, pp. 107671, 2019. Available: https://dx.doi.org/10.1016/j.drugalcdep.2019.107671. DOI: 10.1016/j.drugalcdep.2019.107671.

[12] H. Lu *et al*, "Geographic and temporal trends in fentanyl-detected deaths in Connecticut, 2009–2019," *Annals of Epidemiology*, vol. 79, pp. 32–38, 2023. Available: https://dx.doi.org/10.1016/j.annepidem.2023.01.009. DOI: 10.1016/j.annepidem.2023.01.009.

[13] S. Mukherjee *et al*, "Using internet search trends to forecast short term drug overdose deaths: A case study on connecticut," in Dec 2020, pp. 1332–1339.

[14] B. D. L. Marshall *et al*, "Epidemiology of fentanyl-involved drug overdose deaths: A geospatial retrospective study in Rhode Island, USA," *The International Journal of Drug Policy*, vol. 46, pp. 130–135, 2017. Available: https://www.clinicalkey.es/playcontent/1-s2.0-S0955395917301330. DOI: 10.1016/j.drugpo.2017.05.029.

[15] E. Camizuli and E. J. Carranza, "Exploratory Data Analysis (EDA)," *The Encyclopedia of Archaeological Sciences*, pp. 1–7, 2018. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119188230.saseas0271. DOI: 10.1002/9781119188230.saseas0271.

[16] C. Fan *et al*, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Frontiers in Energy Research*, vol. 9, 2021. Available: https://www.frontiersin.org/articles/10.3389/fenrg.2021.652801/pdf. DOI: 10.3389/fenrg.2021.652801.

[17] K. Sanjar *et al*, "Missing Data Imputation for Geolocation-based Price Prediction Using KNN–MCF Method," *ISPRS International Journal of Geo-Information*, vol. 9, *(4)*, pp. 227, 2020. Available: https://search.proquest.com/docview/2388666805. DOI: 10.3390/ijgi9040227.

[18] H. Kang, "The prevention and handling of the missing data," *Korean Journal of Anesthesiology*, vol. 64, *(5)*, pp. 402–406, 2013. Available: https://www.ncbi.nlm.nih.gov/pubmed/23741561. DOI: 10.4097/kjae.2013.64.5.402.

[19] H. Yi *et al*, "ACP-DL: A Deep Learning Long Short-Term Memory Model to Predict Anticancer Peptides Using High-Efficiency Feature Representation," *Molecular Therapy. Nucleic Acids*, vol. 17, pp. 1–9, 2019. Available: https://dx.doi.org/10.1016/j.omtn.2019.04.025. DOI: 10.1016/j.omtn.2019.04.025.