**Accidental Drug Related Deaths 2012-2020 (Connecticut)**

*Please note that the web interface has been modified since the presentation. The homepage and ‘Missing values’ pages were edited and additional information was added into the new analysis option ‘Dataset View’.*

**Overview**

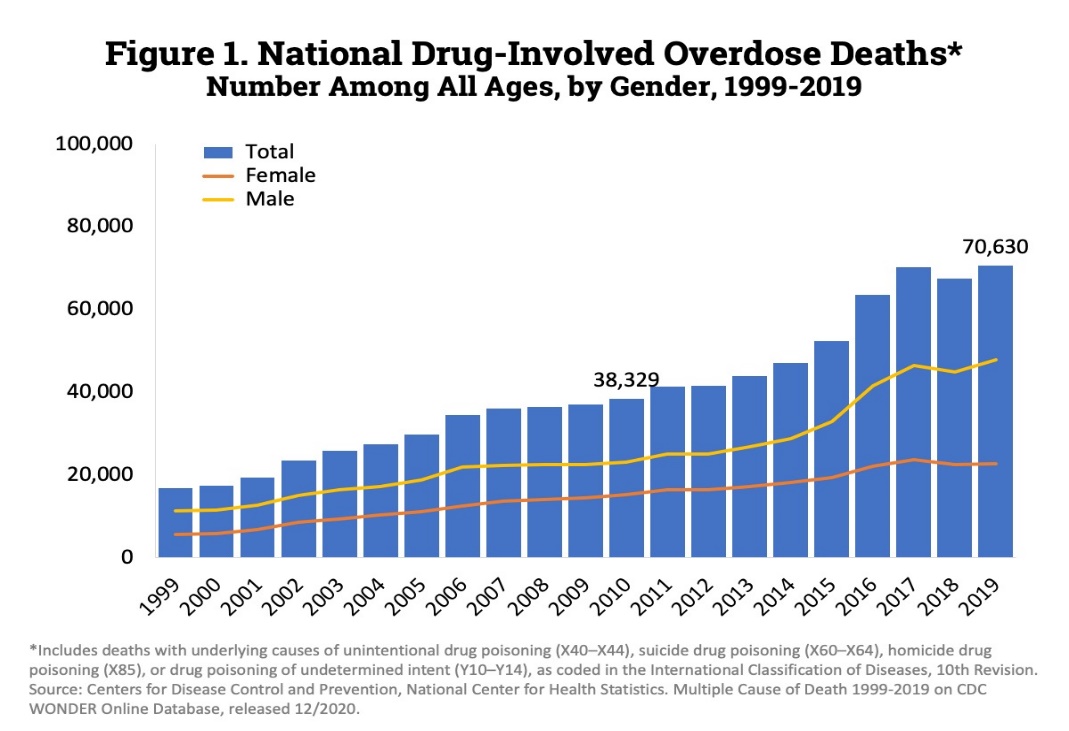
Illegal drug use and drug abuse has plagued the United States since early 19th century, but this issue has likely been present prior to such systematic research and efforts to control drug abuse. The prevalence of drug use in the United States as of 2020 is approximately 19%, with over 31.9 million illegal drug users aged 12 or older, and if including the drugs tobacco and alcohol – which are not illegal, but regardless very harmful to human health – this figure rises to over 53 million (NCDAS, 2021). During the time between the beginning of 2020 and present day, there has been over 700,000 recorded drug overdoses, and in the same year the federal government issued a $35 billion budget for drug control, but the results were not very significant (NCDAS, 2021). The rise of drug-involved overdose deaths increased relatively consistently from 1999 to 2019 (Fig. 1), and there were significantly more incidences occurring among men compared to among women (National Institute on Drug Abuse, 2021).

Figure 1. National Drug-Involved Overdose Deaths. Number among all ages, by gender, 1999-2019. Graph adapted from NCDAS found at: <https://www.drugabuse.gov/drug-topics/trends-statistics/overdose-death-rates>.

More specifically, 22% of American men and 17% of American women have reported use of illegal drugs, and the most common age range in which such incidences occur is 18-25 years of age (NCDAS, 2021). A more detailed review of age distribution of illegal drug use in the United States is highly alarming: approximately 5% of 8th-graders, 20% of 10th-gradesr, 20% of 12th-graders, and 47% of all students who have not yet completed 12th grade have reported using illegal drugs at least once. Moreover, the exposure of children younger than five years of age to marijuana increased by 1.5-fold over a seven-year period (NCDAS, 2021).

Due to the significant impact and prevalence of drug abuse incidences in the United States, I have decided to conduct research on drug-related deaths in the state that I am currently residing in, Connecticut.

**Dataset**

Source

The dataset “Accidental Drug Related Deaths 2012-2020” was acquired from Data.ct.gov, which can be accessed through [clicking this link](https://catalog.data.gov/dataset/accidental-drug-related-deaths-2012-2018). This dataset did not specify any licensing information but was classified as a public dataset and is intended for public access and use. Moreover, the dataset could be downloaded as CSV, RDF, JSON, and XML formats, and the metadata file is available for download on the same webpage as a JSON file. This dataset was published on November 12th of 2020, and the latest update occurred on November 29th of 2021.

The preprocessed dataset included 7,679 entries and 41 columns (variables), including:

* Drug types:
* Heroin
* Cocaine
* Fentanyl
* Fentanyl analogue
* Oxycodone
* Oxymorphone
* Ethanol
* Hydrocodone
* Benzodiazepine
* Methadone
* Hydromorphone
* Xylazine
* Amphet
* Tramad
* Morphine (not heroin)
* Other
* Opiate NOS
* Any Opioid
* Incidence ID
* Incidence recorded date
* Date type
* Age
* Sex
* Race
* Residence city, county, and state
* Death city, county, and state
* Location if ‘other’
* Description of injury
* Injury place, city, county, and state
* Cause of death
* Manner of Death
* Geos of death city, residence city, and injury

|  |  |  |
| --- | --- | --- |
| FAIR Components | Description | Satisfies Component? |
| Findable | The dataset was easily findable on Data.ct.gov, which can also be accessed with Data.gov. The data is downloadable in various human- and machine-readable formats, such as JSON, RDF, and XML, and metadata is in JSON format. | Yes |
| Accessible | The dataset was accessible as it does not require any authorization process for access or download, was classified as a public dataset, and the metadata was equally accessible. | Yes |
| Interoperable | The dataset and metadata are downloadable as formal and broadly applicable languages (CSV, JSON, RDF, and XML). | Yes |
| Reusable | The dataset and metadata were described with accurate and relevant attributes (variables) but did not specify license. | Moderately |

FAIRness (GO FAIR, 2021)

Processing

After data cleaning, the dataset contained 7679 entries and 28 variables. The cleaning process is described below in the subsections: ***Redundant Variables***, ***Data Type***, and ***Missing Values***.

*Redundant Variables*

The dataset variables included multiple violations of PHI according to HIPAA standards, such as GeoSpacing to below the state level for residency, death, and injury, as well as longitude and latitude information provided for these variables. Such violations to PHI were removed from the dataset. Moreover, due to the fact that the dataset is focused on the Connecticut region, there is very little reason to conduct analysis by state or region. Instead, analysis focused on age, gender, missing values, and descriptive statistics of each of the 17 drug types. Removed columns involved the following:

* 'injury city'
* 'injury county'
* 'injury state'
* 'manner of death'
* 'deathcitygeo'
* 'residencecitygeo'
* 'injurycitygeo'
* ‘other significant conditions’
* 'residence city'
* 'residence county'
* 'residence state'
* 'death city'
* 'death county'
* 'location'
* 'location if other'

*Data Type*

The data type of the columns ‘date’ and ‘age’ were changed to dtype = ‘datetime’ and dtype = ‘float’ respectively. The column names were changed to lower case, such that analysis would be more efficient.

*Missing Values*

The drug type columns (Specified in **Dataset 🡪 Source** section) originally only marked detected drug used with a ‘Y’ in the corresponding row, whereas the drug uses not detected or not present were simply left blank. However, this may not be an accurate representation of the data, as in missing value analysis, such blank cells would be interpreted as missing data and not ‘drug not used’. Therefore, all blank cells in the drug type column were filled with ‘N’.

Furthermore, the column ‘other significant conditions’ was removed because the column contained > 90% missing values (‘NaN’). This column did not have its blank cells filled in with ‘N’ as did the drug type columns because it is hard to intuitively determine whether blank cells represented the absence of other significant conditions or the lack of information collected, therefore, to avoid confusion, this column was removed.

Summary Statistics

The summary statistics analysis and some general data explorations were included within the web interface as the analysis options ‘Age’, ‘Age by sex’, and ‘Dataset View’, and they are reproduced below:

**74.37% of cases were of men.**

**25.47% of cases were of women.**

**The other cases are of gender "Unknown".**

**The average age in dataset is 42.48 years old.**

Chart, line chart

Description automatically generated

Figure 2. Time series of the number of accidental drug-related deaths in Connecticut from 2012-2020.

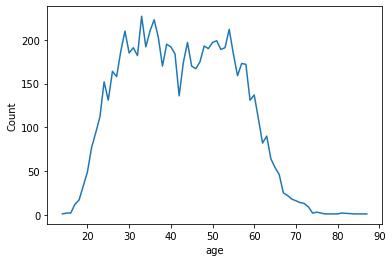


Figure 3. Number of accidental drug-related deaths by age in Connecticut from 2012-2020.

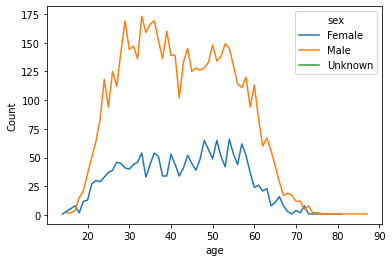


Figure 4. Number of accidental drug-related deaths by gender and age in Connecticut from 2012-2020. Blue line indicates female incidences, orange line indicates male incidences, and green (barely visible, towards older ages) indicates incidences of unknown gender.

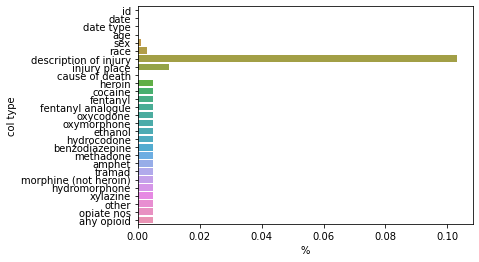


Figure 5. Percent of each column that are missing values. The column type is on the y-axis labeled col-type and the percentage of specific column that is missing values in on the x-axis.

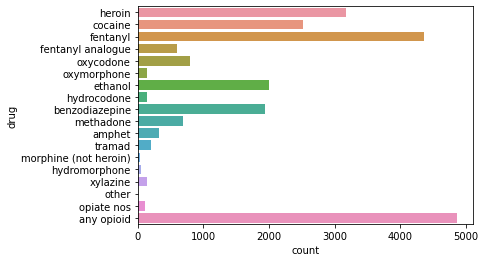


Figure 6. Number of accidental drug-related deaths by drug type in Connecticut from 2012-2020. The drug type is on the y-axis labeled col-type and the number of incidences in on the x-axis.

**Web Interface**

Design and Development

The goal that propelled the design and development of this web interface is to answer the following questions:

* What is the overall trend in accidental drug-related deaths in Connecticut from 2012 to 2020?
* What does the age distribution look like?
* How complete is the dataset?
* What is the distribution of missing data?
* Does age, gender, and race distribution look difference for different types of drug use?
* How many total cases for each drug?
* Which types of drug incidences are more frequent?

The web interface was designed using Flask and written in Python, HTML, CSS, and utilized Jinja2 templates. The main code could be found in the file app.py, and the html templates include: index.html (homepage), info.html (all analysis except dataset view), datasetview.html, and failure.html.

The Home page provides a screenshot (with clickable link to data source website) and a brief description of the data source with a clearly identified link to the data source website.

The webpage is designed to adapt to device width, such that even if one were to view the page on their phones, the words will not shrink to fonts so small that they are illegible.

All pages except the Home page includes a clickable link that redirects user to Home page.

Selection of invalid analysis options or options with no data available (e.g., ‘Drug type: other’) will redirect the user to failure.html.

All data processing and preparation for presentation were dynamically performed in the APIs in app.py, including index(), info(), by\_age(), by\_gender\_age(), missing\_vals(), show\_drug\_type\_stats(drug), and dataset\_view(analysis\_type). The option to select a type of analysis from the 21 options listed below was sent from the Home page (index.html) to the View Information page (info.html) or processed by info() to be sent to datasetview.html if the selected option were Dataset View. The former then receives files sent from the corresponding APIs depending on chosen analysis type via the send\_file() method, and displays the information in html format. The latter renders the datasetview.html page and displays information there instead of in the info.html page.

For more specific information on the processing of data, please refer to the code attached in the appendix section of this report.

Analysis

The analysis that are available to view on the web interface is the following:

* Home:
  + Time series plot of the number of incidences in Connecticut from 2012 to 2020
  + The number of incidences by drug type
* Age: A line plot of the number of incidences by age
* Age by sex: A line plot of the number of incidences by age and gender
* Missing values:
  + The total number of missing values in the cleaned dataset
  + A bar plot of the percent of each column that is occupied by missing values
* Dataset View:
  + The proportion of cases that were of each gender
  + The average age in the dataset
  + The complete cleaned dataset (table format)
* Drug type: {drug}
  + A labeled line plot of the age distribution of incidences for the selected drug
  + The number of total incidences in dataset related to selected drug
  + The number of male and female cases within the total incidence count
  + The average age
  + The most seen race

Results

The results of above analysis are described below:

* Home:
  + The overall trend in drug usage was a relatively consistent increase from 2012 to 2020, with a small dip in 2018 that conforms with the national trend in that year, and a significant increase in the years 2016-2017.
  + The most frequently seen drug in the dataset is opioid, which includes all analogues and derivatives of opioid drugs. The second most seen is fentanyl, and the third most seen is heroin.
* Age:
  + The age range in which most incidences occur is approximately 20-55.
  + Despite the average age in dataset being approximately 42 years old, there was a significant dip in the plot around age 42.
  + The plot is slightly skewed to the right, indicating that more young Americans in Connecticut have accidental drug-related deaths compared to older individuals.
  + There are very few incidences of people under age 20 or over age 65.
* Age by sex:
  + The age distributions for both genders are very similar to the overall age distribution and to each other.
  + There are significantly more incidences of men than of women.
  + The occurrence of incidences in the gender category ‘Unknown’ is insignificant, but surprisingly occurred in ages > 70 years.
* Missing values:
  + There is a total of 1,539 missing values in the cleaned dataset.
  + The most prominent proportion of missing values occurred within the column ‘description of injury’ at approximately 10%.
  + The drug type columns all consistently had 0.35% of their data missing.
  + The columns ‘id’, ‘date’, ‘age’, and ‘cause of death’ all had 0 missing values.
* Dataset View:
  + Almost three-quarters of the dataset are occupied by incidences of men, over one-quarter by incidences of women, and the rest are of gender "Unknown".
  + The average age in dataset is 42.48 years old.
  + The complete dataset without analysis is presented.
* Drug type: {drug}
  + The age distribution for most drug types had the mode age range of approximately 25-55.
  + Some age distributions (heroin, cocaine, fentanyl, oxycodone, ethanol, amphet, and any opioid) were slightly right skewed.
  + The number of total incidences in dataset related to valid drug type selection ranged from 40 to 4366.
  + Age data for drug types of fewer incidences tend to produce graphs that are not as smooth, and hence less informative of a clear trend. Perhaps due to this, some of them seemed left skewed (morphine (not heroin), tramad, methadone, and hydrocodone). However, if such results were accurate, then it implies that such incidences tend to occur among older individuals.
  + The number of male and female cases within the total incidence count varied by drug, but in general there were more male incidences than female incidences.
  + The average age for all drug types were slightly over 40 years old.
  + The most seen race is White for all drug type analyses, but this does not imply that White people are more likely to have such incidences compared to other races, but just that the base population is majority White, hence White people are statistically more likely to appear in this dataset.

Limitations

**References**

GO FAIR. (2021, October 18). *Fair principles*. GO FAIR. Retrieved December 21, 2021, from <https://www.go-fair.org/fair-principles/>

NCDAS. (2021, September 25). *Substance abuse and addiction statistics [2021]*. Drug Abuse Statistics. Retrieved December 18, 2021, from <https://drugabusestatistics.org/>

National Institute on Drug Abuse. (2021, February 25). *Overdose death rates*. National Institute on Drug Abuse. Retrieved December 20, 2021, from <https://www.drugabuse.gov/drug-topics/trends-statistics/overdose-death-rates>

<https://stackoverflow.com/questions/28207761/where-does-flask-look-for-image-files>

<https://www.google.com/url?sa=i&url=https%3A%2F%2Fgiphy.com%2Fgifs%2FYaleAlumni-transparent-yale-yalealumni-YlYX7mrtL1R4Z5huWJ&psig=AOvVaw2a6huJn2rxb6Yo8GubeCVx&ust=1640110782628000&source=images&cd=vfe&ved=0CAsQjRxqFwoTCMDxoPf-8vQCFQAAAAAdAAAAABAD>

<https://www.rawpixel.com/image/1201197/pink-watercolor-background>

<https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.istockphoto.com%2Fvector%2Fwhite-exclamation-mark-symbol-on-red-circle-caution-icon-isolated-on-white-gm1248649489-363706105&psig=AOvVaw384yr0anRu2wlW5SnvqxiB&ust=1640113583889000&source=images&cd=vfe&ved=0CAwQjhxqFwoTCKjqna-J8_QCFQAAAAAdAAAAABAD>