User Guide

Drug Mention with Involvement (DMI) Methodology Tool (DMI2EpiTool) for Identifying Drugs Involved in Drug Poisoning Deaths and Single- vs. Poly- Drug Poisoning Death Classification

Contents

Overview

The DMI2EpiTool contains a suite of programs that allow identification of specific drugs and substances, mentioned by the medical certifiers on death certificate records in the context of involvement in drug poisoning death. The DMI methodology was developed in 2016 by the National Center for Health Statistics (NCHS) and the U.S. Food and Drug Administration (FDA) and described in a paper by Trinidad et al (Trinidad et al., 2016). The CDC/NCHS DMI program is available on GitHub (CDC, 2019). The DMI methodology has been used in several publications for monitoring the most frequently involved drugs in the U.S. drug overdose deaths (Hedegaard et al., 2018; Trinidad et al., 2016; Warner et al., 2016).

DMI2EpiTool

The DMI2EpiTool builds on the DMI methodology and the NCHS DMI program. The current version of the DMI2EpiTool is an upgrade of the NCHS DMI software and has the same core search logic. New processing was added to improve the association of phrases to DMIs, as well as to find complex search terms containing digits and dashes. Performance issues were also addressed.

The ability to monitor trends in drug overdose mortality where a drug of interest was the only drug mentioned with involvement (vs. polydrug involvement) is important for surveillance and for informing policy and regulatory decision making.

The DMI2EpiTool includes a new module that evaluates if a drug poisoning death was a Single-substance or Poly-substance drug poisoning, based on the number of “referent drugs” (see section Terminology) that were mentioned with involvement. The tool identifies and reports alcohol involvement but alcohol was not considered a drug, in line with the ICD-10 coding and classification of drug poisoning (WHO, 2019).

Questions and requests for help could be submitted to DMI2EpiTool@uky.edu

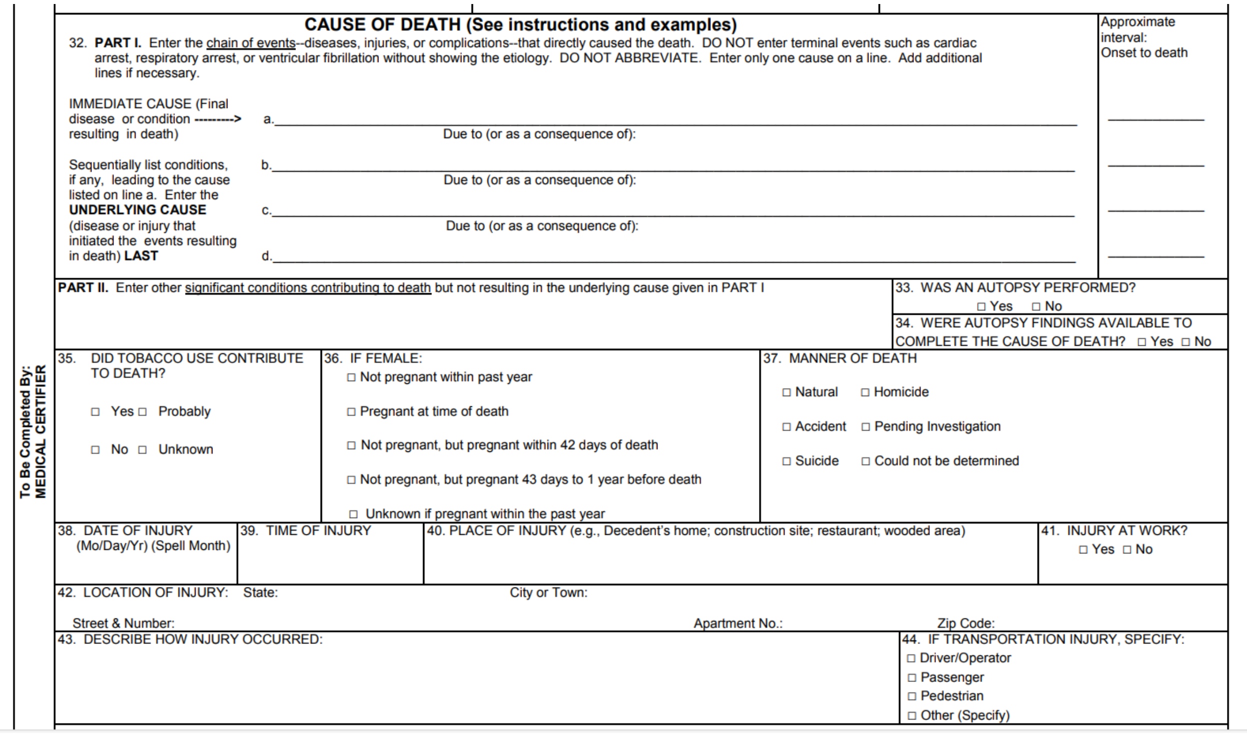
Terminology

Death Certificate (DC)

The death certificate is a legal document, certifying someone’s cause and manner of death. The content of a death certificate follows the U.S. Standard Certificate of Death (NCHS, 2003). The death certificates are filed with the Office of Vital Statistics in the state where the death occurred. A copy of the death certificate is sent electronically to the NCHS, where the cause and manner of the death are coded with ICD-10 codes for an underlying and multiple causes of death. The DMI2EpiTool uses an input death certificate file that contains selected variables/fields to 1) identify drug poisoning (overdose) deaths, and 2) identify drugs mentioned with involvement in the drug poisoning deaths.

Literal text

The DMI2EpiTool works mainly with the three sections on the U.S. Standard Certificate of Death to identify mentioning of drugs:
1) CAUSE OF DEATH, PART I (lines a, b, c, d, describing the events that directly caused the death);
2) CAUSE OF DEATH, PART II (significant conditions contributing to death); and
3) DESCRIBE HOW INJURY OCCURRED (field #43).
The DMI analysis refers to the information in these three sections as *literal text*.



To Be Completed By: MEDICAL CERTIFIER

CAUSE OF DEATH (See instructions and examples)

32. **PART I.** Enter the chain of events--diseases, injuries, or complications--that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest, or ventricular fibrillation without showing the etiology. DO NOT ABBREVIATE. Enter only one cause on a line. Add additional lines if necessary.

Approximate interval: Onset to death

IMMEDIATE CAUSE (Final disease or condition ---------> a.__________ resulting in death) Due to (or as a consequence of): ______

Sequentially list conditions, if any, leading to the cause listed on line a. Enter the **UNDERLYING CAUSE** (disease or injury that initiated the events resulting in death) **LAST**

b.__________ Due to (or as a consequence of): ______

c.__________ Due to (or as a consequence of): ______

d.__________ ______

PART II. Enter other significant conditions contributing to death but not resulting in the underlying cause given in PART I

33. WAS AN AUTOPSY PERFORMED? □ Yes □ No

34. WERE AUTOPSY FINDINGS AVAILABLE TO COMPLETE THE CAUSE OF DEATH? □ Yes □ No

35. DID TOBACCO USE CONTRIBUTE TO DEATH?
□ Yes □ Probably
□ No □ Unknown

36. IF FEMALE:
□ Not pregnant within past year
□ Pregnant at time of death
□ Not pregnant, but pregnant within 42 days of death
□ Not pregnant, but pregnant 43 days to 1 year before death
□ Unknown if pregnant within the past year

37. MANNER OF DEATH
□ Natural □ Homicide
□ Accident □ Pending Investigation
□ Suicide □ Could not be determined

38. DATE OF INJURY (Mo/Day/Yr) (Spell Month)

39. TIME OF INJURY

40. PLACE OF INJURY (e.g., Decedent's home; construction site; restaurant; wooded area)

41. INJURY AT WORK? □ Yes □ No

42. LOCATION OF INJURY: State: City or Town:
Street & Number: Apartment No.: Zip Code:

43. DESCRIBE HOW INJURY OCCURRED:

44. IF TRANSPORTATION INJURY, SPECIFY:
□ Driver/Operator
□ Passenger
□ Pedestrian
□ Other (Specify)

Drug Mention with Involvement (DMI)

DMI is defined as a mention of a drug, a drug class, or drug exposure in the literal text fields, excluding mentions where the contextual information suggested that the drug was not involved in the death.

Drug Search term (Search Term)

A drug search term could be a single word or a combination of words that identifies a drug mention. A drug search term could represent a generic, brand, or street drug name (e.g., “FENTANYL”, “XANAX”, “CRACK COCAINE”), a drug metabolite name (e.g., “NORBUPRENORPHINE”, “6-MAM”), or a drug misspelling (e.g., “FENTANIL”, “HEROINE”).

Special Search Term

A combination of words (e.g., “WITH ANPP PRESENT”) that could be interpolated (inserted between) parts of the literal text, that doesn’t fit the normal sequence of the literal text.

Phrase (Contextual Phrase)

A combination of words and asterisks (e.g., “HISTORY OF * ABUSE”) in which the DMI’s meaning and/or involvement is determined by the adjacent words.

Joining Phrase

A combination of words and asterisks that indicates linkage or exchangeability of DMIs (e.g., “* AS WELL AS *”).

Qualifier (Descriptor)

A word or words that provide information on drug characteristics or characteristics of drug exposure, such as “MULTIPLE” “PRESCRIPTION” or “NON-PRESCRIPTION”.

Principal Variant

The overarching label assigned to a drug, drug class or other exposure from the related search terms as originally described by Trinidad et al (Trinidad et al., 2016). In general, the principal variant represents the generic drug name or recognized chemical name. Search terms for combination products are mapped to principal variants for each component in the product (e.g., “PERCOCET” maps to “OXYCODONE” and “ACETAMINOPHEN” principal variants)

Referent Drug

A *referent drug* is a specific drug that is listed as an ‘*active moiety*’ in the FDA’s Global Substance Registration System (https://precision.fda.gov/uniisearch). If a drug mention is not itself an active moiety, but is rather a prodrug, precursor, contaminant, or metabolite of an active moiety, then that drug mention is assigned to the related active moiety (e.g., “NORBUPRENORPHINE” is assigned to the referent drug “BUPRENORPHINE”).

Note 1: There are 3 types of referent drugs in the DMI2EpiTool: 1) specific referent drugs (e.g., “FENTANYL”), 2) class-level referent drugs (e.g.,”CANNABINOID”, “OPIOID”), or 3) non-specific referent drug (e.g., ”DRUG”, “SUBSTANCE”).

Note 2: There is a referent drug “ALCOHOL”, which is counted separately from the other referent drugs in the analytical output because alcohol is not considered a drug in the ICD-10 classification system framework.

Note 3: There are search terms that are ambiguous and thus cross-walked to a referent drug “AMBIGUOUS” (e.g., “METHETAMINES” is an ambiguous search term that could mean “METHENAMINE” or “METHAMPHETAMINE”, triggering a flag for manual review in the analytical output/ SAS file “literal_matching_output*”)

Single-drug poisoning

Drug poisoning death with exactly one identified referent drug (with or without alcohol involvement).

Poly-drug poisoning

Drug poisoning death with more than one identified referent drug.

DMI2EpiTool General Requirements

These programs were created for use on a Windows PC running Windows 10 or later version.
These programs were tested on SAS Base 9.4; they should be able to run with only minor modifications on SAS Studio and SAS Enterprise guide.

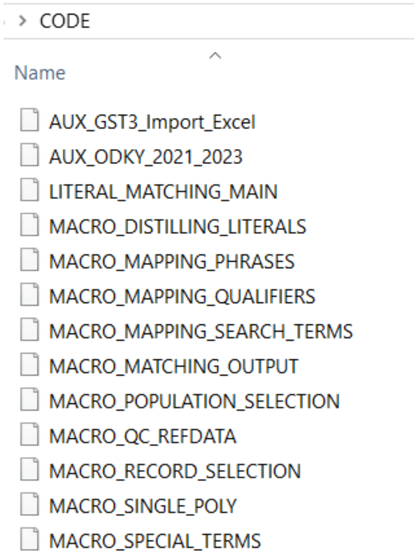
Directory Structure

The root folder could be any local directory or network share. There is no specific requirement about the name of the root folder, while the subfolders are organized in the following way:

CODE folder

Contains SAS code for searching the literal text for drugs/substances:

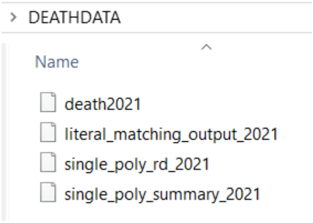
- LITERAL_MATCHING_MAIN.sas – Reads folder locations, prepares, and processes the data, generates main output datasets.
- MACRO_POPULATION_SELECTION.sas – filters the death certificates based on state of residency.



AUX_GST3_Import_Excel
AUX_ODKY_2021_2023
LITERAL_MATCHING_MAIN
MACRO_DISTILLING_LITERALS
MACRO_MAPPING_PHRASES
MACRO_MAPPING_QUALIFIERS
MACRO_MAPPING_SEARCH_TERMS
MACRO_MATCHING_OUTPUT
MACRO_POPULATION_SELECTION
MACRO_QC_REFDATA
MACRO_RECORD_SELECTION
MACRO_SINGLE_POLY
MACRO_SPECIAL_TERMS

- MACRO_RECORD_SELECTION.sas - Selects records based on ICD10 codes criteria (ICD UCD underlying cause of death, ICD MMCD multiple cause of death). Also cleans the literal text for symbols, white space, etc.
- MACRO_MAPPING_SEARCH_TERMS.sas - Locates search terms within literal text.
- MACRO_MAPPING_QUALIFIERS.sas - Locates qualifiers/descriptors within literal text.
- MACRO_DISTILLING_LITERALS.sas - Replaces search terms and various qualifiers with '*', reducing the complexity of the text the phrase search program must execute on.
- MACRO_MAPPING_PHRASES.sas - Searches the literal text for specific phrases.
- MACRO_SPECIAL_TERMS.sas – Searches the DC literal for interpolated text and creates two or more new literals from the separated parts.
- MACRO_SINGLE_POLY.sas – Applies the Single-drug vs. Poly-drug calculation algorithm and produces additional analytical datasets.
- MACRO_QC_REFDATA.sas – Derives general_search_terms dataset from three level data structure (datasets gst_level1, gst_level2, gst_level3) and checks for consistency.

DEATHDATA folder



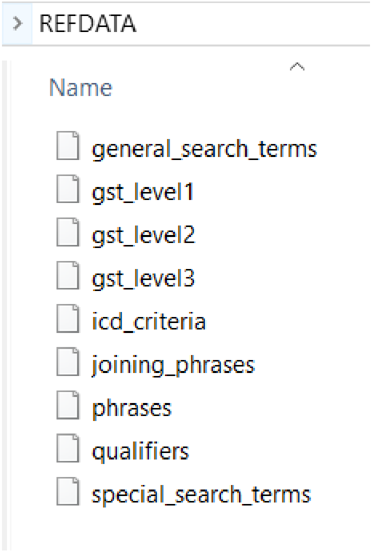
death2021
literal_matching_output_2021
single_poly_rd_2021
single_poly_summary_2021

- DEATHXXXX.sas7bdat – input death certificates dataset for year XXXX
- LITERAL_MATCHING_OUTPUT_XXXX.sas7bdat – mention level output file with all found search terms, qualifiers, and phrases.
- SINGLE_POLY_SUMMARY_XXXX.sas7bdat - person level output file reporting the summary data used for determining if the death was caused by single-drug or poly-drug poisoning.
- SINGLE_POLY_RD_XXXX.sas7bdat - referent drug level output file reporting the summary data used for determining if the death was caused by single-drug or poly-drug poisoning.

REFDATA folder

The folder contains the program settings and lookup datasets used in processing of death certificates data. The user can add new search terms or re-map existing as needed.

- ICD_CRITERIA.sas7bdat - list of ICD-10 codes to filter input death records to drugs considered involved in the death.



REFDATA

Name

general_search_terms
gst_level1
gst_level2
gst_level3
icd_criteria
joining_phrases
phrases
qualifiers
special_search_terms

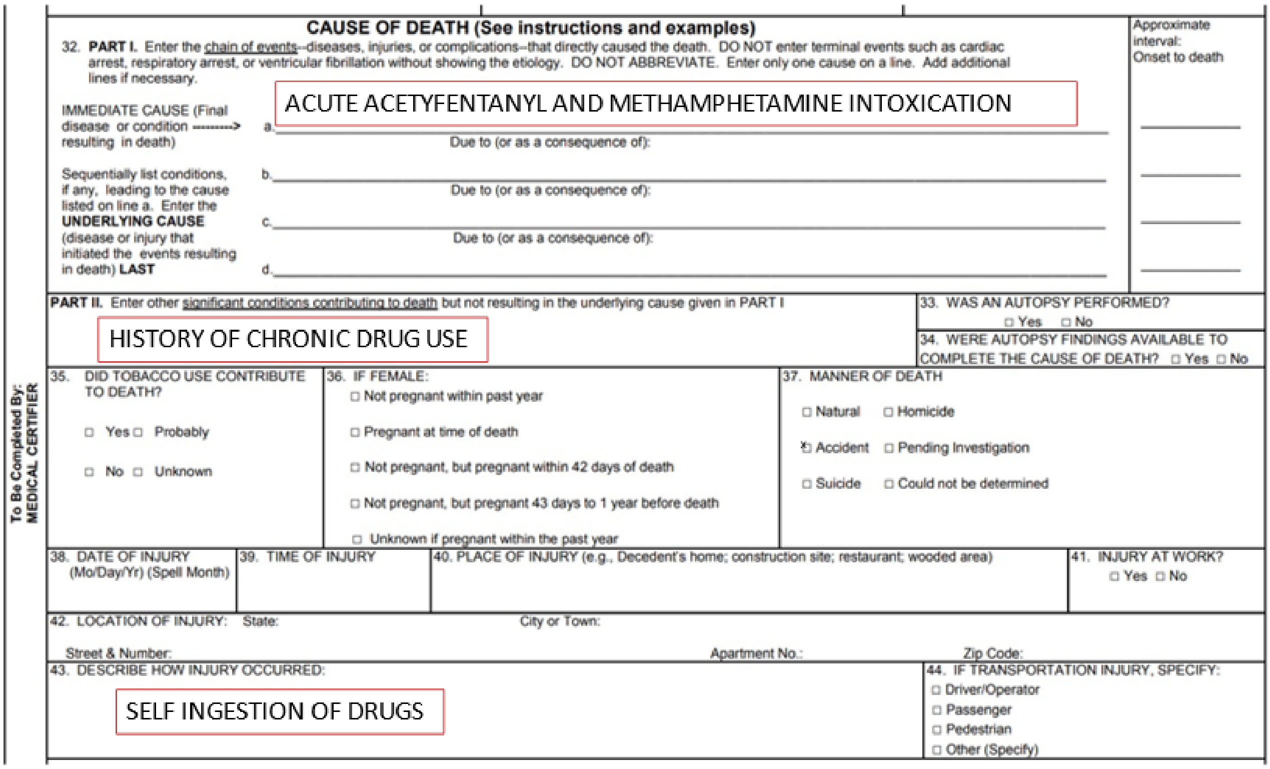
- ○ GENERAL_SEARCH_TERMS.sas7bdat - list of drug names to search against the literal text and mappings to higher level groupings. It is derived from the following 3 datasets:
  - ▪ GST_Level1.sas7bdat – list of search terms and corresponding principal variants
  - ▪ GST_Level2.sas7bdat – list of principal variants and corresponding referent drugs
  - ▪ GST_Level3.sas7bdat – list of unique referent drugs with properties like specificity, common name, ATC codes, etc.
- ○ JOINING_PHRASES.sas7bdat – joining phrases to search against the literal text that link the found drug mentions.
- ○ PHRASES.sas7bdat —phrases to search against the literal text that contain the found drug mention with qualifiers and joining phrases.
- ○ QUALIFIERS.sas7bdat - terms to search against the literal text appearing next to the found drug mentions.
- ○ SPECIAL_SEARCH_TERMS.sas7bdat – special literals that could be inserted between parts of the literal text.

Set-up Instructions

These files are currently provided ‘as is’ and may require modifications; The following instructions should be followed in general, after editing the programs to run in your environment.

I. Inspect all the files and save them in a designated location on your computer.
Your location could be a local PC folder, a File Server share, or a SAS server.

II. Prepare death certificates SAS dataset with at least the following fields (variable names in brackets are the Kentucky naming example, you will see those in the source code LITERAL_MATCHING_MAIN.sas). Below is an example of the text fields in a DC we are interested in (Part I. a,b,c,d ; Part II.; 43.)



To Be Completed By: MEDICAL CERTIFIER

CAUSE OF DEATH (See instructions and examples)

32. PART I. Enter the chain of events--diseases, injuries, or complications--that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest, or ventricular fibrillation without showing the etiology. DO NOT ABBREVIATE. Enter only one cause on a line. Add additional lines if necessary.

Approximate interval: Onset to death

IMMEDIATE CAUSE (Final disease or condition ---------> resulting in death) a. ACUTE ACETYFENTANYL AND METHAMPHETAMINE INTOXICATION

Due to (or as a consequence of):

Sequentially list conditions, if any, leading to the cause listed on line a. Enter the **UNDERLYING CAUSE** (disease or injury that initiated the events resulting in death) **LAST**

b. ________

Due to (or as a consequence of):

c. ________

Due to (or as a consequence of):

d. ________

PART II. Enter other significant conditions contributing to death but not resulting in the underlying cause given in PART I

HISTORY OF CHRONIC DRUG USE

33. WAS AN AUTOPSY PERFORMED? □ Yes □ No

34. WERE AUTOPSY FINDINGS AVAILABLE TO COMPLETE THE CAUSE OF DEATH? □ Yes □ No

35. DID TOBACCO USE CONTRIBUTE TO DEATH?

□ Yes □ Probably

□ No □ Unknown

36. IF FEMALE:

□ Not pregnant within past year

□ Pregnant at time of death

□ Not pregnant, but pregnant within 42 days of death

□ Not pregnant, but pregnant 43 days to 1 year before death

□ Unknown if pregnant within the past year

37. MANNER OF DEATH

□ Natural □ Homicide

☒ Accident □ Pending Investigation

□ Suicide □ Could not be determined

38. DATE OF INJURY (Mo/Day/Yr) (Spell Month)

39. TIME OF INJURY

40. PLACE OF INJURY (e.g., Decedent's home; construction site; restaurant; wooded area)

41. INJURY AT WORK? □ Yes □ No

42. LOCATION OF INJURY: State: City or Town:

Street & Number: Apartment No.: Zip Code:

43. DESCRIBE HOW INJURY OCCURRED:

SELF INGESTION OF DRUGS

44. IF TRANSPORTATION INJURY, SPECIFY:

□ Driver/Operator

□ Passenger

□ Pedestrian

□ Other (Specify)

A. **Required** fields:

1. Unique identifier for the DC

In our environment we generate it by concatenating the death certificate's Year (`dc_year`), Volume Number `(dc_volno)` and Number `(dc_num)` like that:

```
uniq_id=STRIP(dc_year)||STRIP(dc_volno)||STRIP(dc_num);
```

To de-identify the data, you can add SAS code to encrypt this value, it would not affect further processing.

2. Immediate Cause of death (Part I line a) for the DC (`c_under_mc`)
3. Underlying Cause 1 (Part I line b) (`c_supp_mc1`)
4. Underlying Cause 2 (Part I line c) (`c_supp_mc2`)
5. Underlying Cause 3 (Part I line d) (`c_supp_mc3`)
6. Significant contributing conditions (Part II) (`c_sigcc_mc`)
7. Injury Description (filed 43) (`i_descript`)

B. Optional fields – they are often used in reporting and analysis:

1. Underlying cause of death ICD-10 code (`c_under_c`)

2. Multiple/supplemental ICD-10 codes related to the death (`c_supp_c1 .. c_supp_c20`).
3. Manner of Death (`dt_manner`)
4. Certifier (`certifier`)

You can have as many additional optional fields from the DC as you need.

The naming convention for the death certificates dataset is the word “death” followed by four digits year, for example DEATH2021.sas7bdat.

The single- vs poly-drug methodology is developed in the context of drug poisoning deaths. Drug poisoning deaths are captured with an underlying cause-of-death variable values in the following ICD-10 range: X40-X44, X60-X64, X85, Y10-Y14.

The following code identifies the KY residents who died from drug poisoning in year 2021:

```
proc  sql ;
create table death2021  as

SELECT *
FROM Death2021
WHERE (dc_year=2021)
      and  (r_state_fp='KY')
      and  ( UPCASE(c_under_c) IN
      ('X40','X41','X42','X43','X44',
      'X60','X61','X62','X63','X64',
      'X85',
      'Y10','Y11','Y12','Y13','Y14'));
quit;
```

III. Modify LITERAL_MATCHING_MAIN.sas

1. Change the path to your macros SAS files with the actual location for your environment :

```
%include 'c:\dmi\CODE\macro_population_selection.sas';
%include 'c:\dmi\CODE\macro_record_selection.sas';
%include 'c:\dmi\CODE\macro_mapping_search_terms.sas';
%include 'c:\dmi\CODE\macro_mapping_qualifiers.sas';
%include 'c:\dmi\CODE\macro_distilling_literals.sas';
%include 'c:\dmi\CODE\macro_mapping_phrases.sas';
%include 'c:\dmi\CODE\macro_matching_output.sas';
%include 'c:\dmi\CODE\macro_special_terms.sas' ;
%include 'c:\dmi\CODE\macro_single_poly.sas' ;
%include 'c:\dmi\CODE\macro_qc_refdata.sas' ;
```

2. Change the path to your input/output library `indir`, and your reference library `refdir`:

```
libname indir  'c:\dmi\deathdata';
libname refdir 'c:\dmi\refdata';
```

3. Set the year for which the processing will be done in macro variable *year*, for example :

   ```
   %let year=2021;
   ```

   Only one year of data can be processed in a single run of the program unless modifications are made.

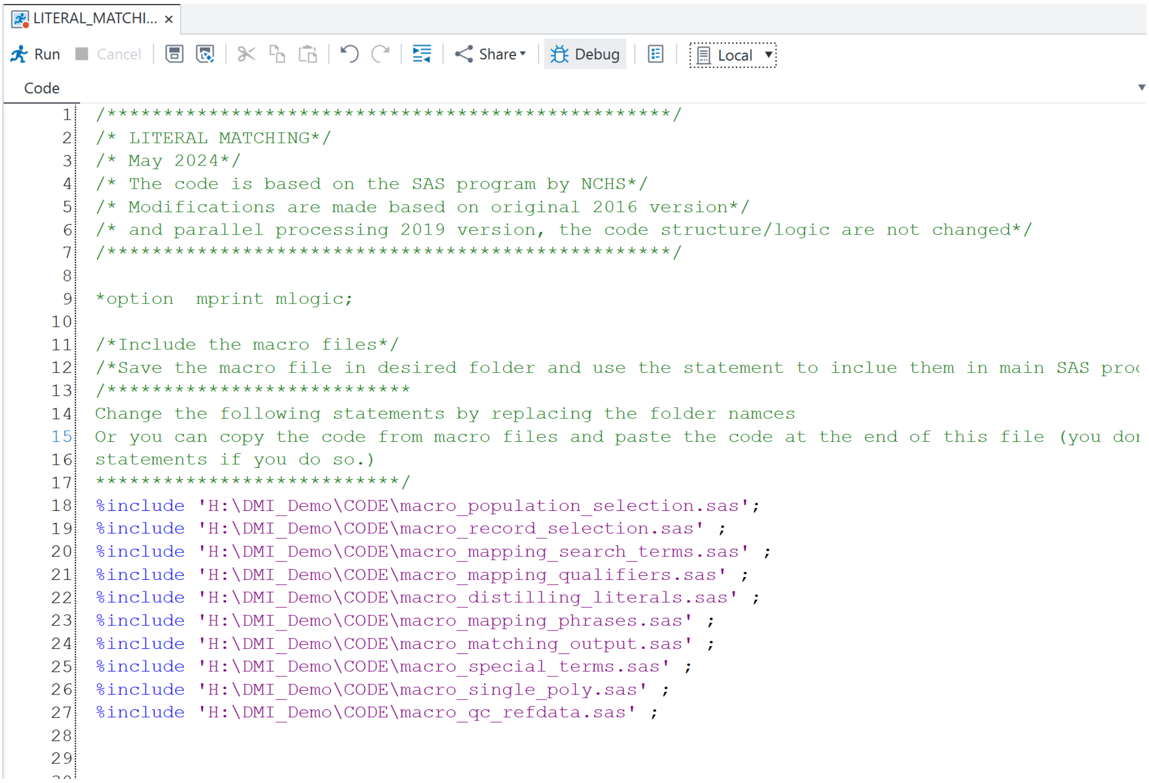
4. Modify the code that uses your specific death certificate dataset variables:

```
uniq_id=STRIP(dc_year)||STRIP(dc_volno)||STRIP(dc_num);

DESCR_LIN1=c_under_mc;
DESCR_LIN2=c_supp_mc1;
DESCR_LIN3=c_supp_mc2;
DESCR_LIN4=c_supp_mc3;

DESCR_LIN5=c_sigcc_mc;

INJ_DESCR =i_descript;
```

Variables on the left of the assignment statement are used across the program to represent DC fields.

IV. Run LITERAL_MATCHING_MAIN.sas

Press ‘Run’ button in your SAS environment.



```
/**************************************************/
/* LITERAL MATCHING*/
/* May 2024*/
/* The code is based on the SAS program by NCHS*/
/* Modifications are made based on original 2016 version*/
/* and parallel processing 2019 version, the code structure/logic are not changed*/
/**************************************************/

*option  mprint mlogic;

/*Include the macro files*/
/*Save the macro file in desired folder and use the statement to inclue them in main SAS proc
/***************************
Change the following statements by replacing the folder namces
Or you can copy the code from macro files and paste the code at the end of this file (you don
statements if you do so.)
***************************/
%include 'H:\DMI_Demo\CODE\macro_population_selection.sas';
%include 'H:\DMI_Demo\CODE\macro_record_selection.sas' ;
%include 'H:\DMI_Demo\CODE\macro_mapping_search_terms.sas' ;
%include 'H:\DMI_Demo\CODE\macro_mapping_qualifiers.sas' ;
%include 'H:\DMI_Demo\CODE\macro_distilling_literals.sas' ;
%include 'H:\DMI_Demo\CODE\macro_mapping_phrases.sas' ;
%include 'H:\DMI_Demo\CODE\macro_matching_output.sas' ;
%include 'H:\DMI_Demo\CODE\macro_special_terms.sas' ;
%include 'H:\DMI_Demo\CODE\macro_single_poly.sas' ;
%include 'H:\DMI_Demo\CODE\macro_qc_refdata.sas' ;
```

Wait until the processing is finished, monitoring the Log messages.
The processing runs in two passes based on the value of the derived variable ‘p’ (from ‘pass’, possible values 1,2) from gst_level1 dataset, variable Search_Term. Search terms without digits and dashes have p=1, while search terms with digits and dashes have p=2.

There is a hardcoded parameter - ‘batchsize’ (the simultaneous number of search terms being processed), which affects the performance:
```
%let batchsize=500;
```

The DMI2EpiTool, version from June 2024, has a search term list of about 12,000 search terms (drugs and non-drugs included in count), which means we will process about 24 batches of search terms. You can follow the progress by looking at the batch number currently in process in SAS Log window.

Upon successful completion, several new output files will be created in **DEATHDATA** folder (Note: XXXX is the 4 digit year (2021 in our demo):

LITERAL_MATCHING_OUTPUT_XXXX.sas7bdat
SINGLE_POLY_SUMMARY_ XXXX.sas7bdat
SINGLE_POLY_RD_XXXX.sas7bdat

Output

A mention level output file with all found search terms, qualifiers for the terms, phrases and other information for each mention will be created as a result of the program execution.

Here is a data dictionary of the important variables in the main output files:

LITERAL_MATCHING_OUTPUT_XXXX.sas7bdat

Name	Type	Length	Sample Value	Description
uniq_id	Character	36	2021123456	unique identifier for the deceased
DC_VOLNO	Numeric	8	123	dc volume number
DC_NUM	Numeric	8	456	dc number
search_term	Character	96	FENTANYL	general search term found in the DC literal
phrase	Character	70	ACUTE * TOXICITY	phrase from the Phrases dataset in which the search term was found and assigned
qualified_term	Character	100	FENTANYL	qualified term (qualifiers from Qualifiers dataset, if any) concatenated with the general search term found)
text_field	Character	32	CLEANED_CHAIN	text only part (punctuation removed) of the DC where term was found, could be CLEANED_CHAIN (Part I), CLEANED_DESCR_LIN5 (Part II), cleaned_INJ_DESCR (43.)
literal_text	Character	490		the original text (punctuation preserved) of the text_field
Principal_Variant	Character	36	FENTANYL	The substance name of the next level to the search term
Delete_Flag	Character	1		Y if the search term was non-drug, blank otherwise
I_DESCRIPT	Character	150		Injury Description (DC field 43.)
C_UNDER_MC	Character	100	ACUTE FENTANYL TOXICITY	Immediate cause of death (DC Part I.a)
C_SUPP_MC1	Character	100		Immediate cause of death (DC Part I.b)
C_SUPP_MC2	Character	67		Immediate cause of death (DC Part I.c)

C_SUPP_MC3	Character	68		Immediate cause of death (DC Part I.d)
C_SIGCC_MC	Character	200		Significant conditions contributing to death (DC Part II.)
C_UNDER_C	Character	4	X40	ICD code for the death
C_SUPP_C1_C20	Character	200	X44,T404,T436,T509	Comma separated supplemental ICD codes list for the death
DT_MANNER	Character	21	ACCIDENT	Manner of death
CERTIFIER	Character	30	DEPUTY CORONER	Name of the certifier
Excl_Part1	Character	1	X	Code for exclusion of the phrase, if found in Part I.
Excl_Part2	Character	1		Code for exclusion of the phrase, if found in Part II.
Excl_InjD	Character	1		Code for exclusion of the phrase, if found in Injury Description

SINGLE_POLY_SUMMARY_ XXXX.sas7bdat

Name	Type	Length	Sample Value	Description
uniq_id	Character	36	2021123456	unique identifier for the deceased
num_spec_drugs	Numeric	8	1	number of specific referent drugs
num_class_level_drugs	Numeric	8	1	number of class-level referent drugs
num_nonspec_drugs	Numeric	8	0	number of non-specific referent drugs
alcohol_involved	Character	1	1	Indicator (1 if ALCOHOL referent drug was involved, 0 otherwise)
ambiguous_involved	Character	1	0	Indicator (1 if AMBIGUOUS referent drug was involved, 0 otherwise)
single_drug_poisoning	Character	1	N	Indicator (Y if exactly one* referent drug was involved, N otherwise)

poly_drug_poisoning	Character	1	Y	Indicator (Y if more than one* referent drug was involved, N otherwise)
list_ref_drugs	Character	1000	ALCOHOL,FENTANYL,OPIOID	Comma separated list of found referent drugs
DC_Part_1	Character	200	FENTANYL TOXICITY	Immediate cause of death (DC Part I.)
DC_Part_2	Character	200	DECEDENT USE OF ALCOHOL	Significant conditions contributing to death (DC Part II.)
DC_Inj_Descr	Character	150	INGESTION OF OPIOIDS	Injury Description (DC field 43.)

SINGLE_POLY_RD_XXXX.sas7bdat

Name	Type	Length	Sample Value	Description
Referent_Drug	Character	100	FENTANYL	name of a particular referent drug found
uniq_id	Character	36	2021123456	unique identifier for the deceased
num_spec_drugs	Numeric	8	1	number of specific referent drugs
num_class_level_drugs	Numeric	8	1	number of class-level referent drugs
num_nonspec_drugs	Numeric	8	0	number of non-specific referent drugs
alcohol_involved	Character	1	1	Indicator (1 if ALCOHOL referent drug was involved, 0 otherwise)
ambiguous_involved	Character	1	0	Indicator (1 if AMBIGUOUS referent drug was involved, 0 otherwise)
single_drug_poisoning	Character	1	N	Indicator (Y if exactly one* referent drug was involved, N otherwise)
poly_drug_poisoning	Character	1	Y	Indicator (Y if more than one* referent drug was involved, N otherwise)
list_ref_drugs	Character	1000	ALCOHOL,FENTANYL,OPIOID	Comma separated list of found referent drugs

DC_Part_1	Character	200	FENTANYL TOXICITY	Immediate cause of death (DC Part I.)
DC_Part_2	Character	200	DECEDENT USE OF ALCOHOL	Significant conditions contributing to death (DC Part II.)
DC_Inj_Descr	Character	150	INGESTION OF OPIOIDS	Injury Description (DC field 43.)
Term_Description	Character	12		Specificity of found referent drug. Possible values are CLASS, NON-SPECIFIC or blank(meaning specific)
Specific_Indicator	Numeric	8	1	Indicator (1 if referent drug is specific, 0 otherwise)
Classlevel_Indicator	Numeric	8	0	Indicator (1 if referent drug is class-level, 0 otherwise)
Nonspec_Indicator	Numeric	8	0	Indicator (1 if referent drug is non-specific, 0 otherwise)
Alcohol_Indicator	Numeric	8	0	Indicator (1 if referent drug is ALCOHOL, 0 otherwise)

The rest of this page was intentionally left blank.

Workflow

- The main processing is done through running LITERAL_MATCHING_MAIN.sas.
- Code is assembled by bringing the separate SAS macros into the current program.
- Input/output and referent data folders are established.
- Annual input DC dataset is configured.
- Referent datasets are uploaded.
- Literal text variables are assigned values from the input DC dataset variables.
- Only non-empty uniq_id records are selected for processing. In case you need specific subset of states, change the code to invoke the macro: e.g., %***POPULATION_SELECTION***(stres=US_Residents)
- If filtering of the death certificates by U.S./state residency is required, make the necessary changes in the file MACRO_POPULATION_SELECTION.sas (the variable name R_STATE_FP is Kentucky-specific and may be renamed)
- Further, the processing is done in two passes. They work on different subsets of general search terms based on the presence of dashes and digits in the term. For example: ‘ANPP’ and ‘4-ANPP’ would produce identical principal variant results, but they are processed in the first and second pass respectively. This is done to ensure the program doesn’t miss any DMI due to typos or misspellings in the literals.
- The input DC fields are grouped and punctuation is removed (the process is referred as ‘cleaning’) in MACRO_RECORD_SELECTION.sas
- Dash symbol ‘-‘ and digits ‘0’..’9’ are removed from the cleaned DC fields on the first pass, but they are not removed on the second pass.
- Special terms are processed to retrieve the interpolated text from the literal text fields. Often, the medical certifier who writes the DC text would insert in between the sequentially ordered substances some other text. For example, in the text ‘ACUTE COMBINED DRUG (ILLEGAL FENTANYL [4-ANPP DETECTED] & METHAMPHETAMINE) | TOXICITY’ we see a note in square brackets [4-ANPP DETECTED], which serves to add information about the substance found, but it breaks the flow of the listed substances FENTANYL and METHAMPHETAMINE, and would prevent the program to correctly associate the terms with the phrase to which they belong.
- As a workaround, we added code to retrieve interpolated text like ANPP DETECTED and process it separately from the rest of the DC literal text. The interpolated texts (a.k.a ‘special search terms’) are listed in the referent dataset SPECIAL_SEARCH_TERMS. Currently we have identified about 60 such special search terms, based on review of recent KY death certificates. This list should be maintained in the current order, which is “grouped by text similarity” and in each group the longer text has higher priority.
- Since we are processing search terms in two passes, the second pass for special terms has some additional processing to keep some special terms found in pass one. For example, if 'MAM' was found on pass one and '6-MAM' is a search term to process in pass two, we would want to keep the word ‘MAM’ in the literal text for the second pass.

- For each special term found in a particular DC (identified by its uniq_id) we create a new linked uniq_id_x (x=1,2 and so on) and process them as if they were regular DC literals.
- We also process the literal text of the original DC with all special terms removed.
- At the end we combine the DMIs found in the DC special terms and the rest of the DC literal text.
- The next step is to find search terms in the DC literal text (sequentially, in each of the three cleaned text fields - cleaned_CHAIN , cleaned_DESCR_LIN5, cleaned_INJ_DESCR). For each found drug search term (mention), a new record is created with the DC uniq_id, the text field, and the starting and the ending position in the field in MENTIONS dataset.
- Note: The search term list is partitioned in blocks of 500 (in our environment this was near optimal value) during the search process. The value could be changed in MACRO_MAPPING_SEARCH_TERMS.sas, but in our experience when its value is above 4,000 it may lead to breaking the SAS environment.
- The next step is to search for phrases immediately before and after the 'mentions' found in the previous step, so called pre- and post- qualifiers. Those are defined in dataset QUALIFIERS. The list could be expanded and based on manual data review. The results are output into QUALIFIED_MENTIONS dataset.
- The next step is to replace all related qualified terms from QUALIFIED_MENTIONS with a wild character "*"; the process is called “distilling literals”. This is done for each of the three cleaned text fields of a particular DC.
- For example, if for the cleaned_CHAIN field ‘ACUTE COMBINED DRUG ILLEGAL FENTANYL METHAMPHETAMINE TOXICITY’, the qualified terms found are ‘ILLEGAL FENTANYL’ and ‘METHAMPHETAMINE’, then the resulting distilled literal will be ‘ACUTE COMBINED DRUG * TOXICITY’.
- Relation patterns are defined in dataset JOINING_PHRASES. Pattern “**” means two consecutive qualified terms (like in the above example), pattern “* AND *” means two qualified terms joined with conjunction “and”, and so on (the list could be expanded based on manual data review).
- The distilled results are output into DISTILLED_LITERALS dataset.
- The next step is to assign the contextual phrase to the search terms it contains. This is an important process, as it provides the opportunity to check the context, in which each search term is used, and decide (automatically - thru PHRASES dataset settings, or manually - thru manual case review) if the contextual phrase, associated with the drug mention, indicates actual involvement in the death.
- For example, we interpret the drug search term “HEROIN” as mentioned without involvement in the drug poisoning death when it was found in “HISTORY OF HEROIN ABUSE”. In contrast, “HEROIN” is interpreted as mentioned with involvement in “ACUTE HEROIN INTOXICATION”
- Phrases are set up in 6 groups (the variable phrase_list takes values from 1 to 6).

  The phrase “*” is included in each of these groups as it represents the qualified mention by itself. That guarantees that the qualified term will be reported as a mention even if no other contextual phrase is found.

  In phrase_list equal to 1 or 2, there are phrases that represent a substance with a measurement of quantity, for example “POSITIVE FOR * NG ML”, “* MCG ML” or “NG ML *”. Note that digits were removed in pass one and kept in pass two.

  In phrase_list equal to 3, there are phrases where the qualified term is in the middle of the phrase, for example “PROBABLE ACUTE * INTOXICATION”.

In phrase_list equal to 4, there are phrases where the qualified term is at the end of the phrase, for example "LETHAL INGESTION OF *" (also, there may be another qualified term in the middle of the phrase). In the program this scenario is considered 'phrase open to the right'.

In phrase_list equal to 5, there are phrases where the qualified term is at the beginning of the phrase, for example "* POISOINING". In the program this scenario is considered 'phrase open to the left'.

In phrase_list equal to 6, there are phrases where the qualified terms are at the beginning and at the of the phrase, for example "* TOXICITY INCLUDING *".

- There are three exclusion settings / parameters for each phrase that can be set in PHRASES dataset, one for each of the three cleaned text field.
  - If the phrase appears in a particular text field and the setting is set to exclude ('x'), then it will be considered as a non-involvement mention in the reporting. This allows flexibility to exclude some mentions from reporting as DMI, due to the phrase they are associated with does not really indicate that the drug is directly involved in the death.
- Phrases may be considered indicating non-involvement, depending on the place they are found in the DC.

For example, the phrase "ABUSE HISTORY OF *" may be considered "non-involvement" if it is found in Part II of the DC (PART II. Enter other significant conditions contributing to death but not resulting in the underlying cause given in PART I), but considered "involvement" if it is found in Part I (PART I. Enter the chain of events--diseases, injuries, or complications--that directly caused the death) or item 43. (43. Describe how injury occurred:).

The current PHRASES exclusion settings are based on manual review of more than 10,000 DC cases for the state of KY, but the user can make changes based on the patterns observed in their jurisdictions.

- This is how the phrase mapping works:

On the first iteration, a sub-list of phrases is taken from PHRASES dataset, where phrases_list=1.

For each (qualified_term,distilled_literal) pair of the DISTILLED_LITERALS dataset, the boundaries of the possible matching phrases are found. If some matching phrases overlap (i.e. one contains the others), only the record with the longest remains. A new dataset PHRASE_MENTIONS is created with that information, each mention sorted by term position in its text_field (left to right).

On the second iteration, a sub-list of phrases is taken from PHRASES dataset, where phrases_list=2.

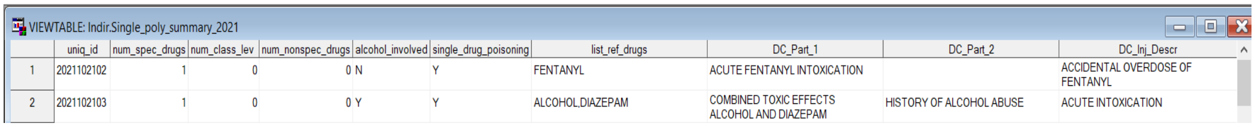
For each (qualified_term,distilled_literal) pair of the created on the previous iteration PHRASE_MENTIONS dataset that still has phrase="*", (reminder, phrase * represents the search term alone), the boundaries of the possible matching phrases are found. If some matching phrases overlap (i.e. one contains the others), only the record with the longest remains. A new version of the PHRASE_MENTIONS is created with that information, each mention sorted by term position in its text_field (left to right).

The third, fourth, fifth, and sixth iteration are similar to the second, they just work on different sublist of phrases. The structure and order of the sublists of phrases is essential because the program tries to find the most logical combinations of phrases and terms, from left to right.

At the end of the sixth iteration the final version of PHRASE_MENTIONS dataset is generated. It still can have a certain number of terms with "*" phrases, it means the program couldn't find any possible phrase on any of the six levels. We can manually review those cases (or subset of them) to determine if there are new common phrases that could be added to the PHRASES dataset to further improve the phrase mapping quality. Adding to PHRASES dataset should be done carefully to the proper level to not disrupt the mapping of already correctly mapped phrases.

- The next step is generation of the output report, where information from the search terms found, mappings and DC data is merged and prepared for further analysis. Each of the two passes on general_search_terms produce its own output.
- After that, the two passes' outputs are combined into a single output for the year and stored in DEATHDATA folder, file "Literal_matching_output_XXXX".
- The next step of the DMI2EpiTool is Single- vs Polydrug analysis.
- Here is how it works:
- Phrase exclusion indicator is derived based on the place in DC where the search term was found and the phrase exclusion settings in PHRASES dataset.
- If the phrase is indicated for exclusion (i.e. the mention of the search term does not indicate true involvement), the mention is removed. At the end of this step only true DMIs remain for analysis.
- Further, for each uniq_id (person) duplicate referent drugs are removed. At the end of this step, we are ready to apply the logic for Single- vs Poly-drug poisoning, which in short can be describes as follows:

1. *Single drug poisoning deaths* are
- Deaths with exactly one specific referent drug as DMI (e.g., "FENTANYL")
or
- Deaths with exactly one referent drug at class level as DMI (e.g., "opioid" overdose)
or
- Deaths with exactly one non-specific referent drug (e.g., "drug" overdose)

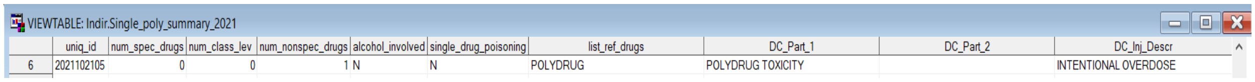

VIEWTABLE: Indir.Single_poly_summary_2021

	uniq_id	num_spec_drugs	num_class_lev	num_nonspec_drugs	alcohol_involved	single_drug_poisoning	list_ref_drugs	DC_Part_1	DC_Part_2	DC_Inj_Descr
1	2021102102	1	0	0	N	Y	FENTANYL	ACUTE FENTANYL INTOXICATION		ACCIDENTAL OVERDOSE OF FENTANYL
2	2021102103	1	0	0	Y	Y	ALCOHOL,DIAZEPAM	COMBINED TOXIC EFFECTS ALCOHOL AND DIAZEPAM	HISTORY OF ALCOHOL ABUSE	ACUTE INTOXICATION

However, there are two additional rules:

1.1 Referent drug 'ALCOHOL' doesn't count as a drug.

1.2 Referent drug 'POLYDRUG' presence always indicates polydrug poisoning (i.e. not a single-drug poisoning), regardless of other counts.


VIEWTABLE: Indir.Single_poly_summary_2021

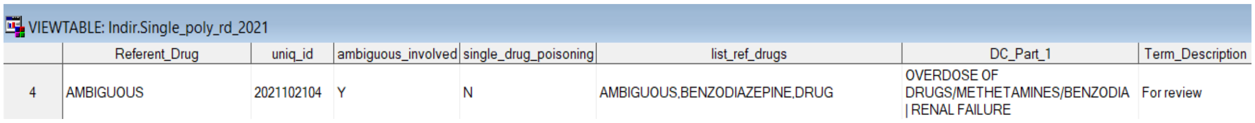
	uniq_id	num_spec_drugs	num_class_lev	num_nonspec_drugs	alcohol_involved	single_drug_poisoning	list_ref_drugs	DC_Part_1	DC_Part_2	DC_Inj_Descr
6	2021102105	0	0	1	N	N	POLYDRUG	POLYDRUG TOXICITY		INTENTIONAL OVERDOSE

2. *Polydrug poisoning deat*hs are all deaths that are not single drug poisoning deaths (i.e. more than one referent drug is listed).

The referent drug specificity properties are derived from gst_level3 dataset, variable 'term_description'. The least specific referent drugs (non-specific), which have term_description value 'NON_SPECIFIC', followed by class-level drugs for which term_description='CLASS'. If a referent drug is not defined as non-specific or class level, it is considered specific and term_description is blank.

Additional indicator for alcohol involvement in the death is derived for each person (uniq_id) based on the presence of referent drug 'ALCOHOL' in the analytical dataset.

Additional indicator for ambiguous drug involvement in the death is derived for each person (uniq_id) based on the presence of referent drug 'AMBIGUOUS' in the analytical dataset. When the ambiguous flag is "Y", the user is prompted to do a manual review of the record. In the example below, the search term" METHETAMINES" is considered AMBIGUOUS.



VIEWTABLE: Indir.Single_poly_rd_2021

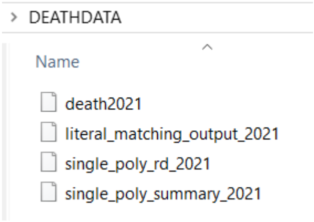
	Referent_Drug	uniq_id	ambiguous_involved	single_drug_poisoning	list_ref_drugs	DC_Part_1	Term_Description
4	AMBIGUOUS	2021102104	Y	N	AMBIGUOUS,BENZODIAZEPINE,DRUG	OVERDOSE OF DRUGS/METHETAMINES/BENZODIA | RENAL FAILURE	For review

Additional indicator for polydrug referent drug involvement in the death is derived for each person (uniq_id) based on the presence of referent drug 'POLYDRUG' in the analytical dataset.

If the analysis shows exactly one referent drug was involved (considering the special rules for 'ALCOHOL' and 'POLYDRUG' described above), indicator Single_drug_poisoning is set to 'Y', which means Single-substance drug poisoning death, while indicator Poly_drug_poisoning is set to 'N'.

Otherwise, indicator Single_drug_poisoning is set to 'N', while indicator Poly_drug_poisoning is set to 'Y', which means Poly-substance drug poisoning death.

After applying the logic for all DCs, two analytic datasets are created and stored in DEATHDATA folder.



death2021
literal_matching_output_2021
single_poly_rd_2021
single_poly_summary_2021

References

CDC. (2019). *National Vital Statistics Mortality Data. Drug Mentions with Involvement Programs. Available from* https://github.com/CDCgov/National-Vital-Statistics-Mortality-Data.

Hedegaard, H., Bastian, B. A., Trinidad, J. P., Spencer, M., & Warner, M. (2018). Drugs Most Frequently Involved in Drug Overdose Deaths: United States, 2011-2016. *Natl Vital Stat Rep*, *67*(9), 1-14. https://www.cdc.gov/nchs/data/nvsr/nvsr67/nvsr67_09-508.pdf

NCHS. (2003). *U.S. Standard Certificate of Death. 2003 revision*. http://www.cdc.gov/nchs/data/dvs/death11-03final-acc.pdf

Trinidad, J. P., Warner, M., Bastian, B. A., Minino, A. M., & Hedegaard, H. (2016). Using Literal Text From the Death Certificate to Enhance Mortality Statistics: Characterizing Drug Involvement in Deaths. *Natl Vital Stat Rep*, *65*(9), 1-15.

Warner, M., Trinidad, J. P., Bastian, B. A., Minino, A. M., & Hedegaard, H. (2016). Drugs Most Frequently Involved in Drug Overdose Deaths: United States, 2010-2014. *Natl Vital Stat Rep*, *65*(10), 1-15.

WHO. (2019). *International classification of diseases, tenth revision: version 2019. World Health Organization.* World Health Organization. https://icd.who.int/browse10/2019/en