

## Drug-Involved Mortality Restricted Data Available at the National Center for Health Statistics Research Data Center

### Introduction

The Drug-Involved Mortality data enhance the National Vital Statistics System, Mortality (NVSS-M) data with information on substances, including prescription and illicit drugs, mentioned on death certificates as having been involved in deaths of U.S. residents of the 50 U.S. States and the District of Columbia. This document describes

- how to access the Drug-Involved Mortality data,
- how the data were extracted from death certificate literal text, and
- the format and structure of the data.

This document is accompanied by an Excel workbook `Search_Terms_and_Principal_Variants.xlsx`.

Throughout this document, uppercase words with quotation marks (e.g., "DRUG") are examples of text that may be found on death certificates, and italicized words (e.g., *substance\_involved*) generally refer to variables in the Drug-Involved Mortality data.

Appendix A of this document lists and describes the variables in the Drug-Involved Mortality data. Appendix B describes the methods used to develop the list of search terms for substances to be identified in death certificate literal text.

We acknowledge the following scientists at the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention, and from the Food and Drug Administration (FDA), for providing input on the content and format of this data documentation:

- James Trinidad, FDA
- Arialdi Minino, NCHS
- Richard (Scott) Swain, FDA
- Yulan Ding, FDA
- Brigham Bastian, NCHS
- Sara Karami, FDA

### Access

The Drug-Involved Mortality data are accessible to the public through the National Center for Health Statistics (NCHS) Research Data Center (RDC). Researchers must submit a proposal for review by NCHS staff. More information is available at <http://www.cdc.gov/rdc>.

### Extraction of information on substance-involved mortality from death certificate literal text

The NVSS-M, in addition to demographic and geographic characteristics, includes a list of codes from the Tenth Revision of the International Classification of Diseases (ICD-10) describing the cause of death for all deaths in the U.S. The Drug-Involved Mortality data supplement the NVSS-M data with information on substances mentioned on death certificates and information on how those substances may have been involved with the death. Methods for identification of drug-involved mortality were expanded to other non-drug substances (e.g., ethylene glycol and helium), but the term 'drug-involved mortality' was kept for sake of consistency. These methods have been previously described in the following publications:

**Updated September 14, 2018**

- Trinidad JP, Warner M, Bastian BA, et al. Using literal text from the death certificate to enhance mortality statistics: Characterizing drug involvement in deaths. National vital statistics reports; vol 65 no 9. Hyattsville, MD: National Center for Health Statistics. 2016.
- Warner M, Trinidad JP, Bastian BA, et al. Drugs most frequently involved in drug overdose deaths: United States, 2010–2014. National vital statistics reports; vol 65 no 10. Hyattsville, MD: National Center for Health Statistics. 2016.

Briefly, the NVSS-M data were linked to electronic files containing literal text information from death certificates. The literal text, located in three fields of the death certificate (i.e., the cause of death from Part I, significant conditions contributing to death from Part II, and a description of how the injury occurred from Box 43), was converted to uppercase, and symbols and numbers were removed. Then, substances mentioned in death certificate literal text were identified using search terms. Lists of search terms were developed using existing drug classification systems as well as from manual review of the literal text. Appendix B describes the methods used to develop the list of search terms. Search terms vary in specificity, with the following classifications: specific substances (e.g., OXYCODONE or FENTANYL), classes of drugs or substances (e.g., OPIOID), non-specific substances (i.e., exposure not otherwise specified) (e.g., DRUG, CHEMICAL, or POLYPHARMACY), ambiguous substances (i.e., possibly referring to two or three different substances), or unknown (i.e., possibly referring to 4 or more different substances); see variable *substance\_specificity* in Appendix A. Each search term was mapped to one or more principal variants. A principal variant is an umbrella term that enables aggregate counts for all search terms that referred to the same substance. Literal text surrounding the identified search terms was searched to ascertain the context (e.g., if and how substances were involved in death) of the substance’s mention.

The case definition of substance-involved mortality (also known as drug-involved mortality) assumes that substances mentioned in the death certificate literal text were involved in the death unless contextual information suggested otherwise (e.g., “METHICILLIN RESISTANT STAPHYLOCOCCUS AUREUS INFECTION” and “DECEDENT DID NOT TAKE DRUG”). Identifying cases of substance-involved mortality is limited by the availability of information provided in death certificate literal text, and the quality and availability of information varies by decedent, jurisdiction, and over time.

Involvement encompasses everything a substance can do or be used for. This includes overdose (poisoning); adverse reactions; being present on board; therapeutic use; being used, misused, or abused; exposure not otherwise specified; complicating health conditions; interacting with other substances; and causing drug dependence or being used by drug dependent persons. Involvement may occur at or prior to time of death (e.g., “HISTORY OF HEROIN ABUSE”). The underlying assumption is that the substances mentioned on the death certificate contributed to death. However, this assumption may not be true in some cases. For example, a substance mentioned as being present on board may have not contributed to death. In addition, a substance could have been used to treat the health conditions that led to death, but also did not contribute to death. Lastly, the death certificate literal text may not contain sufficient information to ascertain exactly how a substance was involved in death. For example, the literal text may simply mention “HEROIN”, with no contextual information on whether heroin resulted in overdose or whether the decedent had a history of heroin abuse. These cases are also assumed to be substance-involved deaths.

As described in the publication by Trinidad et al.,<sup>1</sup> the methods that were developed to extract information on drug-involved mortality have high accuracy in identifying the drugs mentioned and involved in death as well as the corresponding deaths.

---

<sup>1</sup> Trinidad JP, Warner M, Bastian BA, et al. Using literal text from the death certificate to enhance mortality statistics: Characterizing drug involvement in deaths. National vital statistics reports; vol 65 no 9. Hyattsville, MD: National Center for Health Statistics. 2016.

## Format and structure of the data

For every year of data, there are three linkable SAS datasets:

1. A demographics file with a single record for every death of a U.S. resident of the 50 U.S. States and the District of Columbia, and includes demographic, geographic, and cause-of-death variables from the NVSS-M mortality files
2. A mention-level file with each record pertaining to a substance mention that was identified in death certificate literal text
3. A decedent-level file that contains a single record for each decedent with at least one substance mention in death certificate literal text, and that quantifies the number of mentions of involvement of each substance in death

The information on substance-involved deaths in the decedent-level file was derived from data in the mention-level file by counting the number of mentions of substance involvement in death for each decedent and substance. When it is equal to 1, the *substance\_involved* variable in the mention-level file indicates that the substance mentioned in the literal text was assumed to be involved in death; see the section above entitled “Extraction of information on substance-involved mortality from death certificate literal text” regarding the case definition of substance-involved mortality.

The demographic, decedent-level, and mention-level files, are linkable by use of the *decedent\_ID* variable, which is unique within each year.

There are different versions of the decedent-level and mention-level files. Each version is organized by the date when the literal data were processed to identify substance-involved mortality. Search terms and the words and phrases used to ascertain the context in which the substances were mentioned are enhanced with each new version. For example, a new version may have more search terms for specific substances. There are slight differences in the available variables across versions. Appendix A includes descriptions of the variables in the most recent data version, and descriptions of variables in older versions can be made available upon request.

The content of the demographics file does not change across data versions. However, restricted variables in the demographics file must still be requested from and approved by NCHS staff.

The most recent version of the decedent-level and mention-level files will be provided unless older versions are specifically requested. Older versions can be used to replicate previous analyses and to extend those analyses using different data years. Because the most recent version is the most enhanced, use of older versions for *de novo* analyses is discouraged. Conversely, use of the most recent version is encouraged for all analyses, including analyses intended to replicate or extend prior analyses.

However, the results of analyses conducted with different versions may differ. The investigator should consult the Excel workbook, “Search\_Terms\_and\_Principal\_Variants.xlsx,” which details the changes in the queried search terms between the different versions.

The Excel workbook, “Search\_Terms\_and\_Principal\_Variants.xlsx” also lists all search terms used to identify substance mentions, as well as the principal variant mapped to each search term; this workbook is organized by data version. The list of search terms is very comprehensive regardless of data version, and some substances were never mentioned in the death certificate literal text.

To protect from the identification of trade name products, the decedent-level and mention-level files organize substance information by principal variant and not by search term. For example, the mention-level file lists the principal variant METHAMPHETAMINE for every mention of “METHAMPHETAMINE”, “CRYSTAL METH”, and other methamphetamine-related search term found in the literal text. Likewise, the data on substance-involved mortality in the decedent-level file is organized by variable, with each variable representing a specific principal variant. Continuing the previous example, the decedent-level file lists the number of mentions of

methamphetamine-involved mortality in the *METHAMPHETAMINE* variable. These principal variant variables are left blank (“.”) if there were no qualifying mentions of substance-involved mortality in the mention-level file.

With regards to principal variants, newer versions of the decedent-level and mention-level files have been edited to ensure the following:

- Principal variants for specific substances and classes of drugs and substances conform to standardized names.
- Principal variants retain as much substance-related specificity as is possible.

Whenever possible, principal variants for specific substances are the unique preferred substance names listed in the joint Food and Drug Administration (FDA) / United States Pharmacopeia (USP) Substance Registration System (SRS). Each preferred substance name is linked to a unique ingredient identifier (UNII) in the SRS, if available. The preferred name may differ from the term(s) commonly used in everyday practice or even the terms more often mentioned in literal text. For example, although the search term HEROIN is mentioned more often in literal text than the search term DIAMORPHINE, search term HEROIN is mapped to the principal variant DIAMORPHINE. Consult the SRS for preferred names of potential substances of interest. More information about the SRS is available at <https://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII>, and at <https://fdasis.nlm.nih.gov/srs/>.

In many instances, principal variants for classes of drugs and substances are classes listed in the American Hospital Formulary System (AHFS).

Some principal variants are alternatives to or shortened versions of the SRS preferred substance name or AHFS classes. Shortened or alternative principal variants are used because the substance-related SAS variable names in the decedent-level file cannot be longer than 32 characters. For example, the SRS preferred name 2-ETHYLIDENE-1,5-DIMETHYL-3,3-DIPHENYLPYRROLIDINE is 49 characters and is abbreviated as the principal variant EDDP in the mention-level and decedent-level files. Furthermore, some principal variants are simplified versions of the SRS preferred name or AHFS class. For example, the SRS preferred name TOBACCO LEAF is simplified as the principal variant TOBACCO in the mention-level and decedent-level files. The column *Preferred name or class* in the Excel workbook *Search\_Terms\_and\_Principal\_Variants.xlsx* lists the SRS preferred substance name in instances when the principal variants for specific substances are not the SRS preferred name. Researchers who use the SRS or AHFS to identify potential substances of interest should also consult the Excel workbook to ensure comprehensive capture of all mentions of interest.

Except for listing the specific search terms that were identified in literal text, the SAS datasets provide as much substance-specific information as is possible. However, this may result in having to define substances as a composite of various principal variants. For example, an analyst may choose to include the following principal variants to search for deaths involving ethanol: BEER, ETHANOL, TEQUILA, VODKA, and WHISKEY, as well as other ethanol-related principal variants (such as ALCOHOL, which is considered a substance class). Furthermore, starting from data versions dated July 7, 2017, and onward, search terms that are potential metabolites are no longer mapped to potential parent drugs. For example, the search term BENZOYLECGONINE – a potential metabolite of cocaine – is now mapped to the principal variant BENZOYLECGONINE, and not the principal variant COCAINE. This provides additional substance specificity and circumvents two issues: 1) metabolites may arise from different biochemical pathways, and 2) the substance may not be a metabolite, but rather the actual substance involved in death (i.e., there was no parent drug). In addition, most metabolites have a preferred substance name in the SRS. An analyst may choose to define substances as a composite of principal variants to include potential metabolites.

Some search terms could not be mapped to a unique principal variant, but otherwise are for substances mentioned in death certificates. When a search term may refer to two or three isomeric forms (e.g., a mention of

DIFLUOROETHANE may refer to either 1,1-DIFLUOROETHANE or 1,2-DIFLUOROETHANE), the search term is mapped to a principal variant consisting of a common name and a plus sign (e.g., DIFLUOROETHANE+). When a search term may refer to two or three different substances which are not isomers, the search term is mapped to the principal variant AMBIGUOUS. The variables *ambiguous\_variant\_1*, *ambiguous\_variant\_2*, and *ambiguous\_variant\_3* in the mention-level files provide potential principal variants for the search term. The Excel workbook, "Search\_Terms\_and\_Principal\_Variants.xlsx," also lists the potential principal variants in the ambiguous variant variables, as well as the UNII for the potential principal variants (if available) in the columns *ambiguous\_variant\_UNII\_1*, *ambiguous\_variant\_UNII\_2*, and *ambiguous\_variant\_UNII\_3*. An analyst may choose to define substances as a composite of principal variants to include ambiguous mentions, which can only be identified using the mention-level file. Note: search terms that definitely refer to two or more substances are mapped to each appropriate principal variant and are not considered ambiguous (e.g., the search term LORTAB is mapped to both the principal variant ACETAMINOPHEN and the principal variant HYDROCODONE).

In some instances, the search terms are mapped to the principal variant UNKNOWN, indicating that the search term may refer to four or more different substances.

**Because principal variants can contain symbols and spaces, every SAS program that aims to analyze the decedent-level file should include the following global SAS option: OPTIONS VALIDVARNAME = ANY. This option allows for the use of non-conforming SAS variable names. In addition, refer to each principal variant using the name literal convention (i.e., refer to substance-specific variables using "*substance*"n (quotes required); for example, "-gamma.hydroxybutyric acid"n).**

The mention-level file contains information that can be used to help describe the context of the substance's mention. The most granular information on the context of the substance's mention is recorded in the descriptor variables (*descriptor\_before* and *descriptor\_after*) and the *phrase* variable.

The descriptor variables provide information about substance characteristics or characteristics of exposure to the substance according to words immediately adjacent to the substance mention. For example, the descriptors "SYNTHETIC" or "ILLEGAL" may appear before mentions of "FENTANYL", potentially affecting the analyst's decision whether to include those mentions of "FENTANYL". Note: the list of search terms was constructed to exclude descriptors to ensure so that the list could be used to efficiently identify mentions of substances; descriptors were identified after identifying mentions of substances. Care was taken to ensure comprehensive identification of descriptors in the death certificate literal text. However, it is possible that some descriptors were not identified. Furthermore, the literal text may not have sufficient descriptive information for more granular analyses. For example, in the absence of any descriptors, a mention of "FENTANYL" in death certificate literal text may refer to prescription fentanyl or to illicitly manufactured fentanyl. The analyst should determine whether and how to use descriptor information in identifying deaths of interest.

The *phrase* variable in the mention-level file lists contextual phrases, which help to describe whether and how a substance was involved in mortality. Every contextual phrase is deidentified. Furthermore, contextual phrases are formatted as an asterisk (\*) surrounded by words, where the asterisk represents the one or more substances identified in the death certificate literal text. For example, if a literal text field states "OPIOID (HEROIN) OVERDOSE", then the contextual phrase '\* OVERDOSE' would be listed in the *phrase* variable for the mention of "OPIOID" and the mention of "HEROIN". A literal text field that only states "HEROIN" with no other context would be associated with the contextual phrase '\*.

In the mention-level file, many of the contextual phrases have been categorized to enable efficient querying of contexts associated with each substance mention. For example, the *overdose* variable is flagged ("1") for each substance mention associated with a contextual phrase related to overdose. A substance mention may be associated with more than one category of contextual phrase. When equal to 1, the *substance\_involved* variable

indicates that the substance mentioned in death certificate literal text was assumed to have been involved in death. Moreover, this variable has high accuracy in identifying substance-involved mortality compared with manual review of the literal text (see Appendix B). As with the descriptor variables, care was taken to ensure comprehensive capture of contextual phrases, but some contextual information may have not been identified or may be missing in literal text. Furthermore, the categories of contextual phrases may not be as comprehensive as necessary for analyses. In addition, the death certificate literal text may not contain sufficient information to ascertain the true involvement of a substance. For example, a mention of "NALOXONE" associated with the contextual phrase '\* OVERDOSE' may indicate an overdose of naloxone or treatment of overdose with naloxone. The analyst should determine whether and how to use contextual phrase information in identifying deaths of interest.

## APPENDIX A: Data Dictionary

### Drug-involved mortality data, decedent-level file

The decedent-level file contains a single record for each decedent with at least one substance mention in death certificate literal text. Substances were identified in literal text using search terms and were classified by level of specificity of information (e.g., specific substances, classes of drugs or substances, or non-specific substances (i.e., exposure not otherwise specified)). The data on substance-involved mortality in the decedent-level file is organized by variable, with each variable representing a specific principal variant. A principal variant is an umbrella term that enables aggregate counts for all search terms that referred to the same substance. The substances mentioned in the death certificate literal text were assumed to be involved in the death unless contextual information in the literal text suggested otherwise.

Variable Name	Variable description	Data Values
decedent_ID	<b>Numeric.</b> Randomly assigned ID, enabling linkage between the decedent-level file, the mention-level file, and the demographics file	>0
count_PV_involved	<b>Numeric.</b> Quantifies the number of unique principal variants involved in death, where principal variants include specific substances (e.g., OXYCODONE or FENTANYL), classes of drugs or substances (e.g., OPIOID), non-specific substances (i.e., exposure not otherwise specified) (e.g., DRUG, CHEMICAL, or POLYPHARMACY), ambiguous substances (i.e., possibly referring to two or three different substances), or unknown (i.e., possibly referring to 4 or more different substances). If the <i>count_PV_involved</i> variable is equal to 0, then all substances being mentioned in the death certificate literal text were categorized as not being involved in death.	≥0
count_specific_subs_involved	<b>Numeric.</b> Quantifies the number of unique specific substances (e.g., HEROIN) involved in death.	≥0
count_classes_involved	<b>Numeric.</b> Quantifies the number of unique classes (e.g., OPIOID) involved in death.	≥0
count_non_specific_sub_involved	<b>Numeric.</b> Quantifies the number of non-specific substances (e.g., DRUG, CHEMICAL, or POLYPHARMACY) involved in death.	≥0
'principal variant'	<b>Numeric.</b> Each principal variant variable quantifies the number of mentions of involvement of the referent substance per decedent. Substances that were not categorized as having been involved in death have a null value (".").  Some of these variables do not conform to the ordinary SAS name standards. Therefore, use the following SAS option: OPTION VALIDVARNAME = ANY.	>0 or .

### Drug-involved mortality data, mention-level file

In the mention-level file, each record pertains to a substance mention that was identified in the following death certificate literal text fields: the cause of death from Part I, significant conditions contributing to death from Part II, and a description of how the injury occurred from Box 43. Substances were identified in literal text using search terms and were classified as specific substances, classes of drugs or substances, or non-specific substances

**Updated September 14, 2018**

(i.e., exposure not otherwise specified). Each search term is mapped to one or more principal variants. A principal variant is an umbrella term that enables aggregate counts for all search terms that referred to the same substance. The mention-level file also contains information that can be used to help describe the context of the substance’s mention. The most granular information on the context of the substance’s mention is recorded in the descriptor variables (*descriptor\_before* and *descriptor\_after*) and the *phrase* variable. The descriptor variables provide information about substance characteristics or characteristics of substance exposure according to words immediately adjacent to the substance mention. The *phrase* variable lists contextual phrases, which help to describe whether and how a substance was involved in mortality. Many of the contextual phrases have been categorized to enable efficient querying of contexts associated with each substance mention.

Note: Flag variables, such as the *substance\_involved* variable, are equal to 1 when the variable applies, and are equal to null (“.”) when the variable does not apply.

Variable Name	Variable description	Data Values
decedent_ID	<b>Numeric.</b> Randomly assigned ID, enabling linkage between the decedent-level file, the mention-level file, and the demographics file	>0
text_field	<b>Character.</b> Part of death certificate the search term was mentioned on the death certificate (i.e., the cause of death from Part I, significant conditions contributing to death from Part II, and a description of how the injury occurred from Box 43)	PART 1 PART 2 BOX 43
term_position_beg	<b>Number.</b> Start position of term in text field	>0
term_position_end	<b>Number.</b> End position of term in text field	>0
principal_variant	<b>Character.</b> Principal variant in which all related search terms are mapped. For example, the search terms ACETAMINOPHEN and APAP are both mapped to the principal variant ACETAMINOPHEN.	Substance name, such as ACETAMINOPHEN
descriptor_before	<b>Character.</b> Descriptive word(s) identified immediately before the mention of the search term, such as the descriptors “ILLCIT” or “PRESCRIPTION” prior to the mention of “DRUG”.	Descriptive words, such as the word PRESCRIPTION
descriptor_after	<b>Character.</b> Descriptive word(s) identified immediately after the mention of the search term, such as the descriptor “TABLET” after the mention of “DRUG”.	Descriptive words, such as the word “TABLET”
term_not_principal	<b>Numeric.</b> Flag for when the search term identified in the literal text is not the same as the principal variant.	1 or .
substance_specificity	<b>Character.</b> Level of specificity of substance mention. Substance specificity classifications are as follows: specific substances (e.g., OXYCODONE or FENTANYL), classes of drugs or substances (e.g., OPIOID), or non-specific substances (i.e., exposure not otherwise specified) (e.g., DRUG, CHEMICAL, or POLYPHARMACY), ambiguous substances (i.e., possibly referring to two or three different substances), or unknown (i.e., possibly referring to four or more different substances)	AMBIGUOUS CLASS NON-SPECIFIC SUBSTANCE SPECIFIC SUBSTANCE UNKNOWN
number_mapped_principal	<b>Number.</b> Number of principal variants mapped to the identified search term. For example, the value of <i>number_mapped_principal</i> would equal 2 for the search term “PERCO CET”, which is mapped to the principal variants “ACETAMINOPHEN” and “OXYCODONE”.	>0
UNII	<b>Character.</b> Unique Ingredient Identifier (UNII) of principal variant, according to the Substance Registration System (SRS).	Alphanumeric identifier, such as 36209ITL9D, which is the UNII for acetaminophen

original_preferred_name	<b>Character.</b> Provides the Substance Registration System (SRS) preferred name or the American Hospital Formulary System (AHFS) class when the principal variant is not the SRS preferred name or AHFS class, respectively. For example, when the SRS preferred name is greater than 32 characters, the principal variant is shortened to 32 characters or less. Similarly, when the name of the AHFS class is greater than 32 characters, the class name is listed here.	Substance name, such as 2-ETHYLIDENE-1,5-DIMETHYL-3,3-DIPHENYLPYRROLIDINE
ambiguous_variant_1 through ambiguous_variant_3	<b>Character.</b> A potential principal variant that could be mapped to the search term. Applies only if the search term ambiguously refers to two or three potential principal variants. For example, the search term ACETAMIN may refer to ACETAMINOPHEN, ACETAMINOSALOL, or ACETAMINOTADALAFIL. The principal variant for ambiguous search terms is AMBIGUOUS. Note: Some search terms refer to substances upon manual review of literal text, but could not be mapped to a principal variant because they may refer to four or more substances. In these instances, the principal variant is UNKNOWN.	Substance name, such as ACETAMINOPHEN
ambiguous_variant_UNII_1 through ambiguous_variant_UNII_3	<b>Character.</b> The UNII for ambiguous variant #1 through #3. Applies only if the search term ambiguously refers to two or three potential principal variants.	Alphanumeric identifier, such as 36209ITL9D, which is the UNII for acetaminophen
phrase	<b>Character.</b> The contextual phrase in which the substance was mentioned. Substances are symbolized as an asterisk (“*”) in these phrases. For example, in the case of the text “ACUTE OPIOID OVERDOSE”, the search term OPIOID is mapped to contextual phrase ACUTE * OVERDOSE.	Contextual phrases, each containing an asterisk, such as ACUTE * OVERDOSE
phrase_beg	<b>Number.</b> Beginning position of the contextual phrase within the literal text field after symbols, numbers, and double-spaces have been removed from the text field and the search terms and their descriptors have been symbolized with an asterisk (“*”).	>0
phrase_end	<b>Number.</b> Ending position of the contextual phrase within the literal text field after symbols, numbers, and double-spaces have been removed from the text field and the search terms and their descriptors have been symbolized with an asterisk (“*”).	>0
substance_involved	<b>Number.</b> Flag for when the substance mention meets the case definition of substance-involved mortality (a.k.a. drug-involved mortality). This case definition assumes that substances mentioned in the death certificate literal text were involved in the death unless contextual information suggested otherwise, according to the <i>not_substance_involved</i> variable.	1 or .
not_substance_involved	<b>Number.</b> Flag for when the contextual phrase suggests no substance involvement. <i>Not_substance_involved</i> is equal to 1 when any of the following ancillary variables are equal to 1: <i>negation</i> , <i>diabetes</i> , or <i>ABX</i>	1 or .
negation	<b>Number.</b> Flag for when the contextual phrase indicates no substance involvement (e.g., “DECEDENT DID NOT ADMINISTER DRUG”)	1 or .
diabetes	<b>Number.</b> Flag for when the contextual phrase refers to a type of diabetes (e.g., “INSULIN DEPENDENT DIABETES”)	1 or .

ABX	<b>Number.</b> Flag for when the contextual phrase refers to a type of antibiotic resistance or sensitivity (e.g., “METHICILLIN RESISTANCE STAPHYLOCOCCUS AUREUS”). ABX is equal to 1 when any of the following ancillary variables are equal to 1: <i>ABX_staph_aureus</i> , <i>ABX_enterococcus</i> , <i>ABX_pseudomonas</i> , or <i>ABX_ecoli</i> . ABX is also equal to 1 when the contextual phrase refers to medication resistance or sensitivity for organisms not otherwise specified.	1 or .
ABX_staph_aureus	<b>Number.</b> Flag for when the contextual phrase refers to a Staphylococcus aureus resistance or sensitivity to medications	1 or .
ABX_enterococcus	<b>Number.</b> Flag for when the contextual phrase refers to a Enterococcus resistance or sensitivity to medications	1 or .
ABX_pseudomonas	<b>Number.</b> Flag for when the contextual phrase refers to a Pseudomonas resistance or sensitivity to medications	1 or .
ABX_ecoli	<b>Number.</b> Flag for when the contextual phrase refers to a Escherichia coli resistance or sensitivity to medications	1 or .
admin	<b>Number.</b> Flag for when the contextual phrase refers to administration of a substance. Admin is equal to 1 when any of the following ancillary variables are equal to 1: <i>admin_oral</i> , <i>admin_inject</i> , <i>admin_inhale</i> , <i>admin_rectal</i> , or <i>admin_topical</i> . Admin is also equal to 1 when the context phrase refers to administration by routes not otherwise specified	1 or .
admin_oral	<b>Number.</b> Flag for when the contextual phrase refers to oral administration of a substance	1 or .
admin_inject	<b>Number.</b> Flag for when the contextual phrase refers to IV administration of a substance	1 or .
admin_inhale	<b>Number.</b> Flag for when the contextual phrase refers to nasal or inhalational route of administration of a substance	1 or .
admin_rectal	<b>Number.</b> Flag for when the contextual phrase refers to rectal administration of a substance	1 or .
admin_topical	<b>Number.</b> Flag for when the contextual phrase refers to topical administration of a substance	1 or .
exposure_neonatal	<b>Number.</b> Flag for when the contextual phrase refers to <i>in utero</i> exposure to a substance	1 or .
exposure	<b>Number.</b> Flag for when the contextual phrase refers to exposure to a substance. Does not include <i>in utero</i> exposure.	1 or .
use	<b>Number.</b> Flag for when the contextual phrase refers to use of a substance	1 or .
misuse	<b>Number.</b> Flag for when the contextual phrase refers to substance misuse or noncompliance	1 or .
recreation	<b>Number.</b> Flag for when the contextual phrase refers to recreational intent.	1 or .
abuse	<b>Number.</b> Flag for when the contextual phrase refers to abuse of a substance	1 or .
addiction	<b>Number.</b> Flag for when the contextual phrase refers to substance addiction	1 or .
dependence	<b>Number.</b> Flag for when the contextual phrase refers to substance dependence	1 or .
withdrawal	<b>Number.</b> Flag for when the contextual phrase refers to substance withdrawal	1 or .
intoxication	<b>Number.</b> Flag for when the contextual phrase refers to substance intoxication or impairment	1 or .
toxicity	<b>Number.</b> Flag for when the contextual phrase refers to substance toxicity	1 or .

toxic_effect	<b>Number.</b> Flag for when the contextual phrase refers to substance toxic effect or toxic reaction	1 or .
allergy	<b>Number.</b> Flag for when the contextual phrase refers to substance allergy or anaphylaxis	1 or .
stevens_johnson_syn	<b>Number.</b> Flag for when the contextual phrase refers to Stevens Johnson Syndrome	1 or .
gvfd	<b>Number.</b> Flag for when the contextual phrase refers to Graft vs Host Disease	1 or .
reaction	<b>Number.</b> Flag for when the contextual phrase refers to adverse reaction or effect. Does not necessarily include phrases classified under toxicity, toxic_effect, allergy, stevens_johnson_syn, GVHD, sedation, poisoning, overdose, interaction, complication, or coagulopathy	1 or .
sedation	<b>Number.</b> Flag for when the contextual phrase refers to sedation or analgesia from substances	1 or .
poisoning	<b>Number.</b> Flag for when the contextual phrase refers to poisoning from substances	1 or .
overdose	<b>Number.</b> Flag for when the contextual phrase refers to overdose from substances	1 or .
interaction	<b>Number.</b> Flag for when the contextual phrase refers to substance interaction or synergism	1 or .
complication	<b>Number.</b> Flag for when the contextual phrase refers to substance complication or exacerbation of a health condition	1 or .
toxicology	<b>Number.</b> Flag for when the contextual phrase refers to toxicology or screening	1 or .
toxicology_quantified	<b>Number.</b> Flag for when the contextual phrase refers to a quantified toxicology or screening result, e.g., "BLOOD LEVEL 10 MG PER ML". Note: The numbers are not kept in the analytical dataset, but are in the raw dataset	1 or .
therapy	<b>Number.</b> Flag for when the contextual phrase refers to therapy with substance	1 or .
access	<b>Number.</b> Flag for when the contextual phrase refers to substance receipt, dispensed, or obtained.	1 or .
coagulopathy	<b>Number.</b> Flag for when the contextual phrase refers to coagulopathy related to substances	1 or .

### Demographics file: variables that are not for public use

The demographics files contain a single record for every death of a U.S. resident of the 50 U.S. States and the District of Columbia, and includes demographic, geographic, and cause-of-death variables from the NVSS-M mortality files. The variables below can be requested to enhance the demographics file. These additional variables are from the NVSS-M files and are not-for-public-use.

Variable Name	Variable description	Data Values
---------------	----------------------	-------------

stocc	<b>Character.</b> State of occurrence, i.e., the state or territory where death occurred, according to the Federal Information Processing Standards (FIPS)	United States AL ... Alabama AK ... Alaska AZ ... Arizona AR ... Arkansas CA ... California CO ... Colorado CT ... Connecticut DE ... Delaware DC ... District of Columbia FL ... Florida GA ... Georgia HI ... Hawaii ID ... Idaho IL ... Illinois IN ... Indiana IA ... Iowa KS ... Kansas KY ... Kentucky LA ... Louisiana ME ... Maine MD ... Maryland MA ... Massachusetts MI ... Michigan MN ... Minnesota MS ... Mississippi MO ... Missouri MT ... Montana NE ... Nebraska NV ... Nevada	United States continued NH ... New Hampshire NJ ... New Jersey NM ... New Mexico NY ... New York NC ... North Carolina ND ... North Dakota OH ... Ohio OK ... Oklahoma OR ... Oregon PA ... Pennsylvania RI ... Rhode Island SC ... South Carolina SD ... South Dakota TN ... Tennessee TX ... Texas UT ... Utah VT ... Vermont VA ... Virginia WA ... Washington WV ... West Virginia WI ... Wisconsin WY ... Wyoming
exstocc	<b>Character.</b> State of occurrence, i.e., the state or territory where death occurred, according to FIPS, with New York City as a separate (non-FIPS) code	Same abbreviations used for the <i>stocc</i> variable, plus: YC ... New York City	
cntyocc	<b>Character.</b> County of occurrence, stating the county where death occurred, according to FIPS. Counties and county equivalents (independent and coextensive cities) are numbered alphabetically within each State and identify each county. (Note: To uniquely identify a county, both the state and county codes must be used.)	001-nnn ... Code range	
pcntyocc	<b>Character.</b> Population size of county of occurrence, based on the 2000 Census data	0 ... Place of 1,000,000 or more 1 ... Place of 500,000 to 1,000,000 2 ... Place of 250,000 to 500,000 3 ... Place of 100,000 to 250,000 4 ... Place of 50,000 to 100,000 5 ... Place of 25,000 to 50,000 6 ... Place of 10,000 to 25,000 9 ... Place of less than 10,000	
stres	<b>Character.</b> State of residence, which should be restricted to all 50 states, according to FIPS	Same codes used for the <i>stocc</i> variable	

extstres	<b>Character.</b> State of residence, which should be restricted to all 50 states, according to FIPS, with New York City as a separate (non-FIPS) code	Same abbreviations used for the <i>stocc</i> variable, plus: YC ... New York City
cntyres	<b>Character.</b> County of residence (FIPS) . Counties and county equivalents (independent and coextensive cities) are numbered alphabetically within each State and identify each county. (Note: To uniquely identify a county, both the state and county codes must be used.)	001-nnn ... Code range
pcntyres	<b>Character.</b> Population size of county of residence, based on 2000 Census data	0 ... Place of 1,000,000 or more 1 ... Place of 500,000 to 1,000,000 2 ... Place of 250,000 to 500,000 3 ... Place of 100,000 to 250,000 4 ... Place of 50,000 to 100,000 5 ... Place of 25,000 to 50,000 6 ... Place of 10,000 to 25,000 9 ... Place of less than 10,000
day	<b>Character.</b> Day of occurrence (death); Note: Month and year of occurrence variables are public use	01 - 31 ... day of month 99 ... missing
monbirth	<b>Character.</b> Month of birth	01 ... January 02 ... February 03 ... March 04 ... April 05 ... May 06 ... June 07 ... July 08 ... August 09 ... September 10 ... October 11 ... November 12 ... December 99 ... missing
daybirth	<b>Character.</b> Day of birth	01 - 31 ... day of month 99 ... missing
yrbirth	<b>Character.</b> Year of birth	1868-nnnn ... year 9999 ... missing
tobacco	<b>Character.</b> Whether or not tobacco use contributed to death	N ... No P ... Probably U ... Unknown Y ... Yes Blank ... Not on certificate
preg	<b>Character.</b> Whether or not decedent was pregnant (at occurrence or recently)	1 ... Not pregnant within the past year 2 ... Pregnant at time of death 3 ... Not pregnant, but pregnant within 42 days of death 4 ... Not pregnant, but pregnant 43 days to 1 year before death 7 ... Not on certificate 8 ... Not applicable 9 ... Unknown if pregnant within last year
whcbox	<b>Character.</b> Checkbox for white	N ... No Y ... Yes Blank ... missing

blcbox	<b>Character.</b> Checkbox for black or African American	N ... No Y ... Yes Blank ... missing
aiancbox	<b>Character.</b> Checkbox for American Indian or Alaska Native	N ... No Y ... Yes Blank ... missing
aindcbox	<b>Character.</b> Checkbox for Asian Indian	N ... No Y ... Yes Blank ... missing
chincbox	<b>Character.</b> Checkbox for Chinese	N ... No Y ... Yes Blank ... missing
filcbox	<b>Character.</b> Checkbox for Filipino	N ... No Y ... Yes Blank ... missing
japcbox	<b>Character.</b> Checkbox for Japanese	N ... No Y ... Yes Blank ... missing
korcbox	<b>Character.</b> Checkbox for Korean	N ... No Y ... Yes Blank ... missing
vietcbox	<b>Character.</b> Checkbox for Vietnamese	N ... No Y ... Yes Blank ... missing
othacbox	<b>Character.</b> Checkbox for other Asian	N ... No Y ... Yes Blank ... missing
hawcbox	<b>Character.</b> Checkbox for native Hawaiian	N ... No Y ... Yes Blank ... missing
guamcbox	<b>Character.</b> Checkbox for Guamanian or Chamorro	N ... No Y ... Yes Blank ... missing
samcbox	<b>Character.</b> Checkbox for Samoan	N ... No Y ... Yes Blank ... missing
opicbox	<b>Character.</b> Checkbox for other Pacific Islander	N ... No Y ... Yes Blank ... missing
othcbox	<b>Character.</b> Checkbox for other race	N ... No Y ... Yes Blank ... missing
mexcbox	<b>Character.</b> Checkbox for Mexican	N ... No, not specified Hispanic origin H ... Yes, specified Hispanic origin U ... Unknown
prcbox	<b>Character.</b> Checkbox for Puerto Rican	N ... No, not specified Hispanic origin H ... Yes, specified Hispanic origin U ... Unknown
cubcbox	<b>Character.</b> Checkbox for Cuban	N ... No, not specified Hispanic origin H ... Yes, specified Hispanic origin U ... Unknown
ohspcbox	<b>Character.</b> Checkbox for other Spanish/Hispanic/Latino	N ... No, not specified Hispanic origin H ... Yes, specified Hispanic origin U ... Unknown

## Demographics file: variables for public use

The demographics files contain a single record for every death of a U.S. resident and includes demographic, geographic, and cause-of-death variables from the NVSS-M mortality files. The following variables are standard.

Variable Name	Variable description	Data Values
rectype	<b>Character.</b> Record type, stating whether the state and county of occurrence and residence are the same or different	1 ... Resident; state and county of occurrence and residence are the same 2 ... Nonresident; state and/or county of occurrence and residence are different
resstat	<b>Character.</b> Resident status, stating whether the state and county of occurrence are the same, are different by county only, or are different but are still in the U.S.	1 ... Resident; state and county of occurrence and Residence are the same 2 ... Intrastate nonresident; same state of occurrence and residence, but different county 3 ... Interstate nonresident; U.S. state of occurrence and residence are different
educ	<b>Character.</b> Education level, 1989 revision	00 ... No formal education 01-08 ... Years of elementary school 09 ... 1 year of high school 10 ... 2 years of high school 11 ... 3 years of high school 12 ... 4 years of high school 13 ... 1 year of college 14 ... 2 years of college 15 ... 3 years of college 16 ... 4 years of college 17 ... 5 or more years of college 99 ... Not stated
educ2003	<b>Character.</b> Education level, 2003 revision	1 ... 8th grade or less 2 ... 9 - 12th grade, no diploma 3 ... high school graduate or GED completed 4 ... some college credit, but no degree 5 ... Associate degree 6 ... Bachelor's degree 7 ... Master's degree 8 ... Doctorate or professional degree 9 ... Unknown
educflag	<b>Character.</b> Education flag, either the 1989 or 2003 revision	0 ... 1989 revision of education item on certificate 1 ... 2003 revision of education item on certificate
month	<b>Character.</b> Month of occurrence (death)	01 ... January 02 ... February 03 ... March 04 ... April 05 ... May 06 ... June 07 ... July 08 ... August 09 ... September 10 ... October 11 ... November 12 ... December
year	<b>Numeric.</b> Year of occurrence (death)	nynn ... year
sex	<b>Character.</b> Sex	M ... Male F ... Female

ageID	<b>Character.</b> Age unit (e.g., years). Use with the <i>detage</i> variable to ascertain detailed age.	1 ... Years 2 ... Months 4 ... Days 5 ... Hours 6 ... Minutes 9 ... Age not stated
detage	<b>Numeric.</b> Detailed age value. Use with the <i>ageID</i> variable to ascertain detailed age.	nnn ... age value 999 ... age not stated
agesub	<b>Character.</b> Age substitution flag, stating whether calculated age was substituted for unknown reported age using dates of birth and death	Blank ... Calculated age is not substituted for reported age 1 ... Calculated age is substituted for reported age
place	<b>Character.</b> Place of death and decedent's status (e.g., hospital, clinical or medical center, and dead on arrival)	1 ... Hospital, clinic or medical center - Inpatient 2 ... Hospital, clinic or medical center - Outpatient or admitted to emergency room 3 ... Hospital, clinic or medical center - Dead on arrival 4 ... Decedent's home 5 ... Hospice facility 6 ... Nursing home/long term care 7 ... Other 9 ... Place of death unknown
marital	<b>Character.</b> Marital status	S ... Never married, single M ... Married W ... Widowed D ... Divorced U ... Marital status unknown
injwork	<b>Character.</b> Whether or not an injury occurred at work	Y ... Yes N ... No U ... Unknown
mandth	<b>Character.</b> Manner of death (e.g., accident, suicide)	1 ... Accident 2 ... Suicide 3 ... Homicide 4 ... Pending investigation 5 ... Could not determine 7 ... Natural Blank ... Not specified
autopsy	<b>Character.</b> Whether an autopsy was performed	Y ... Yes N ... No U ... Unknown
actcode	<b>Character.</b> Coded information on the activity decedent was engaged in at occurrence (e.g., engaged in sports activity)	0 ... While engaged in sports activity 1 ... While engaged in leisure activity 2 ... While working for income 3 ... While engaged in other types of work 4 ... While resting, sleeping, eating (vital activities) 8 ... While engaged in other specified activities 9 ... During unspecified activity Blank ... Not applicable
plofinj	<b>Character.</b> Place of injury for causes W00-Y34, except Y06.- and Y07.-; example of places of injury include home, residential institution, and street and highway	0 ... Home 1 ... Residential institution 2 ... School, other institution and public administrative area 3 ... Sports and athletics area 4 ... Street and highway 5 ... Trade and service area 6 ... Industrial and construction area

		7 ... Farm 8 ... Other Specified Places 9 ... Unspecified place Blank ... Causes other than W00-Y34, except Y06.- and Y07.-
cause	<b>Character.</b> Underlying cause of death, according to the 10 <sup>th</sup> Revision of the International Classification of Diseases (ICD-10)	See the International Classification of Diseases, Tenth Revision, Volume 1.
nenaxcon	<b>Character.</b> Number of entity axis conditions	00-20 ... number of entity axis conditions
entax1 through entax20	<b>Character.</b> ICD-10-coded condition, including part and line number that condition was identified on death certificate	Each condition consists of six positions. Entax variables are blank when conditions are not listed.  Position 1: Part/line number on certificate 1 ... Part I, line 1 (a) 2 ... Part I, line 2 (b) 3 ... Part I, line 3 (c) 4 ... Part I, line 4 (d) 5 ... Part I, line 5 (e) 6 ... Part II  Position 2: Sequence of condition within part/line 1-6 ... Code range  Position 3 - 6: Condition code
nrcaxcon	<b>Character.</b> Number of record axis conditions	00-20 ... number of record axis conditions
recax1 through recax20	<b>Character.</b> Conditions sorted according to the rules of ICD-10	Each condition consists of four positions. Recax variables are blank when conditions are not listed.  Position 1 - 4: Condition code
detrace	<b>Character.</b> Decedent race. This item is a) the reported race for States that are reporting single race, or b) the bridged race for States that are reporting multiple race.	01 ... White 02 ... Black 03 ... American Indian (includes Aleuts and Eskimos) 04 ... Chinese 05 ... Japanese 06 ... Hawaiian (includes Part-Hawaiian) 07 ... Filipino 18 ... Asian Indian 28 ... Korean 38 ... Samoan 48 ... Vietnamese 58 ... Guamanian 68 ... Other Asian or Pacific Islander in areas reporting codes 18-58 78 ... Combined other Asian or Pacific Islander, includes codes 18-68 for areas that do not report them separately
bridged	<b>Character.</b> Bridged race flag	Blank ... Race is not bridged 1 ... Race is bridged
raceimp	<b>Character.</b> Race imputation flag	Blank ... Race is not imputed 1 ... Unknown race is imputed 2 ... All other races is imputed

rac3	<b>Character.</b> Race record (white, races other than white or black, and black)	1 ... White 2 ... Races other than White or Black 3 ... Black
rac5	<b>Character.</b> Race record (white, black, American Indian, and Asian or Pacific Islander)	1 ... White 2 ... Black 3 ... American Indian 4 ... Asian or Pacific Islander
hispor	<b>Character.</b> Hispanic origin	100-199 ... Non-Hispanic 200-209 ...Spaniard 210-220 ... Mexican 221-230 ... Central American 231-249 ...South American 250-259 ...Latin American 260-269 ... Puerto Rican 270-274 ... Cuban 275-279 ...Dominican 280-299 ... Other Hispanic 996-999 ... Unknown
hisprc	<b>Character.</b> Hispanic/race recode	1 ... Mexican 2 ... Puerto Rican 3 ... Cuban 4 ... Central or South American 5 ... Other or unknown Hispanic 6 ... Non - Hispanic white 7 ... Non - Hispanic black 8 ... Non - Hispanic other races 9 ... Hispanic origin unknown

## APPENDIX B: Developing a search term list for substances

As described in the section “Extraction of information on substance-involved mortality from death certificate literal text,” a search term list is required to identify mentions of substances in death certificate literal text. This appendix describes the methods used to develop the list of search terms for substances.

A substance is defined as any potentially exogenous non-living single chemical or combination of chemicals. Substances exclude body organs, devices and other objects, dosage forms, food, genes, organisms (i.e., animals, bacteria, parasites, persons, and plants), places, and viruses. Fungi (e.g., POISONOUS MUSHROOMS) are not considered food. Substances also excludes behaviors (e.g., SMOKING), procedures (e.g., SURGERY), and medical conditions. Chemicals that can either be endogenous or exogenous were included for further manual review, e.g., VITAMIN D and IRON.

The search term list was constructed in three phases. In the first two phases, the list was intended to ensure capture of drugs involved in death. Drugs includes substances used for therapeutic purposes or for abuse or misuse. The list was expanded to enable capture of substances as defined above.

In the first phase, the search term list consisted of single-word generic names listed in the Substance Abuse and Mental Health Services Administration’s (SAMHSA) Drug Abuse Warning Network (DAWN) Drug Reference Vocabulary (DRV), published in 2012.<sup>2</sup> DAWN DRV is a drug vocabulary and classification system that lists licit substances and substances that can be misused or abused. The first phase excluded the following categories of substances: major substances of abuse, nutritional products, alternative medicines, medical gases, biologicals, immune globulins, immunostimulants, sterile irrigating solutions, and drugs unknown. Products in these categories had generic names that were difficult to condense into a single word denoting a drug product. The list also excluded combination products, nearly all of which could be identified by their components. The search term list that resulted from the first phase of development did not include names of classes or non-specific substances.

In the second phase, the search term list was expanded by adding terms for specific drugs not identified in the first phase, including illicit drugs; classes; non-specific substances; terms containing more than one word; brand names; and obvious, frequently occurring misspellings. Most of the search terms added during the second phase were identified through nonsystematic manual reviews and queries of the 2003–2014 literal text. Additional search terms were added by consulting the Drugs@FDA website and the FDA Adverse Event Reporting System data. A few search terms were also identified using other approaches, including comparison with ICD-10 codes.

The search term list that resulted from the second phase excluded foods and food additives (e.g., starch), excipients, gases (e.g., helium and carbon monoxide), airborne contaminants (e.g., soot), industrial chemicals (e.g., ethylene glycol), periodic table elements (e.g., lithium and iodine), and substances with unknown industrial or pharmaceutical applications.

The third phase relied on a systematic manual review of death certificate literal text to ensure a sensitive capture of substances involved in death according to death certificate literal text. The third phase is described here in detail to ensure replicability of the methods.

For the first step of phase three, decedent records were selected using record axis and entity axis ICD-10 codes that likely pertained to substance-involved mortality. These codes included ICD-10 codes referring to mental or behavioral disorders due to psychoactive substance use, poisonings, adverse effects due to substance exposure,

---

<sup>2</sup> Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. Drug Abuse Warning Network methodology report, 2010 update. 2010.

and ICD-10 codes whose title or definition explicitly indicated substance involvement (e.g., P04.4 Fetus and newborn affected by maternal use of drugs of addiction) (Table 1).

Table 1: Record axis and entity axis ICD-10 codes that likely pertained to substance-involved mortality

F55, G211, G240, G25[.1,.4,.6], G312, G444, G620, G720, H263, H406, I427, I952, J70[.2-.4], K85[.2,.3], K860, L105, L23[.0-.5,.8], L24[.0,.2-.5,.8], L25[.0-.3,.8], L27[.0,.1,.8,.9], L432, L56[.0,.1], L640, M102, M320, M342, M804, M814, M835, M871, N14[.0-.3], O35[.4,.5], P04[.0-.4,.8,.9], P584, P93, P96[.1,.2], Q86[.0-.2], R502, R78, R825, R83[.2,.3], R84[.2,.3], R85[.2,.3], R86[.2,.3], R87[.2,.3], R89[.2,.3], T36-T60, T64-T65, T88[.0,.1,.6,.7], T96, T97, U82-U85, X40-X49, X60-X69, X85-X90, Y10-Y19, Y35[.2,.5], Y367, Y40-Y59, Y63[.0,.1,.4,.6,.8.9], Y64, Y651, Y880, Y90-Y91
---

The second step prepares the literal text of the selected records by removing symbols, numbers, and double spaces; formatting the literal text in uppercase letters; and removing mentions of search terms already listed from phase two. The second step creates a list of the unique words ever listed in the remaining literal text.

The third step flags words for manual review. The flagged words were those that are or could be substances. Words were not flagged if they were expected to be used to describe substances (e.g., PRESCRIPTION or TABLET), or the involvement of those substances in death (e.g., OVERDOSE). Because the list includes unique words without context, substance names that consist of two or more words were identified by their constituent words. For example, the words ETHYLENE and GLYCOL would be flagged because they could be found together as mentions of ETHYLENE GLYCOL. Several lists aided in flagging words. The prior list of search terms was consulted to identify potential constituent words that may be found in substance mentions (e.g., "GAMMA", as in gamma-hydroxybutric acid). The Substance Registration System (SRS) was consulted to assist in flagging potential brand or trade names. A list of single word preferred terms in the SRS was also consulted for this step.

The fourth step identified substance mentions in the literal text of the selected records, using the words flagged in the previous step to select the literal text for manual review. This step was necessary to identify the full substance names (e.g., "ETHYLENE GLYCOL") and to identify instances when potential terms do not refer to substances (e.g., "PCP" referring to pneumocystis pneumonia rather than phencyclidine). Substance mentions in deaths with no substance-involved mortality were included in this step (e.g., "DRUG" would have been included in cases where literal text stated "DECEDENT DID NOT TAKE DRUG"). For efficiency, up to ten randomly selected mentions of each flagged word were reviewed. New substances mentioned in the literal text but not initially flagged for review were noted, and up to ten randomly selected mentions of the words comprising the substance mentions were reviewed.

Some mentions in the literal text may not always meet the substance definition. For example, a mention of "AMMONIA" often referred to endogenous ammonia and in fewer cases referred to exogenous ammonia. Mentions that did not meet the substance definition more than 75% of the time were excluded as potential substance mentions.

The manual review also noted words and phrases that provided context to the substance mentions (i.e., descriptors and contextual phrases). See the section above entitled "Format and structure of the data" regarding descriptors and contextual phrases. Joining phrases, which are words that act as conjunctions, were also identified during the manual review.

The fifth step categorized the newly identified search terms by mapping the appropriate principal variant and unique ingredient identifier (UNII), and by categorizing search terms by substance specificity (e.g., specific substance, class, or non-specific substance). See the section above entitled "Format and structure of the data" regarding principal variants and UNIs.

The final step evaluated the updated lists of search terms, descriptors, joining phrases, and contextual phrases for the identification of mentions of substance-involved mortality and deaths of substance-involved mortality. The aim of the evaluation was to inform any additional editing of these lists.

For the evaluation, SAS programs were run to identify mentions of substances and ascribe context to those substance mentions, as described previously.<sup>3</sup> Then, two subsets of the data were evaluated. The first subset included 500 random death records that had one or more record axis or entity axis ICD-10 codes that likely pertained to substance-involved mortality (Table 1). The second subset included 1,000 random death records with substance mentions, but did not have record axis or entity axis ICD-10 codes that likely pertained to substance-involved mortality. Manual review of the literal text for these two subsets ascertained how well the SAS programs identified mentions of substance-involved mortality and deaths with these mentions. Some search terms that were identified did not appear to refer to substances in the literal text of the second subset of deaths, and some substance mentions were newly identified during the evaluation. A random sample of up to 50 deaths were further evaluated for each newly identified substance mention, and search terms that were determined to be substances in over 75% of deaths were kept in the final list of search terms. Furthermore, additional descriptors, joining phrases, and contextual phrases were also identified during the evaluation.

Among the first subset of deaths having one or more of the selected ICD-10 record axis or entity axis codes, 99.3% of the substance mentions that were identified as mentions of substance involvement were correct (Table 2), and 100% of deaths that were identified as substance-involved deaths were correct (Table 3). Among the second subset of deaths with at least one substance mention but none of the selected record axis or entity axis codes, 84.2% of the substance mentions that were identified as mentions of substance involvement were correct (Table 4), and 83.3% of deaths that were identified as substance-involved deaths were correct (Table 5). A few substances were often mentioned but not involved in death: ANTIBIOTIC, CARBAPENEM, INSULIN, METHICILLIN, OXYGEN, STEROID, VANCOMYCIN, and VITAMIN D. After excluding deaths with these mentions from the second subset, 343 (91.7%) of 374 deaths were correctly identified as substance-involved deaths.

Table 2: Identifying mentions of substance involvement among a random sample of 500 deaths having one or more ICD-10 record axis or entity axis codes: U.S. residents, 2010-2016

Evaluation	Mentions of substance involvement identified from manual review		
	Yes	No	Total
Substance-involved mentions identified by the SAS programs	284	2	286
Substance-involved mentions not identified by the SAS programs	3	...	...

NOTES: See Table 1 for list of record axis and entity axis codes. Positive predictive value calculated as: 284 mentions / 286 mentions = 99.3%

SOURCE: NCHS, National Vital Statistics System, Mortality files linked with death certificate literal text

Table 3: Identifying substance-involved deaths among a random sample of 500 deaths having one or more ICD-10 record axis or entity axis codes: U.S. residents, 2010-2016

Evaluation	Substance-involved deaths identified from manual review		
	Yes	No	Total
Substance-involved deaths identified by the SAS programs	140	0	140
Substance-involved deaths not identified by the SAS programs	1	359	360

<sup>3</sup> Trinidad JP, Warner M, Bastian BA, et al. Using literal text from the death certificate to enhance mortality statistics: Characterizing drug involvement in deaths. National vital statistics reports; vol 65 no 9. Hyattsville, MD: National Center for Health Statistics. 2016

NOTES: See Table 1 for list of record axis and entity axis codes. Positive predictive value calculated as: 140 deaths / 140 deaths = 100.0%

SOURCE: NCHS, National Vital Statistics System, Mortality files linked with death certificate literal text

Table 4: Identifying mentions of substance involvement among a random sample of 1,000 deaths having one or more substance mentions, but no selected ICD-10 record axis or entity axis codes: U.S. residents, 2010-2016

Evaluation	Mentions of substance involvement identified from manual review		
	Yes	No	Total
Substance-involved mentions identified by the SAS programs	441	83	524
Substance-involved mentions not identified by the SAS programs	12	...	...

NOTES: See Table 1 for list of record axis and entity axis codes. Positive predictive value calculated as: 441 mentions / 524 mentions = 84.2%

SOURCE: NCHS, National Vital Statistics System, Mortality files linked with death certificate literal text

Table 5: Identifying substance-involved deaths among a random sample of 1,000 deaths having one or more substance mentions, but no selected ICD-10 record axis or entity axis codes: U.S. residents, 2010-2016

Evaluation	Substance-involved deaths identified from manual review		
	Yes	No	Total
Substance-involved deaths identified by the SAS programs	403	81	484
Substance-involved deaths not identified by the SAS programs	6	510	516

NOTES: See Table 1 for list of record axis and entity axis codes. Positive predictive value calculated as: 403 deaths / 484 deaths = 83.3%

SOURCE: NCHS, National Vital Statistics System, Mortality files linked with death certificate literal text

The high percentage of agreement between the programs and manual review suggest that the lists used by the programs were generally comprehensive, and updates to the lists may yield better agreement.

The development and maintenance of the lists used by the SAS programs to identify mentions of substance involvement requires an understanding of the substances of interest as well as iterative manual reviews of the literal text.

Development and maintenance of the lists also requires some subjectivity. Subjectivity arises in determining whether to include potential search terms with multiple meanings. For example, in some records, "POLYPHARMACY" means procurement of many pharmaceutical drugs, which is a behavior that does not meet the substance definition. However, in other records, "POLYPHARMACY" loosely translates to 'many pharmaceutical drugs', which meets the substance definition. The decision to include a search term that has multiple meanings is done on a case-by-case basis, and some potential search terms require further evaluation (e.g., the word CHEMOTHERAPY, which is currently considered a medical procedure; and the word ALCOHOLIC, which is currently considered an adjective or a medical condition). Note: POLYPHARMACY and ANTICOAGULATION have been included as search terms.

Subjectivity also plays a role in the development and maintenance of lists used to ascribe context to the substance mentions. For example, whether contextual phrases containing the word DEPENDENT should be categorized as either assuming or not assuming substance involvement depended on other contextual factors. For example, in the text "OPIOID DEPENDENT", the word dependent was used to convey an opioid addiction, so the contextual phrase "\* DEPENDENT" was categorized to assume substance involvement. Note: In these contextual phrases, the asterisk represents the one or more substances identified in the death certificate literal

text. In the text “INSULIN DEPENDENT DIABETES”, the word dependent was used to convey a type of diabetes, so the contextual phrase ‘\* DEPENDENT DIABETES’ was categorized to assume no substance involvement. Because of this distinction, a mention of “INSULIN DEPENDENT” in literal text, but no mention of diabetes, would be categorized to assume substance involvement. Analysis of the drug-involved mortality data should carefully consider whether the phrases used to assume substance involvement or no substance involvement make sense.

A careful understanding of the process to ascribe context to the substance mentions is also important in the development and maintenance of the lists. For example, some descriptors help quantify the amount of exposure to a substance (e.g., EXCESSIVE, LITTLE, MANY, and SUBSTANTIAL). However, the word NO was not listed as a descriptor because it was determined to be more useful in contextual phrases to better categorize whether the substances should be assumed or not be assumed to be involved in death.