Drugs in Death Certificates: Automatic Identification and Classification

Soon Jye, Kho
Lister Hill National Center for Biomedical Communications, National Library of Medicine,
Bethesda, MD, 20864, USA

Introduction

The National Vital Statistics System (NVSS) is an inter-governmental effort for forming the standardized mechanism by which the National Center for Health Statistics (NCHS) collects and disseminates the Nation's official vital statistics[1]. It provides a means of vital data collection through contracts between NCHS and vital registration systems which are responsible for the registration of vital events. The collected mortality data are a fundamental source of demographic, geographic, and cause-of-death information[2]. It has been used to characterize the decedent population, determine life expectancy, infant mortality rate, and mortality trends.

Mortality data have been used to determine drug overdose trends as well. Causes of death are classified in accordance with the *International Classification of Disease, Tenth Revision* (ICD-10). ICD-10 codes X40-X44 are assigned to unintentional drug poisoning or overdose deaths, while drug-specific overdose deaths are identified by the contributory causes of death indicated by "T" codes (E.g., T40.1 indicates death due to poisoning by, adverse effect of and underdosing of heroin). Using the ICD-10 codes assigned, researchers have determined drug overdose trends over a time period [1,2,3].

However, merely using ICD-10 codes for analyzing drug overdose trends does not provide enough granularity (i.e., does not support reporting at the level of individual drugs) and does not provide sufficient flexibility for aggregating drug overdose cases into meaningful classes. Our investigation addresses these two issues. We researched methods for automatically identifying drugs in death certificates and investigated drug classification resources for aggregation purposes. This report is divided into two sections: automatic identification and classification.

AUTOMATIC IDENTIFICATION OF DRUGS IN DEATH CERTIFICATES

For every drug overdose case, an ICD-10 code is assigned based on the drug involved in the cause of death. However, ICD-10 codes are not fine-grained enough to record all the information found on death certificates. Though some drugs have a unique ICD-10 code, most of them do not. For example, drug overdose cases caused by heroin and methadone are assigned distinct ICD-10 codes, T40.1 and T40.3, respectively. In contrast, drug overdose cases caused by fentanyl and tramadol are clustered together and assigned the same code, T40.4 (poisoning by synthetic narcotics). This granularity issue also affects all the other opioids (T40.2), as well as barbiturates (T42.3) and benzodiazepines (T42.4).

To mitigate the granularity issues, researchers have utilized the literal text portion of death certificates to identify the number of overdose cases related to a specific drug [4]. The death certificates in which

[1] https://www.cdc.gov/nchs/nvss/index.htm
[2] https://www.cdc.gov/nchs/nvss/deaths.htm

a specific drug is mentioned can be retrieved and trends can be monitored for specific drugs. Yet, identifying drug mentions in death certificates remains challenging as a drug entity could be represented by different terminology variants, including brand names and synonyms. Moreover, drug names are frequently misspelled in death certificates. It is important to identify these variants in order to have complete and accurate retrieval of death certificates.

Trinidad et. al. [5] manually inspected death certificates over a 5-year period (2010-2014) and identified a list of search terms for drugs. These terms were manually identified from the literal text of death certificates and include synonyms, abbreviations, brand names, and misspellings. For a given drug, one of the terms ("principal variant") was designated as the representative term for that drug.

This manually compiled drug list covers a large number of drugs involved in the death of decedents. Manual identification of drug name variants from death certificates is expected to maximize recall in the retrieval of death certificates. On the other hand, because it requires a significant manual effort from domain experts, it does not constitute a sustainable approach. Therefore, our goal here was not so much to outperform the retrieval obtained from the manually compiled drug list. It was rather to assess whether such a drug list can be established automatically. In this preliminary investigation, we focused on acquiring drug name variants for the principal variants from the manual list. We evaluated the performance of the automatically compiled list of variants in comparison with the manually compiled list.

Methods

The manually compiled list of principal variants includes one term for each drug, along with the Unique Ingredient Identifier (UNII) codes to which they were mapped. Starting from the principal variants, we enriched the list with terminology variants for every principal variant using existing drug reference terminologies. The purpose is to identify as many terminology variants as possible using an automated approach. The workflow of identifying terminology variants is described below, followed by the evaluation method.

Mapping principal variants to reference terminologies

RxNorm[3] is a resource that covers prescription drugs in the U.S. and is used in this project. Each drug in RxNorm is represented by a unique identifier, RxCUI. We used the RxNorm API to map each principal variant (drug name or UNII code) to its respective RxCUI.

Identifying additional variants

There are three types of terminology variants: synonyms, brand names and misspellings. RxNorm is used as a source of synonyms and brand names for the principal variants from the manually compiled list. In practice, synonyms and brand names are queried using the RxNorm API for the RxCUIs to which the principal variants mapped.

As drug names are often misspelled in death certificates, it is important to identify misspellings as well. Potential misspellings are identified by adopting the algorithm proposed by Pimpalkhute et. al. [7]. The algorithm provides a consistent and automatic approach to identifying potential misspellings using edit distance and phonetic-based filtering. In this preliminary investigation, we used pre-

[3] https://www.nlm.nih.gov/research/umls/rxnorm/

generated misspellings for common drug names from [7]. These list only provide partial coverage of the drugs from the list of principal variants.

The list of variants obtained from adding to the list of principal variants synonyms and brand names from RxNorm and misspellings for common drug names constitutes our automatically enriched list.

Evaluation

We compared the number of documents retrieved using the manually compiled list and automatically enriched list. The number of documents retrieved is aggregated for every principal variant. For the manually compiled list, the number of documents is the number of unique documents retrieved by all principal variants and their manually collected variants. For the automatically enriched list, it is the number of unique documents retrieved by all principal variants and their synonyms, brand names and misspellings

Results

RxNorm Mapping

Of the total of 2477 principal variants, 1699 (68.59%) could be mapped to RxNorm and represent the prescription drugs in U.S. The remaining 778 principal variants could not be mapped to RxNorm as these are mostly illicit drugs (e.g. cathinone), drug class terms (e.g. antidepressants) and non-medication substance (e.g. petroleum), which do not exist in RxNorm, as RxNorm only covers prescription drugs. Only those principal variants with RxCUI are enriched with terminology variants, as shown in Figure 1.

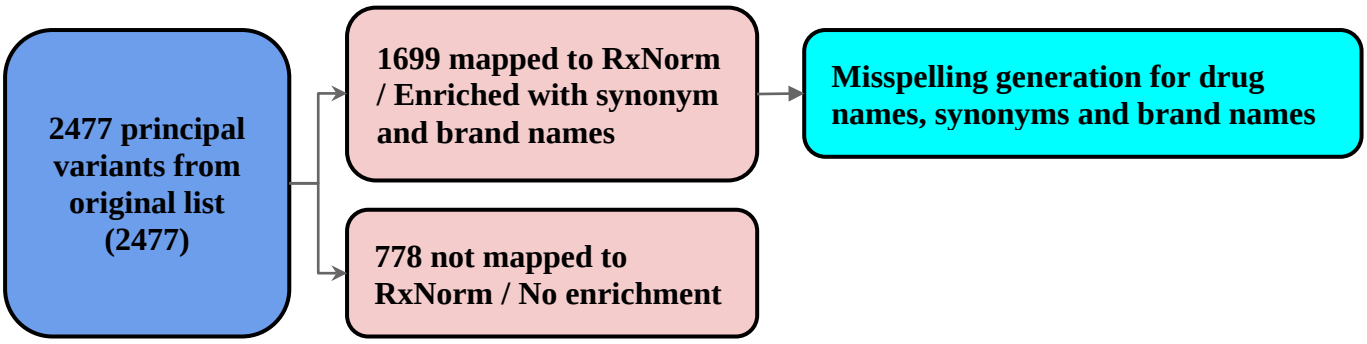

Figure 1: The workflow for automatic enrichment of principal variants

Evaluation

The comparison is performed in two different parts: for principal variants that could be mapped to RxNorm and for those that could not. Figure 2 shows the number of documents retrieved by those 1699 principal variants that could automatically be enriched using RxNorm. Figure 3 shows the number of documents retrieved for the remaining 778 principal variants that could not be automatically enriched and only the name of each principal variant is used for retrieval.

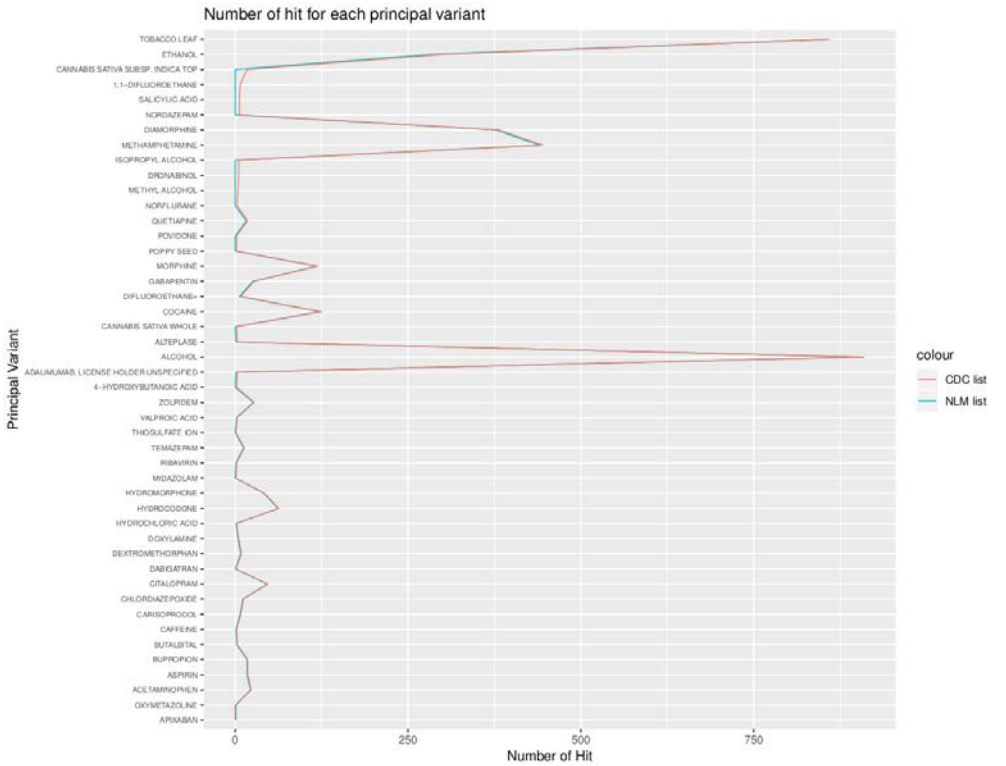

Figure 2: Number of death certificate retrieved for 1699 principal variants that could be automatically enriched using RxNorm.

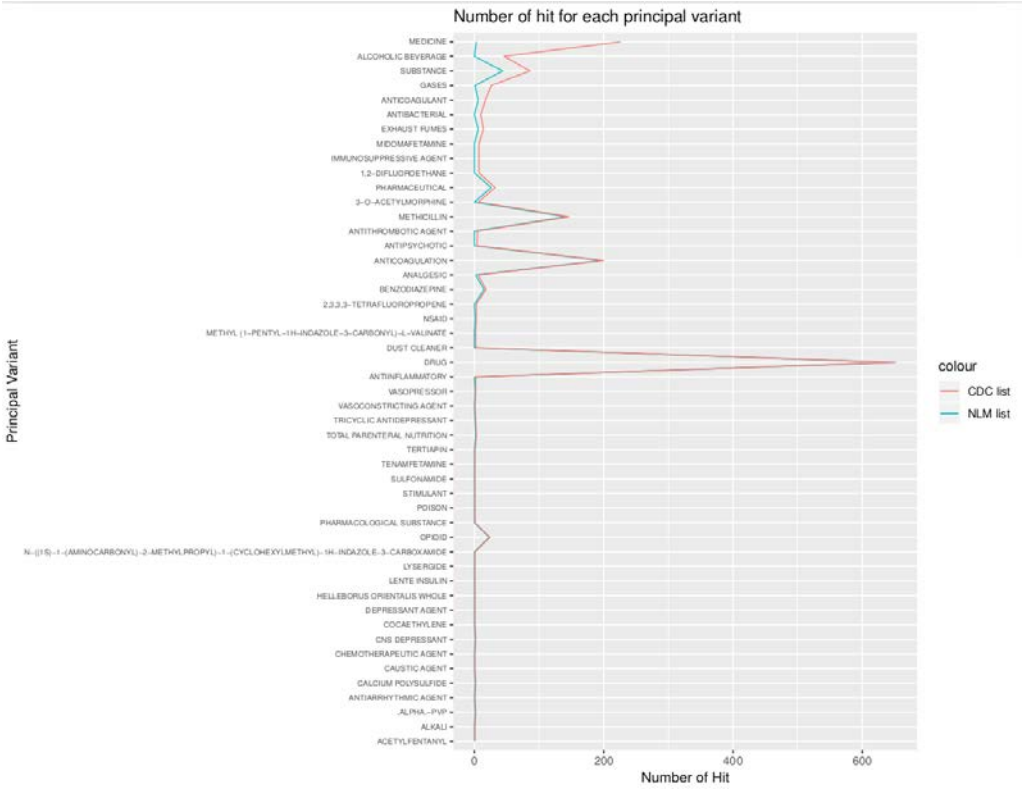

Figure 3: Number of death certificate retrieved for 778 principal variants that could not be automatically enriched.

Table 1 shows the total number of documents retrieved using the manually compiled list and automatically enriched list. For the 1699 principal variants that could be enriched automatically, recall is 2.5% lower; while for the remaining 778 principal variants, recall is 19.8% lower.

Table 1: Difference in the total number of documents retrieved between the manually compiled list and the automatically enriched list.

	Prescription drugs	Illicit drugs and higher level terms
Number of principal variants	1699	778
Total number of documents retrieved by the manually compiled list	5010	2332
Total number of documents retrieved by the automatically enriched list	4883	1871*
Difference (loss in recall)	127 (2.5%)	461 (19.8%)

* These principal variants could not map to any resource and hence no enrichment. Only the original name is used in document retrieval.

Discussion

The retrieval results show that the automatically enriched list is not performing as well as the manually compiled list. For the prescription drugs, recall is 2.5% lower. Given that the list is enriched automatically and misspelling generation is only performed for a subset of drugs, the automatically enriched list is doing fairly well with a slight loss in recall. The automatic enrichment approach could potentially eliminate the need for manual effort in collecting drug variants.

The potential benefits of automatically enriching the drug list lie in the aspects of sustainability, scalability and maintenance. In term of sustainability, the manually compiled list is doing well in the data scope of the year 2015, but it is not guaranteed it would be working as well for future data as it is now. Besides, manual effort is needed for maintenance of the manually compiled list as new drugs are introduced to the market periodically. Using RxNorm as a reference source for drug enrichment would eliminate the need for human experts updating the drug list, and thus resolve the maintenance issue.

However, for the remaining 778 principal variants, which mostly include illicit drugs, drug class terms and non-medicinal substances, recall is 19.8% lower. We performed an error analysis to identify which variants were critical to retrieval. The manually collected variants that contributed to the first 75% of missing retrieval are shown in Table 2. The first four terminology variants are ‘medication’ (for ‘medicine’), ‘beverage alcohol’ (for ‘alcoholic beverage’), ‘polysubstance’ (for ‘substance’), and ‘product of combustion’ (for ‘gases’). This shows that human effort is still needed to identify higher-level class terms. The last three terminology variants are ‘antiplatelet’ and ‘anti platelet’ (both for ‘anticoagulant’), and ‘antibiotic’ (for ‘antibacterial’). These variants are not exactly equivalent, but rather subclasses of their respective principal variants. This indicates that drug class terms could potentially be enriched by considering the subclass relationship in UMLS, which we will investigate in future work.

Table 2: Error analysis for non-prescription drug terms.

Principal variant	**Terminology variant**	**Total number of documents retrieved**
Medicine	Medication	222
Alcoholic Beverage	Beverage alcohol	46
Substance	Polysubstance	42
Gases	Product of Combustion	24
Anticoagulant	Antiplatelet	5
	Anti platelet	3
Antibacterial	Antibiotic	10

The manually compiled list clusters search terms with their principal variants without retaining the type of variant. In contrast, our automatic enrichment method retains the type of each terminology variant and this would help in determining the contribution of each type of variant. For example, the brand name ‘Coumadin’ is used as frequently as its generic drug name ‘Warfarin’ and this shows the significance of identifying brand names for Warfarin. A complete evaluation of the specific contribution of each type of variant will be performed in the future.

In this project, the comparison is only performed for principal variants found in the manually compiled list. In the future, we would like to incorporate all the drugs in RxNorm (and other sources to cover drug classes and illicit drugs) to assess whether such a drug list could be established with a completely automated approach. Other future work includes completing misspelling generation for all the drugs, as well as exploring resources such as UMLS for drug class terms enrichment.

CLASSIFICATION

ICD-10 is not flexible enough for aggregating death overdose cases into meaningful classes. Most of the analyses present drug overdose trends in terms of meaningful drug classes, instead of individual drugs. However, this presents a challenge as drugs that belong to the same drug class could have different ICD-10 codes. For examples, heroin (T40.1), opioids (T40.2), methadone (T40.3) and synthetic narcotics (T40.4) are all opioid-related drugs, but they all have different ICD-10 codes. Moreover, these codes are grouped together with non-narcotics drugs (e.g. LSD) under the group 'narcotics and psychodysleptics' (T40) making it impossible to determine opioid overdose trend directly from ICD-10 codes. Ruhm [9] identified the number of opioid-involved drug poisoning death by aggregating death certificate assigned any one of these six ICD-10 codes: T40.0, T40.1, T40.2, T40.3, T40.4 and T40.6.

The issue is less serious in a situation where a drug class is represented by several different ICD-10 codes as researchers could just aggregate drug overdose cases with those targeted ICD-10 codes. A more difficult situation would be when a researcher is interested in a drug class that is not represented by any ICD-10 code. One notable example of these drug classes would be fentanyl and its analogs.

In the year 2015, CDC released a health advisory reporting an increase in fentanyl-related unintentional overdose fatalities in multiple states across the U.S[4]. The number of fentanyl encounters has more than doubled in the US from 5,343 in 2014 to 13,882 in 2015[5]. Given the increase in fentanyl-related overdoses, it is important to specifically monitor fentanyl-related overdose trends. However, the ICD-10 code T40.4 (poisoning by synthetic narcotic) is not fine-grained enough to identify fentanyl-related overdose trends as it also includes other synthetic narcotics, such as tramadol. On the other hand, utilizing the literal text to identify overdose death involved with fentanyl is simply not adequate because most of the fentanyl deaths are not related to prescription fentanyl, but illicitly manufactured fentanyl analogs [8]. Thus, it is important to have a classification hierarchy in grouping all fentanyl products and their analogs to facilitate the analysis of fentanyl-related overdose trends.

In this project, we utilized two classification schemes to classify prescription drugs and drugs of abuse. We showed that these classification schemes provide sufficient flexibility for aggregation in producing meaningful higher-level statistics, and facilitate drug overdose trend analysis at the drug class level that is more useful in real life scenario.

Metabolite identification for enhanced retrieval

A metabolite is a molecule that is a byproduct of breaking down a drug or substance. Most of the drug toxicity tests detect the presence of metabolites in the body as the consumed drug is broken down quickly after entering the body. As such, physicians would record the presence of metabolites on death certificate as a proxy for the consumed drug.

Since the presence of the metabolite suggests the consumption of a specific drug, it is important to consider both in determining drug overdose incidence [5]. Horon et. al. [6] used enhanced surveillance to identify heroin overdose-related deaths. They considered a case as a heroin overdose-related death if the drug 'heroin' or its metabolite '6-monoacetymorphine' was mentioned in the death

[4] https://emergency.cdc.gov/han/han00384.asp

[5] https://www.cdc.gov/drugoverdose/data/fentanyl-le-reports.html

certificate, along with other criteria. They identified twice as many heroin overdose-related cases as found through the standard approach.

The manually compiled list does not have the information of metabolite-parent drug relationship. A metabolite is treated as an individual principal variant. This could cause incomplete retrieval of death certificates since the mention of the metabolite suggests the consumption of the corresponding drug. In this project, we identified metabolites among the principal variants and assessed the contribution of each metabolite in determining drug overdose trends of its parent drug.

Methods

Prescription drugs classification

RxNorm is integrated with RxClass which contains information about drug classes and drug-class membership. For prescription drugs, we utilized RxClass to obtain Anatomic Therapeutic Chemical (ATC) codes for each principal variant. ATC is a drug classification system that classifies drugs into a hierarchy with five different levels[6]. This hierarchy provides a means for aggregating drugs class.

Classification of drugs of abuse

ATC codes only cover prescription drugs and thus are not appropriate for classifying drugs of abuse. As drugs of abuse are highly involved in drug overdoses, we implemented a classification scheme for these drugs. The classification are based on the Drug Enforcement Agency (DEA)[7] drug resource guide and there is a total of eight classes:

- NARCOTICS
- STIMULANTS
- DEPRESSANTS
- HALLUCINOGENS
- CANNABIS
- INHALANTS
- STEROIDS
- DESIGNER DRUGS

Each class has its own characteristic properties but drugs within the same class produce a similar effect. The mapping of drugs of abuse to these classes is done partially by manual effort and partially by mapping from the DEA drug list. The DEA drug list is a drug list that is used within the agency, but not published. It has shallow groupings of drugs but the grouping does not follow a standard classification and is not consistent with the drug resource guide. We mapped the groups in the DEA drug list to the respective eight classes shown above. Death certificates retrieval is then performed for each class.

Metabolite identification

The principal variants consist of drug names as well as drug class terms, illicit drugs, non-medicinal substances and metabolites. Principal variants that are metabolites are identified via the PubChem[8] resource. Using the name of the principal variants as query, each of them are mapped to their respective compound identifier (CID), a unique identifier used in PubChem. PubChem is a database

[6] https://www.whocc.no/atc/structure_and_principles/
[7] https://www.dea.gov/pr/multimedia-library/publications/drug_of_abuse.pdf
[8] https://pubchem.ncbi.nlm.nih.gov/

of chemical molecules and it provides mapping of chemical molecules to different resources with the Human Metabolome Database (HMD) being one of them[9]. A metabolite is mentioned explicitly in the summary provided in HMD and could be detected using simple word matching. The metabolites and their parent drugs are extracted from the summary. Every metabolite is then used to retrieve relevant death certificates.

Results and Discussion

Prescription drugs classification

ATC codes enable grouping of drugs by drug class. We describe a use case that would showcase the significance of ATC codes. Suppose a researcher is interested in determining drug overdoses for different subclasses of antiepileptic drugs. Barbiturates and derivatives (N03AA) constitutes one of the subclasses and this subclass has a unique ICD-10 code, T42.3. Determining the drug overdose trend of this subclass is simple as the researcher could just retrieve death certificates which are assigned code T42.3. However, retrieval of other subclasses of antiepileptic drugs is less straightforward. For example, succinimide derivatives (N03AD) constitutes a subclass of antiepileptic drugs that does not have a unique ICD-10 code. To analyze the drug overdose trend for this subclass, a researcher has to manually identify all the drugs within the subclass before performing retrieval. Using this use case, we showed that utilizing ATC codes enables a researcher to readily identify the drugs within each subclass of antiepileptic drugs. As an illustration, the number of death certificates retrieved for each subclass are shown in Table 3.

Table 3: Number of documents retrieved for each subclass of antiepileptic drugs

Antiepileptic Drugs Subclass (ATC code)	Drug (ATC code)	Number of hits for each drug	Number of hits for each subclass
Barbiturates and derivatives (N03AA)	Methylphenobarbital (N03AA01)	0	4
	Phenobarbital (N03AA02)	4	
	Primidone (N03AA03)	0	
Hydantoin derivatives (N03AB)	Ethotoin (N03AB01)	0	1
	Phenytoin (N03AB02)	1	
	Mephenytoin (N03AB04)	0	
	Fosphenytoin (N03AB05)	0	
Oxazolidine derivatives (N03AC)	Paramethadione (N03AC01)	0	0
	Trimethadione (N03AC02)	0	

[9] http://www.hmdb.ca/hml

Benzodiazepine derivatives (N03AE)	Clonazepam (N03AE01)	11	11
Carboxamide derivatives (N03AF)	Carbamazepine (N03AF01)	4	5
	Oxcarbazepine (N03AF02)	1	
	Rufinamide (N03AF03)	0	
Fatty acid derivatives (N03AG)	Valproic acid (N03AG01)	2	2
	Vigabatrin (N03AG04)	0	
	Tiagabine (N03AG06)	0	
Other antiepileptics (N03AX)	Phenacemide (N03AX07)	0	38
	Lamotrigine (N03AX09)	11	
	Felbamate (N03AX10)	0	
	Topiramate (N03AX11)	1	
	Gabapentin (N03AX12)	25	
	Levetiracetam (N03AX14)	0	
	Zonisamide (N03AX15)	0	
	Pregabalin (N03AX16)	1	
	Lacosamide (N03AX18)	0	
	Carisbamate (N03AX19)	0	
	Retigabine (N03AX21)	0	

Classification of drugs of abuse

Drugs of abuse are classified into eight classes defined by DEA drug resource guide. Table 3 shows a examples of principal variants classified into the respective eight classes of drugs of abuse. Class-based retrieval was performed and the number of death certificates retrieved by all the drugs within a same class is shown in Table 4.

Table 4: Examples of principal variants for each drug of abuse class and number of death certificates retrieved by each class

Class of Drugs of abuse	Few Example of Principal Variants Classified Under the Class	Number of Death Certificates
Narcotics	OPIOID, OPIATE AGONIST, MORPHINE, METHADONE, CODEINE, ...	1200
Stimulants	STIMULANT, .ALPHA.-PVP, METHAMNETAMINE, BREPHEDRONE, ...	595
Depressants	DEPRESSANTS, BARBITURATE, PENTOBARBITAL, BENZODIAZEPINE, DIAZEPAM, ...	281
Hallucinogens	HALLUCINOGEN, DESCHLOROKETAMINE, NBOME, PSILOCIN, ...	1
Cannabis	CANNABIS, CANNABINOID AGONIST, JWH CANNABINOID, MDMB-CHMICA, ...	13
Inhalants	INHALANT, DEODORIZER, ORGANIC SOLVENT, REFRIGERANT, PAINT REMOVER, GLUE, ...	14
Steroids	STEROID, ANABOLIC STEROID, THENODONE ENANTHATE	67
Designer Drugs	1-(1-BENZOFURAN-6-YL)-N-METHYLPROPAN-2-AMINE, BATH SALTS	0

Table 4 shows that narcotics are responsible for the most drug overdose cases among the eight classes, followed by stimulants and depressants. Due to the large number of cases involving narcotics, we proposed several subclasses of narcotics. The subclasses of narcotics, their respective ICD-10 codes and the number of death certificates retrieved are shown in Table 5.

Table 5: Subclasses of narcotics with their respective ICD-10 and the total number of documents retrieved for each subclass.

Narcotics Subclass	ICD-10 Code	Principal Variants (PV)	No. of PV	Number of Hits
Opium	T40.0	Opium	1	0
Opiates		Opiate	1	235
Heroin	T40.1	Diamorphine	1	377
Oxycodone	T40.2	Oxycodone	1	157
Hydrocodone		Hydrocodone	1	62
Morphine		Morphine	1	117
Hydromorphone		Hydromorphone	1	41
Oxymorphone		Oxymorphone	1	14
Codeine		Codeine	1	21
Others		Levomethorphan, Apomorphine, Nicomorphine, Morphinone, Dextrorphan, Racemorphan, Levorphanol	7	0
Methadone	T40.3	Methadone	1	138
Others		Phenadoxone, Dimepheptanol	2	0
Fentanyl	T40.4	Fentanyl (36), Remifentanil (1), Carfentanil, Alfentanil, 3-methylfentanyl, Sufentanil, ...	23	37
Tramadol		Tramadol	1	36

Table 5 emphasizes the granularity issue in ICD-10 codes. The code T40.2 (poisoning by opioid) is shared by several different drugs: hydromorphone, oxycodone, morphine, and others. Each drug is responsible for a significant proportion of opioid overdoses and identifying the overdose trends for each drug might give important insights for policy making. The same issue affects the ICD-10 code T40.4 (poisoning by synthetic narcotic) as well. Fentanyl and its analogs present a recent concern of drug overdose in U.S. in recent years, but all these drugs share the same code with another synthetic narcotics, i.e. Tramadol. The proposed classification scheme allows aggregation of fentanyl and its analogs into one subclass which can facilitate the monitoring of fentanyl-related overdose cases.

The classification scheme we proposed for drugs of abuse not only facilitates aggregation into higher level drug classes (e.g., narcotics), but also supports aggregation into lower level drug subclasses (fentanyl and analogs), which helps facilitate the overdose trend analysis of drug classes that are in needs of immediate attention. As for now, the classification scheme is shallow and not refined enough. We had three meetings with NCHS team during the summer, but additional feedback is needed from domain experts in order to have a more practical and intuitive classification scheme. Future work includes revision of the classification schemes. We also want to perform retrieval using death certificates over a 10-year period to assess reproducibility of our methods.

Metabolite identification

Fifty principal variants are identified as metabolites. Among these 50 metabolites, only one of them (desmethylsertraline) retrieves one document. The increase in recall is marginal. Investigating death certificates over a 10-year period will help assess the contribution of metabolites to the retrieval of overdose cases.

Acknowledgements

I would like to thank Olivier Bodenreider for his patience and guidance throughout the summer project. I would also like to thank Margaret Warner and Kate Brett for their valuable feedback and insight in initiating and guiding the project.

References

1. Ruhm, C. J. (2016). Drug poisoning deaths in the United States, 1999–2012: a statistical adjustment analysis. *Population health metrics*, *14*(1), 2.
2. Ruhm, C. J. (2017). Drug involvement in fatal overdoses. *SSM-population health*, *3*, 219-226.
3. Warner, M., Chen, L. H., Makuc, D. M., Anderson, R. N., & Miniño, A. M. (2011). Drug poisoning deaths in the United States, 1980-2008. *NCHS data brief*, (81), 1-8.
4. Ossiander, E. M. (2014). Using textual cause-of-death data to study drug poisoning deaths. American journal of epidemiology, 179(7), 884-894.
5. Trinidad, J. P., Warner, M., Bastian, B. A., Miniño, A. M., & Hedegaard, H. (2016). Using literal text from the death certificate to enhance mortality statistics: Characterizing drug involvement in deaths.
6. Horon, I. L., Singal, P., Fowler, D. R., & Sharfstein, J. M. (2018). Standard Death Certificates Versus Enhanced Surveillance to Identify Heroin Overdose–Related Deaths. *American journal of public health*, *108*(6), 777-781.
7. Pimpalkhute, P., Patki, A., Nikfarjam, A., & Gonzalez, G. (2014). Phonetic spelling filter for keyword selection in drug mention mining from social media. *AMIA Summits on Translational Science Proceedings*, *2014*, 90.
8. Drug Enforcement Administration. (2016). Counterfeit prescription pills containing fentanyls: a global threat. US Department of Justice, DEA Strategic Intelligence Section, Springfield (VA).
9. Ruhm, C. J. (2018). Corrected US opioid-involved drug poisoning deaths and mortality rates, 1999–2015. Addiction.