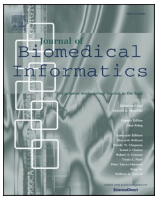
Journal of Biomedical Informatics 119 (2021) 103824

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

Detection of emerging drugs involved in overdose via diachronic word embeddings of substances discussed on social media

Austin P. Wright [a,b], Christopher M. Jones [b], Duen Horng Chau [a], R. Matthew Gladden [c], Steven A. Sumner [b,*]

[a] School of Computational Science & Engineering, Georgia Institute of Technology, Atlanta, USA
[b] Office of Strategy and Innovation, National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, Atlanta, USA
[c] Division of Overdose Prevention, National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, Atlanta, USA

ARTICLE INFO

Keywords:
Drug overdose
Fentanyl
Social media
Word embeddings
Natural language processing

ABSTRACT

Substances involved in overdose deaths have shifted over time and continue to undergo transition. Early detection of emerging drugs involved in overdose is a major challenge for traditional public health data systems. While novel social media data have shown promise, there is a continued need for robust natural language processing approaches that can identify emerging substances. Consequently, we developed a new metric, the relative similarity ratio, based on diachronic word embeddings to measure movement in the semantic proximity of individual substance words to 'overdose' over time. Our analysis of 64,420,376 drug-related posts made between January 2011 and December 2018 on Reddit, the largest online forum site, reveals that this approach successfully identified fentanyl, the most significant emerging substance in the overdose epidemic, >1 year earlier than traditional public health data systems. Use of diachronic word embeddings may enable improved identification of emerging substances involved in drug overdose, thereby improving the timeliness of prevention and treatment activities.

1. Introduction

1.1. Emerging causes of overdose

Reducing overdose deaths is a top public health priority in the United States.[1] In 2018, 67,367 Americans died from a drug overdose, nearly double the number of individuals dying from a drug overdose in 2009.[2] Indeed, the rapid rise in deaths due to drug overdose over the past decade contributed to decreases in overall life expectancy in the U.S. in 2015, 2016, and 2017.[3-6] One major challenge to reducing fatalities from drug overdose is that the substances involved in fatalities are continually shifting, creating substantial challenges with matching evidence-based interventions and resource deployment to shifting on-the-ground epidemiology in communities.[7,8] Research analyzing overdose fatalities over a multi-decade period has demonstrated that the U.S. has experienced various sub-epidemics primarily driven by different drugs over time.[8] In recent years, deaths from overdose in the U.S. have experienced a rapid epidemiologic transition from deaths involving prescription opioids to heroin to highly-potent synthetic opioids.[7] Illicitly manufactured synthetic opioids, namely fentanyl and its analogs, are now the most common drug class involved in overdose deaths in the U.S.[9]

1.2. Challenges for early detection

The recent and rapid emergence of fentanyl as the drug most commonly involved in overdose deaths in the U.S. illustrates the challenges with prevention of drug overdose as a result of emerging and evolving substances in the illicit drug supply.[10] Historically, fentanyl, which was primarily used as a prescription product in patch or oral lozenge form for the control of chronic pain[11,12] or as an injectable product in the operative setting for analgesia,[13] had relatively lower rates of misuse compared to other prescription opioids [14] and was not a primary driver of overall overdose mortality.[15] However, since 2013 there has been a rapid, substantial influx of illicitly manufactured fentanyl into the illicit drug supply in the U.S. that has corresponded with a

[*] Corresponding author at: National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, 4770 Buford Highway, Mailstop S106-9, Atlanta, GA 30341, USA.
E-mail address: hvo5@cdc.gov (S.A. Sumner).

https://doi.org/10.1016/j.jbi.2021.103824
Received 13 November 2020; Received in revised form 21 May 2021; Accepted 23 May 2021
Available online 26 May 2021
1532-0464/Published by Elsevier Inc.

rapid rise in overdose deaths involving synthetic opioids, especially fentanyl. [15-17]

National information on which substances are involved in overdose deaths is derived from death certificates and published by CDC. Unfortunately, owing to the complexity of post-mortem toxicologic testing, the increased time needed for investigation of deaths not from natural causes, rising numbers of overdose deaths needing investigation, inadequately staffed medical examiner and coroner offices, and the decentralization of death records, national information on which substances are involved in overdose deaths has traditionally lagged by 1 or more years. [18] Hence, major nationwide attention to the emergence of illicitly manufactured fentanyl and fentanyl analogs in the illicit drug supply and in overdose deaths did not occur as rapidly as possible given information delays on substances involved in overdose deaths. [19]

1.3. Novel data sources and emerging drug detection

Within the past decade there has been a growing body of literature examining the potential of novel social media data sources to better understand substance use patterns and trends.[20-25] However, major challenges persist in the development of a quantitative approach to successfully identify emerging drugs, with two predominant approaches existing. The first approach seeks to identify conversations about substances through keyword lists or lexicons and primarily focuses on examining the frequency or proportion that a given drug is mentioned. [19,26] The major limitation of this approach is that it requires knowledge of which drugs to search for by name and that examining counts or proportions is often inadequate as emerging substances are, by definition, mentioned infrequently during the early years of their emergence. The second leading approach to emerging drug detection focuses on building machine learning classifiers to detect types of posts—such as those discussing substance misuse—and then observe which substances are being mentioned in such conversations.[27] While machine learning models have progressed markedly in their ability to be able to accurately classify posts, the majority of substance misuse related posts discuss popular drugs. Emerging drugs, because they are initially rarer in mention, become hard to detect in the larger volume of conversations about commonly misused substances. Therefore, a need exists to identify improved quantitative approaches for emerging drug detection; in this manuscript we adapt and translate recent methodological approaches in computational linguistics to the task of emerging drug detection.

1.4. Measuring semantic shifts in natural language

Recent work in computational linguistics has explored “diachronic word embeddings,” for quantifying shifts in the meaning of words over time.[28,29] “Word embeddings” are complex numerical representations of words generated from a neural network model and “diachronic” refers to the fact that shifts in these numerical representations are being measured over time. For example, the meaning or context of the word “broadcast” has evolved over time, moving from a family of words used in farming (i.e., sowing seeds) to one nearer to media entertainment, and this shift can be both discovered in an automated fashion and quantitively described. [29] Researchers have recently shown that beyond studying language evolution, use of diachronic word embeddings can enable study of social phenomenon, such as change in gender bias over time through study of natural language.[30]

We hypothesized that use of diachronic word embeddings to measure semantic shifts in drugs over time from large volumes of public social media posts about substance use could potentially enable improved detection of emerging drugs involved in overdose faster than is currently possible with traditional public health surveillance systems. Using the emergence of fentanyl as a primary example, we demonstrate the applicability of diachronic word embeddings for such early detection. These findings can inform future innovative surveillance strategies and the development and implementation of more timely and targeted prevention and response strategies.

2. Methods

2.1. Data source

We analyzed anonymous, public posts from Reddit, the largest forum messaging site, which has over an estimated 400 million users and has been rated the third most visited website in the United States. [31] On the platform, users create various public message boards known as ‘subreddits’, where posts and comments are made. There exist a large number of forums dedicated to substance use-related topics.

We identified 225 drug-related subreddits from public lists of such forums (https://www.reddit.com/r/Drugs/wiki/subreddits) and manual qualitative searching of the site. Publicly available, historic Reddit data is accessible from multiple sources, including the Google BigQuery repository [32] and the Pushshift Reddit repository. [33] We downloaded and utilized data from January 2011 through 2018 through the application programming interface (API) of https://pushshift.io/.

2.2. Text processing and word embeddings

As the intent of the project was to measure finely grained temporal shifts in the meaning of words, we separated posts on Reddit into one-month intervals. We used all posts made to Reddit on 225 drug-related forums, including top-level posts, known as “submissions” on the platform, and all replies to those posts, known as “comments”. Each post was pre-processed using standard natural language processing techniques, including lowercasing text, removing hyperlinks, and removing punctuation.

We then constructed word embeddings using the word2vec model, a dominant method in natural language processing for this task, originally authored by researchers at Google. [34] Specifically, each word is represented by a string of numbers (in our case, 50 numbers) that specify an exact coordinate position or vector in 50-dimensional space. Prior studies in this field of research have used vector lengths ranging from 20 to 1000 dimensions [34]; we chose a 50-dimensional embedding for two main reasons. Firstly, the dataset used in our analysis—while containing tens of millions of posts—is significantly smaller than the original word2vec datasets using a vector length of 300 or more dimensions; reducing dimensionality in these cases helps to avoid overfitting. In Supplementary Table 1 we present formal evaluation of 20, 50, 150, and 300 dimensional vectors and present evidence of overfitting at larger dimensions. In addition to increased computational time with larger vector lengths, the dimensionality reduction techniques we use to visualize these embeddings (i.e., to reduce the information to just 2 dimensions for plotting as discussed below) are more challenging to compute in higher dimensions.

We implemented the word2vec model using the skip-gram architecture, given its generally better performance when working with less common words, which is of particular interest when trying to identify emerging trends. Additional parameters utilized for the word2vec model included a window size of 5, minimum required word count of 5 to compute a vector, 5 epochs for training, and a negative sampling value of 0. Word2vec is a neural network model that scans though the text of a given post and trains a model to predict the surrounding context given a word, as illustrated in Fig. 1. This process produces a vector representation for each word (the “word embeddings”). The resulting embeddings are shown to capture semantic similarity through distance metrics in vector space, as discussed below. We trained embeddings for every month from January 2011 through December 2018, resulting in a set of 50 dimensional vectors representing each word’s meaning in that month.

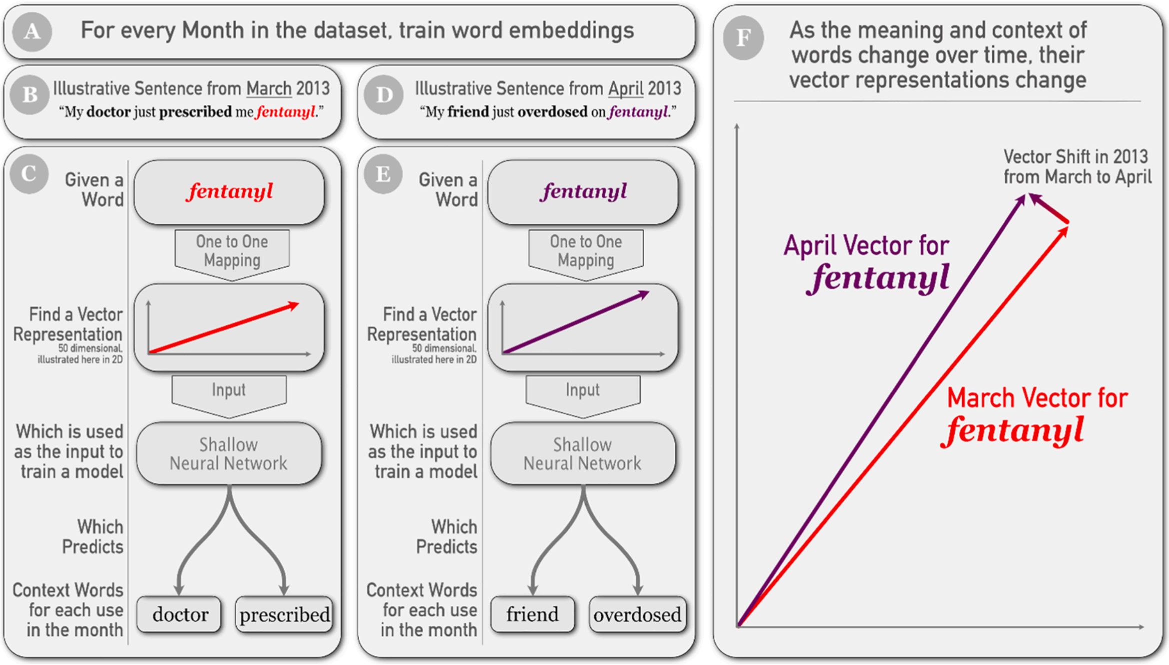

A For every Month in the dataset, train word embeddings
B Illustrative Sentence from March 2013
"My doctor just prescribed me fentanyl."
D Illustrative Sentence from April 2013
"My friend just overdosed on fentanyl."
C Given a Word
fentanyl
One to One Mapping
Find a Vector Representation
50 dimensional, illustrated here in 2D
Input
Which is used as the input to train a model
Shallow Neural Network
Which Predicts
Context Words for each use in the month
doctor
prescribed
E Given a Word
fentanyl
One to One Mapping
Find a Vector Representation
50 dimensional, illustrated here in 2D
Input
Which is used as the input to train a model
Shallow Neural Network
Which Predicts
Context Words for each use in the month
friend
overdosed
F As the meaning and context of words change over time, their vector representations change
Vector Shift in 2013 from March to April
April Vector for fentanyl
March Vector for fentanyl

Fig. 1. Visual Schematic of Word2Vec Processes for Measuring Temporal Shifts in Word Embeddings. Note: Figure presents a visual explanation of how word embeddings are constructed from natural language and used to assess temporal shifts in word meaning.

2.3. Measuring semantic change

The next challenge in analyzing semantic shift of words is ensuring that the embedding space that is learned for each time interval is comparable across time. To accomplish this we use a Procrustes transformation for each month to align with the previous month; this approach is widely used in computational linguistics for this task. [29]

Once we have developed this embedding space of word meaning over time, we can then develop ways to further extract specific insights related to words of interest. In particular, we hypothesized that measuring the semantic shift over time of a given drug word, such as 'fentanyl', could be a feasible approach to automate detection of emerging substances causing overdose, as drugs that are increasingly discussed in the context of overdose events would demonstrate increasing semantic proximity to 'overdose' related words.

In order to measure this type of shift, we define a new metric, which we refer to as the *Relative Similarity Ratio (RSR)*. The RSR represents how the similarity of a given word or set of words to another group of words can be calculated, taking into account a reference group. This general quantitative approach forms the basis of linguistic research using diachronic word embeddings. [30] However, while many other methods for analysis of semantic shift over time use generally static vocabularies and pure cosine similarity between words as an analysis metric, in our data the number of new unique words increases significantly over time. This causes pure cosine similarity to undergo 'inflation' as more words in the same space affects the relative angle between words, which requires correction to appropriately measure change over time. Our approach explicitly takes this into account by calculating a *relative* metric that finds similarity of words compared to a reference group. This method controls for rapidly changing dynamics in the size of the datasets used to populate the diachronic embedding space, such as constantly shifting, social media forums.

In order to calculate the RSR for a given word (w), the formula takes a given set of n reference words (R), and a set of m target words (T), as shown:

$$RSR(w) = \left(\frac{1}{m}\sum_{i=1}^{m}\frac{\langle w, T_i\rangle}{\|w\|\|T_i\|}\right) \Big/ \left(\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\langle R_j, T_i\rangle}{\|R_j\|\|T_i\|}\right)\right) \quad (1)$$

For each word w in a given timespan, the RSR is the average cosine similarity of the embedding of word w to a set of target words, which is then compared to the average cosine similarity of a set of reference words to the set of target words. This allows us to generate a flexible single metric on how similar any given word is to a set of target words, relative to a specified baseline.

Our main analysis for validation purposes examines semantic shift in the drug 'fentanyl'. In the formula presented above for the RSR the following set of ten fentanyl-related words, drawn from published literature [19] (fent, fents, fentanyl, fentynal, fentanil, fentanils, fentanyls, fentnyl, fetanyl, fentantyl), are each entered as w in the formula to capture semantic usage of the word fentanyl in posts. We calculated the RSR of each word in the fentanyl set every month and used this distribution of values in subsequent analyses using the statistical tests described below. The following set of overdose related terms (od, overdose, ods, oding, overdosing, overdosed, overdoses) are entered as *T,* the target words to which cosine similarity of the fentanyl words is measured.

Lastly, we created a reference group of words R by which the cosine similarity between these words and the target words are also calculated and then compared to the cosine similarity of fentanyl to the target words, to yield a ratio as presented in the RSR. For the reference group we selected ten common prescription opioid terms that were present in each month of the entire dataset and broadly covered common generic/brand formulations (oxycodone, oxy, roxy, percocet, oxycontin, hydrocodone, vicodin, hydromorphone, dilaudid, morphine); these are entered as *R*, the reference words by which the cosine similarity between these words and the target words are also calculated and then compared as a ratio to the cosine similarity of fentanyl to the target words as presented in the RSR. Of note, for the common prescription opioid terms, we did not enter extensive lists of spelling variants given the sizeable number of drugs already represented in the list and the fact that previous research on prescription opioids has revealed that correct spellings are the most prevalent variant and reference words must be present in the text over the entire time period studied. [25] Furthermore, spelling variants/errors already exist in close semantic proximity to the base words and thus do not meaningfully alter our measures of semantic distance. Sensitivity analyses testing the inclusion of additional variants revealed no change in results. Including potential spelling variants mainly aids in identifying a sufficient number of mentions of a given drug if that substance is relatively rare in the overall corpus, which aids in tightening confidence intervals around estimates for rare substances.

For illustrative purposes and to show the semantic movement of fentanyl relative to drugs other than prescription opioids, we calculate RSR values over time for the following additional substances/categories. Up to 10 spelling or colloquial variants for each word are included, drawn from previous research on these substances. [24,26,35,36]

- **Heroin**: heroin, heroins, herion, h, heroine, heorin, tar
- **Cannabis:** cannabis, cannibis, marijuana, mj, marajuana, marihuana, ganja, pot, weed
- **Methamphetamine:** methamphetamine, methamphetamines, methamp, methampetamine, crank, meth, speed, ice, shards, crystal
- **Cocaine:** cocaine, cocain, blow, coke, crack, crakc, coca, yayo

2.4. Statistical analysis

To describe crude trends in word frequency over time in Reddit posts, we first calculated the proportion of words in a particular drug category out of all words used in a given year from 2011 to 2018 (i.e., how many times fentanyl words are used out of all words in a given year). For added context, we also calculated overdose deaths involving prescription opioids and overdose deaths involving synthetic opioids annually from 2011 to 2018, the most recent year for which data are available, based on data from the National Center for Health Statistics' National Vital Statistics System. [18] Overdose deaths were those with an ICD-10 underlying cause of death code: X40-44, X60-64, X85, Y10-14. Overdoses involving prescription opioids had T40.2-T40.3 in the multiple-cause-of-death field. Overdoses involving synthetic opioids had T40.4 in the multiple-cause-of-death-field.

In order to verify that the signal we measure is a meaningful indicator of emerging overdose trends, we examined the case of fentanyl and compared it to other prescription opioids, with the hypothesis being that the RSR of fentanyl to overdose would initially be within the range of other prescription opioids, and over time it would diverge from the reference group as illicitly manufactured fentanyl is increasingly involved in overdoses. We calculated the RSR of each word in the fentanyl set every month and compared this distribution over the course of each year to the RSR values generated from the prescription opioid set using a Mann Whitney U Test. P values < 0.05 were considered statistically significant.

2.5. Visualizing word shifts

Diachronic embeddings contain a large amount of information regarding the changing meaning of words over time, however accessing the full breadth of that information is still difficult since we cannot visualize vectors in 50 dimensions. We created 2 plots to visually explore information from word embeddings. The first (Fig. 2) presents a higher-level exploration of the movement of the word vector for fentanyl over the study period. This plot uses a leading dimensionality reduction algorithm called Uniform Manifold Approximation and Projection (UMAP) to compress the 50 dimensional vectors into 2 dimensions. [37] For ease of visualization in this figure, we plot movement in the word "fentanyl" and do not include spelling variants as these exist in close proximity to the correctly spelled word. The trajectory for "fentanyl" is calculated using a smoothed univariate cubic spline over the set of points for every month. For general context, in this figure we also plot a subset of other drug words representing some of the most common prescription opioids and illicit substances for general context.

The second plot we create (Fig. 3) shows changes in the RSR values over time, which is a more complex calculation that takes into account multiple drug words and spelling variants, along with a reference group, as described above. Plots are made using the Python Seaborn library and the distribution of the RSR for fentanyl and its spelling variants are plotted with a polynomial trendline and 95% confidence interval.

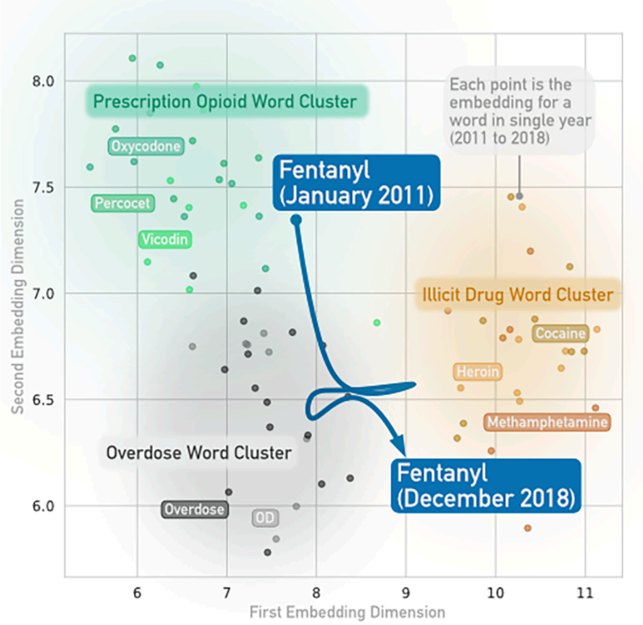


Fig. 2. Semantic Movement of Fentanyl in Two-Dimensional Space, 2011 to 2018. Note: Fig. 2 plots the trajectory of the word 'fentanyl' (blue arrow) from 2011 to 2018, as calculated from the positions of each month's word embeddings by the Uniform Manifold Approximation and Projection (UMAP) algorithm. UMAP compresses the 50 dimension word vectors into just 2 dimensions for visualization on a standard coordinate plane; each UMAP axis is an arbitrary unitless number representing position on the best fitting two dimensional structure and is intended to allow for the visualization of the relative position of nearby words and their clusters. For ease of visualization, only select drug words are plotted and spelling variants are not included as they exist in very close vector space to the correctly spelled substance word. Words illustritive of prescription opioids (such as oxycodone and Percocet), other illicitly used substances (cocaine, methamphetamine, heroin), and overdose (overdose, od) are also plotted with their yearly values to reveal the general semantic space in which these words exist. Fentanyl moves from close proximity to other prescription opioids toward illicitly used substances and overdose. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Data availability

Post data is publicly available via Pushshift.io; the trained word vectors used in this research are available upon request.

4. Results

A total of 64,420,376 drug-related posts between January 2011 and December 2018 were included in our final analysis (Table 1). Over the time period examined, the total number of posts made in each year increased from 3,836,916 in 2011 to 13,704,634 in 2018. As a proportion of all words used in drug-related posts in a given year, fentanyl increased several fold from 0.54 per 100,000 words in 2011 to 15.57 per 100,000 words in 2018. Common prescription opioid words also increased in usage from 9.81 per 100,000 words in 2011 to 29.73 per 100,000 words in 2018. Word usage trends are also presented in Table 1 for the additional drug categories studied.

Fig. 2 plots the trajectory of the UMAP dimensionality-reduced word vectors for fentanyl from 2011 to 2018. As shown by the blue arrow, fentanyl shows two important patterns during the study period. First, fentanyl moves from close proximity to other prescription opioids such as oxycodone and Percocet, toward illicitly used substances, such as

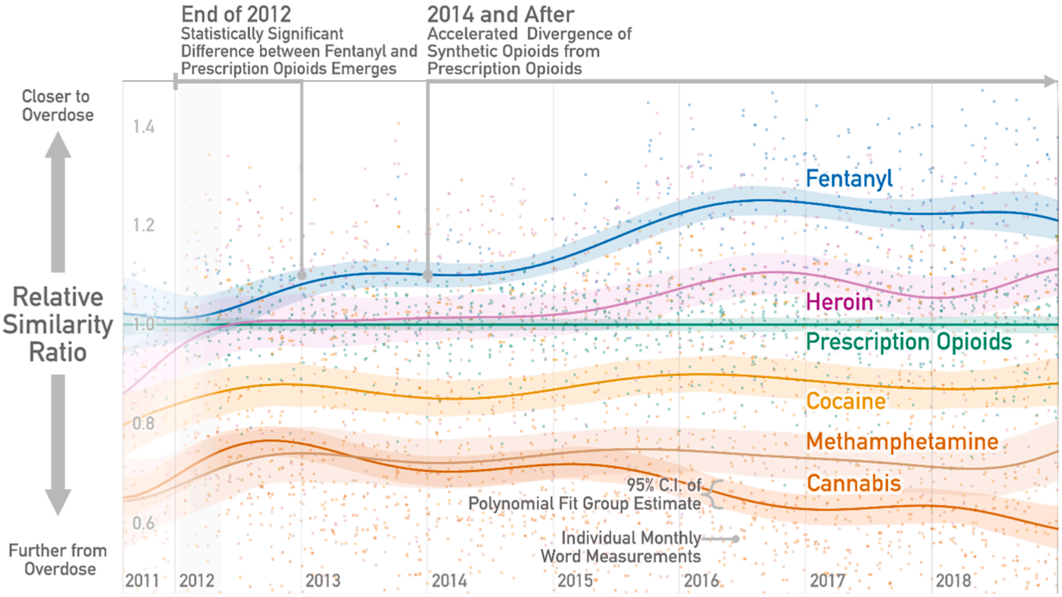


Fig. 3. Relative Similarity Ratio Displaying Semantic Proximity of Various Substances to Overdose Over Time. Note: Y-axis displays the Relative Similarity Ratio (RSR) metric, which compares the strength of the semantic association between a given substance (i.e., fentanyl) and overdose to the strength of the semantic association between the reference group (common prescription opioids, centered at 0) and overdose. The degree to which a substance is above the green horizontal center line reveals its semantic proximity to overdose, relative to the common prescription opioid terms. Trendlines are drawn through the monthly RSR values for each word in a given category (i.e., fentanyl and its spelling variants) with a 10th order polynomial approximation for the group trend and shading displaying a surrounding 95% confidence interval. Words used for each category are included in the methods section. Partial year data shown for 2011 due to sparse data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Frequency of Substance Mentions in Reddit Posts Over Time, 2011–2018.

				Proportion of Words from a Given Category per 100,000 Words							
Year	Total Post Volume (N)	Total Words	Unique Words Per Year (N)	Fentanyl[1]	Common Prescription Opioids[2]	Methamphetamine	Cocaine	Heroin	Cannabis	Synthetic Opioid Deaths[3]	Prescription Opioid Deaths[4]
2011	3,836,916	74,884,763	383,261	0.54	9.81	26.36	34.29	17.16	354.45	2666	15,140
2012	6,915,352	140,974,586	545,869	1.43	18.55	27.13	36.12	24.08	340.47	2628	14,240
2013	3,990,656	98,962,910	463,896	2.63	29.15	39.79	46.42	39.95	227.26	3105	14,145
2014	7,265,859	176,198,238	652,449	1.93	15.14	36.44	43.14	26.45	271.29	5544	14,838
2015	8,538,950	223,211,789	748,023	5.96	24.98	42.54	45.95	39.51	230.12	9580	15,281
2016	9,821,619	262,394,595	806,880	11.45	28.36	50.02	48.84	41.82	201.28	19,413	17,087
2017	10,346,390	269,206,729	838,493	17.63	30.70	49.74	50.50	44.65	161.68	28,466	17,029
2018	13,704,634	360,365,525	912,460	15.57	29.73	51.58	54.41	41.51	208.20	31,335	14,975

[1] The following spelling variants are used to identify fentanyl posts: fent, fents, fentanyl, fentynal, fentanil, fentanils, fentanyls, fentnyl, fetanyl, fentantyl

[2] oxycodone, oxy, roxy, percocet, oxycontin, hydrocodone, vicodin, hydromorphone, dilaudid, morphine

[3] Deaths in United States as determined by ICD-10 underlying cause of death code of X40-44, X60-64, X85, or Y10-14 and multiple-cause-of-death code of T40.4. Fentanyl is the leading contributor to synthetic opioid deaths.

[4] Deaths in United States as determined by ICD-10 underlying cause of death code of X40-44, X60-64, X85, or Y10-14 and multiple-cause-of-death code of T40.2 or T40.3

heroin and cocaine. Second, fentanyl moves more closely to overdose and overdose-related terms.

Fig. 3 plots the Relative Similarity Ratio (RSR) displaying the semantic proximity of various substances to overdose over time, in relation to the reference category of common prescription opioids. The figure show both the individual measurements of the RSR for each word in the substance groups, as well as the overall group trendline calculated by a 10th order polynomial. Both fentanyl and heroin exhibit an upward trend in RSR over time, suggesting an increasing semantic proximity to overdose. The RSR for fentanyl increases beyond that of heroin, indicating that semantic proximity of fentanyl to overdose is greater than that of heroin to overdose. The RSR of fentanyl peaks in 2016 at 1.25 (95% CI 1.22–1.28) and plateaus thereafter, relative to the reference group of prescription opioids. The RSR for other illicit drugs including cocaine, methamphetamine, and cannabis is lower than the reference category of common prescription opioids.

Annual differences in the distribution of the RSR for fentanyl and overdose compared to prescription opioids and overdose are depicted in Table 2. On an annual basis, statistically significant differences between fentanyl and other prescription opioids emerge by 2012. After 2012, the strength of the statistical difference markedly increases. This emergence of fentanyl from real-time online data from 2012 was >1–2 years earlier

Table 2
Mann Whitney U Test for Significant Differences by Year Between Fentanyl and Common Prescription Opioids[1] as Measured by the Relative Similarity Ratio (RSR).

Year	2011	2012	2013	2014	2015	2016	2017	2018
p-value	3.88e−01	1.01e−03	6.97e−09	1.36e−10	3.31e−19	1.48e−22	1.53e−24	1.07e−21

Note: The Relative Similarity Ratio compares the strength of the semantic association between fentanyl and overdose to the strength of the semantic association between the common prescription opioids and overdose.

[1] Common prescription opioid terms include: oxycodone, oxy, roxy, percocet, oxycontin, hydrocodone, vicodin, hydromorphone, dilaudid, morphine

than information available to public health professionals at the time; the first increase in deaths from fentanyl occurred in 2013 but this information was not observed by public health professionals until 2014 owing to the 1 year lag in mortality data reporting nationally.

5. Discussion and conclusion

In this research, we adapt a mathematical approach from computational linguistics based on diachronic word embeddings and demonstrate that such a measure holds promise for enabling automated detection of emerging drugs involved in overdose. Indeed, our findings yield intelligible results on substance word shifts that can be assessed both quantitatively and visually.

Although multiple studies have indicated that real-time social media data on illicit substances reveal potentially useful correlations with findings from slower, traditional epidemiologic systems, [19,26,38,39] there have been a number of challenges to-date in fully realizing the potential of social media data for emerging drug detection. As noted in the introduction, drug detection efforts using a purely lexicon-based approaches rely on a keyword list of drugs of interest to search large volumes of social media text for matching mentions and track frequencies and proportions of drug terms overtime. Lexicon-based approaches, while widely used, present challenges for emerging drug detection as emerging drugs are rare in frequency earlier in their emergence. [19,26] Machine learning based approaches, which attempt to build linguistic classifiers that can detect whether a given post is discussing substance use, [20] also struggle with emerging drug detection as initially the raw frequency of such substances may be very small relative to commonly used substances, causing initially rarer, novel compounds to be difficult to identify among the large volume of discussions about other drugs.

The diachronic word embedding approach which we apply in this manuscript aims to further techniques to overcome a number of these challenges. Perhaps most importantly, this approach is independent of the frequency of mentions of a given substance. Although a larger number of drug mentions helps to increases confidence around estimates, the semantic distance between a given substance and overdose can be calculated for any quantity of posts. Secondly, through developing our measure as a ratio and with the inclusion of a reference group, we are able to examine the substances most closely associated with overdose in spite of myriad semantic shifts that occur over time.

While the current study focuses on retrospectively validating the ability of diachronic word embeddings to identify changes in the semantic context of fentanyl, a key emerging drug of the past decade, ultimately prospective deployment and detection of emerging drugs is needed. While a longitudinal, prospective validation is beyond the scope of the current manuscript, there are a couple approaches by which we believe such a system could operate in a semi-supervised way. First, because emerging drug names are often not known a priori, we could extract a large corpus of potential drug words by querying for cosine-similarity to known substances. Although emerging drugs could conceivably arise in a completely new pharmacologic class, in general, emerging drugs are variants of existing drug classes and compounds, such as carfentanil and other fentanyl analogs emerging following fentanyl. After extracting a large number of potential candidate drugs, these drug names would then be passed through the RSR metric to measure their semantic proximity to overdose overtime. A human with expertise in the subject matter, such as a toxicologist would then examine the drugs which show the closest proximity to overdose or with rapid movement toward overdose and use this information to inform further study of these compounds using traditional public health and laboratory systems. It should be noted that it is also possible to pass every word in the corpus (which would encompass drug words as well as non-drug words) through the RSR calculation. This approach would be expected to increase sensitivity of detection of new substances but at the potential added cost of additional person-time to inspect an increased number of candidate words. Substances of concern from social media derived signals, could be verified by early field studies, [40] undergo validation by inspection of trends from other data systems such as prescribing data, Emergency Department based syndromic surveillance, or illicit drug seizures by law enforcement, [16] or aid in the development and deployment of relevant laboratory assays.

Some limitations of this work and areas for future research should be noted. Overall, generalizability of these models to future emerging substances and applicability to other social networking sites given the unique nature of language on each site is not known and merits future research. Further prospective validation is also needed to fully refine the most successful target word(s) for emerging drug detection. In our analyses, we used a set of overdose related terms as the target words to which to cosine similarity of emerging drugs was measured. In our testing, this target set was intuitive and worked well for synthetic opioids such as fentanyl; however, detection of other emerging or re-emerging substances (which may not be primary causes of overdose but rather are co-used with opioids or simply used for psychoactive effects but are not highly fatal) will likely require an alternate or expanded set of target words. For example, if one wished to identify emerging drugs of use that are not necessarily associated with overdose, the target set of words might include the terms “high”, “amped”, “stoned”, etc. Nonetheless, our measure is flexible to these adjustments. Additionally, macro-level secular changes in word usage has the potential to alter the results of such an approach. For example, towards the end of the time period studied, we note a decrease in the RSR of fentanyl, though it still is significantly higher than the common prescription opioids reference group. Although fentanyl remains the single leading drug causing excess overdose mortality in the U.S., fentanyl is now widely appreciated as such and is no longer an “emerging drug.” Thus, posts discussing fentanyl have likely, to some degree, shifted to other topics beyond overdose concerns, causing a slight decrease in RSR. It is also important to note that beyond the semantic proximity of fentanyl to overdose, the strength of the association of the reference words with overdose also influences the RSR for fentanyl. We suspect that our approach is most useful for detecting emerging drugs or detecting shifts in drug use patterns.

Nonetheless, this research helps advance computational approaches to substance use and overdose prevention and may have broader applicability for other public health use cases. For example, better quantifying language change over time may aid in the study of public perception of health-related policies or in understanding large-scale public shifts in norms, behaviors, and beliefs. For example, understanding public shifts in beliefs about corporal punishment for children is a key area of interest for violence prevention researchers that lacks large scale quantitative information and could be further explored through these methods. Similarly, understanding changing levels of stigmatization and stigmatizing language over time surrounding discussion of mental health could be explored through these methods. The use of prescription and illicit drugs continues to undergo evolution in the U.S., including among opioids, stimulants, cannabinoids, and other substances. [42-43] While social media data hold promise for emerging threat detection, the development of robust and scalable mathematical approaches are urgently needed to extract insights in a way that is manageable and actionable for public health professionals. Automated assessment of semantic shifts in substances as detected in large volumes of unstructured text may improve efforts at early detection of emerging drugs and thereby accelerate early recognition and prevention and response efforts by public health and clinical professionals.

CDC disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

CRediT authorship contribution statement

Austin P. Wright: Conceptualization, Data Curation., Formal analysis, Methodology, Visualization, Writing - original draft. **Christopher M. Jones:** Methodology, Project administration, Resources, Supervision, Writing - review & editing. **Duen Horng Chau:** Methodology, Project administration, Resources, Supervision, Visualization, Writing - review & editing. **R. Matthew Gladden:** Methodology, Project administration, Resources, Supervision, Writing - review & editing. **Steven A. Sumner:** Conceptualization, Data curation, Formal analysis, Supervision, Methodology, Writing - original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103824.

References

[1] National Drug Control Strategy. Office of National Drug Control Policy, Feb 2020. Available at: https://www.whitehouse.gov/wp-content/uploads/2020/02/2020-NDCS.pdf.

[2] Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. Web-based Injury Statistics Query and Reporting System. Available at: http://www.cdc.gov/injury/wisqars/index.html, 2019.

[3] S.L. Murphy, J. Xu, K.D. Kochanek, E. Arias, Mortality in the United States, 2017. NCHS Data Brief. No. 328. Nov 2018, 2018.

[4] J. Xu, S. Murphy, K. Kochanek, E. Arias, Mortality in the United States, 2015. NCHS data brief, no 267. Hyattsville, MD: US Department of Health and Human Services, CDC. National Center for Health Statistics, 2016.

[5] K.D. Kochanek, S.L. Murphy, J. Xu, E. Arias, Mortality in the United States, 2016. NCHS Data Brief. No 293, December 2017.

[6] D. Dowell, E. Arias, K. Kochanek, et al., Contribution of Opioid-Involved Poisoning to the Change in Life Expectancy in the United States, 2000–2015, JAMA 318 (11) (2017) 1065–1067, https://doi.org/10.1001/jama.2017.9308.

[7] W.M. Compton, C.M. Jones, Epidemiology of the US opioid crisis: the importance of the vector, Ann. NY Acad. Sci. 1451 (2019) 130–143.

[8] H. Jalal, J.M. Buchanich, M.S. Roberts, L.C. Balmert, K. Zhang, D.S. Burke, Changing dynamics of the drug overdose epidemic in the United States from 1979 through 2016, Science 361 (6408) (2018), https://doi.org/10.1126/science.aau1184.

[9] Wilson N. Drug and Opioid-Involved Overdose Deaths—United States, 2017–2018. MMWR Morbidity and Mortality Weekly Report 2020;69.

[10] H. Hedegaard, B.A. Bastian, J.P. Trinidad, M. Spencer, M. Warner, Regional differences in the drugs most frequently involved in drug overdose deaths: United States, 2017. National Vital Statistics Reports, vol. 68, no. 12, October 25, 2019.

[11] Duragesic Prescribing Information. Food and Drug Administration. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/019813s079lbl.pdf.

[12] Fentora Prescribing Information. Food and Drug Administration. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/021947s029lbl.pdf.

[13] Fentanyl Citrate Prescribing Information. Food and Drug Administration. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/019115s033lbl.pdf.

[14] S.F. Butler, R.A. Black, T.A. Cassidy, T.M. Dailey, S.H. Budman, Abuse risks and routes of administration of different prescription opioid compounds and formulations, Harm Reduct. J. 8 (1) (2011) 29.

[15] M.R. Spencer, M. Warner, B.A. Bastian, J.P. Trinidad, H. Hedegaard, Drug overdose deaths involving fentanyl, 2011–2016, National Vital Statistics Reports National Center for Health Statistics National Vital Statistics System 68 (3) (2019) 1–19.

[16] NFLIS Brief: Fentanyl, 2001–2015. U.S. Department of Justice, Drug Enforcement Agency, Diversion Control Division. Mar 2017. Available at: https://www.nflis.deadiversion.usdoj.gov/DesktopModules/ReportDownloads/Reports/NFLISFentanylBrief2017.pdf.

[17] J.K. O'Donnell, R.M. Gladden, P. Seth, Trends in deaths involving heroin and synthetic opioids excluding methadone, and law enforcement drug product reports, by census region—United States, 2006–2015, MMWR Morb. Mortal. Wkly Rep. 66 (34) (2017) 897.

[18] Centers for Disease Control and Prevention. CDC WONDER. Multiple Cause of Death Data. Available at: https://wonder.cdc.gov/mcd.html.

[19] D.A. Bowen, J. O'Donnell, S.A. Sumner, Increases in Online Posts About Synthetic Opioids Preceding Increases in Synthetic Opioid Death Rates: a Retrospective Observational Study, J. General Int. Med. (2019) 1–3.

[20] A. Sarker, K. O'Connor, R. Ginn, et al., Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter, Drug Saf. 39 (3) (Mar 2016) 231–240, https://doi.org/10.1007/s40264-015-0379-4.

[21] J. Kalyanam, T. Katsuki, G.R. Lanckriet, T.K. Mackey, Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using unsupervised machine learning, J. Addictive Behaviors 65 (2017) 289–295.

[22] T. Katsuki, T.K. Mackey, R. Cuomo, Establishing a Link Between Prescription Drug Abuse and Illicit Online Pharmacies: Analysis of Twitter Data, J. Med. Internet Res. 17 (12) (2015), e280, https://doi.org/10.2196/jmir.5144.

[23] T.K. Mackey, J. Kalyanam, T. Katsuki, G. Lanckriet, Twitter-Based Detection of Illegal Online Sale of Prescription Opioid, Am. J. Public Health 107 (12) (Dec 2017) 1910–1915, https://doi.org/10.2105/AJPH.2017.303994.

[24] M. Chary, N. Genes, C. Giraud-Carrier, C. Hanson, L.S. Nelson, A.F. Manini, Epidemiology from Tweets: Estimating Misuse of Prescription Opioids in the USA from Social Media, J. Med. Toxicol. 13 (4) (Dec 2017) 278–286, https://doi.org/10.1007/s13181-017-0625-5.

[25] A. Sarker, G. Gonzalez-Hernandez, Y. Ruan, J. Perrone, Machine Learning and Natural Language Processing for Geolocation-Centric Monitoring and Characterization of Opioid-Related Social Media Chatter, JAMA Netw Open. 2 (11) (2019) e1914672, https://doi.org/10.1001/jamanetworkopen.2019.14672.

[26] S.A. Sumner, T. Haegerich, C.M. Jones, Temporal Trends in Online Posts about Vaping of Cannabis Products, J. Addict. Med. (2020).

[27] A. Sarker, R. Ginn, A. Nikfarjam, et al., Utilizing social media data for pharmacovigilance: a review, J. Biomed. Inform. 54 (2015) 202–212.

[28] W.L. Hamilton, J. Leskovec, D. Jurafsky, Cultural shift or linguistic drift? comparing two computational measures of semantic change. NIH Public, Access (2016) 2116.

[29] W.L. Hamilton, J. Leskovec, D. Jurafsky, Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change (2016) 1489–1501.

[30] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proc. Natl. Acad. Sci. USA 115 (16) (2018) E3635–E3644, https://doi.org/10.1073/pnas.1720347115.

[31] C. Nguyen, Reddit beats out Facebook to become the third-most-popular site on the web. Digital Trends. May 30, 2018. Available at: https://www.digitaltrends.com/computing/reddit-more-popular-than-facebook-in-2018/.

[32] Google Cloud. BigQuery public datasets. Available at: https://cloud.google.com/bigquery/public-data.

[33] Pushshift. Available at: https://pushshift.io/.

[34] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781, 2013.

[35] S. Chancellor, G. Nitzburg, A. Hu, F. Zampieri, M. De Choudhury, Discovering alternative treatments for opioid use recovery using social media, (2019) 1–15.

[36] Drug Facts. Drug Enforcement Agency. Available at: https://www.dea.gov/factsheets.

[37] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv e-prints, 2018:arXiv:1802.03426. Accessed February 01, 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv180203426M.

[38] M.J. Paul, M.S. Chisolm, M.W. Johnson, R.G. Vandrey, M. Dredze, Assessing the Validity of Online Drug Forums as a Source for Estimating Demographic and Temporal Trends in Drug Use, J. Addict. Med. 10 (5) (2016) 324–330, https://doi.org/10.1097/ADM.0000000000000238.

[39] M. Chary, N. Genes, C. Giraud-Carrier, C. Hanson, L.S. Nelson, A.F. Manini, Epidemiology from Tweets: Estimating Misuse of Prescription Opioids in the USA from Social Media, J. Med. Toxicol. 13 (4) (2017) 278–286.

[40] M.C. Mercado, S.A. Sumner, M.B. Spelke, M.K. Bohm, D.E. Sugerman, C. Stanley, Increase in drug overdose deaths involving fentanyl—Rhode Island, January 2012–March 2014, Pain Med. 19 (3) (2017) 511–523.

[41] S. Elliott, R. Sedefov, M. Evans-Brown, Assessing the toxicological significance of new psychoactive substances in fatalities, Drug Test. Anal. 10 (1) (2018) 120–126.

[42] C.M. Jones, Patterns and Characteristics of Methamphetamine Use Among Adults—United States, 2015–2018. MMWR Morbidity and Mortality Weekly Report 2020, 69.

[43] M. Kariisa, L. Scholl, N. Wilson, P. Seth, B. Hoots, Drug Overdose Deaths Involving Cocaine and Psychostimulants with Abuse Potential - United States, 2003–2017, MMWR Morb. Mortal Wkly. Rep. 68 (17) (2019) 388–395, https://doi.org/10.15585/mmwr.mm6817a3.