Estimating Nonmedical Use of Prescription Opioids in US from Social Media

Michael Chary PhD[1], Nicholas Genes, MD, PhD[2], Christophe Giraud-Carrier, PhD[3], Carl Hanson, PhD[4], Lewis Nelson, MD[5], Alex Manini, MD, MS, FACMT[2,6]

[1]Icahn School of Medicine, NY; [2]Dept. of Emergency Medicine, Mt. Sinai NY; [3]Dept. of Computer Science Brigham Young University; [4]Dept. of Health Science Brigham Young University;
[5] Dept. of Emergency Medicine New York University;[6]Div. Of Medical Toxicology, Mount Sinai Hospital

Introduction

• The non-medical use of prescription drugs (NMUPD) is a significant public health burden.
• Social media provide data that may help us understand NMUPD in the general population.
• Until now, social media have played a limited role in public health research, partially owing to a lack of validated methods for estimating essential epidemiological quantities from social media.

Objective

To demonstrate that the point prevalence of opioid NMUPD can be accurately and rapidly estimated from publicly available data from Twitter.

Methods

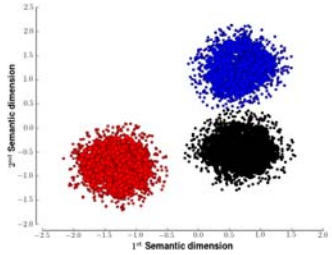
Design. Cross-sectional study of opioid abuse via Twitter.

Data Source. We analyzed three months of tweets from Twitter's streaming API for tweets that mentioned the words in **Table 1**. We excluded tweets that only contained links to websites.
Preprocessing. Once acquired, all text was converted to lowercase, non-ASCII characters ignored, stopwords removed, and all words lemmatized. Lemmatization refers to converting variations on a word, such as 'ran','run', and 'running' into one base form, e.g. 'run'.
Geocoding. To geolocate the tweet we used latitude and longitude coordinates in the metadata of the tweet. Since only 1-2% of tweets contain explicit information, we used Carmen— a program that infers location from the text and metadata of a tweet— to approximate the location of more tweets.
Semantic Distance. The semantic distance between two words is the average of the minimal path lengths between all combinations of senses of those two words.

Comparison We calculated the location quotient for each state in the continental US.We validated our estimate by calculating its correlation with the location quotient calculated from the 2012 National Survey on Drug Use and Health.

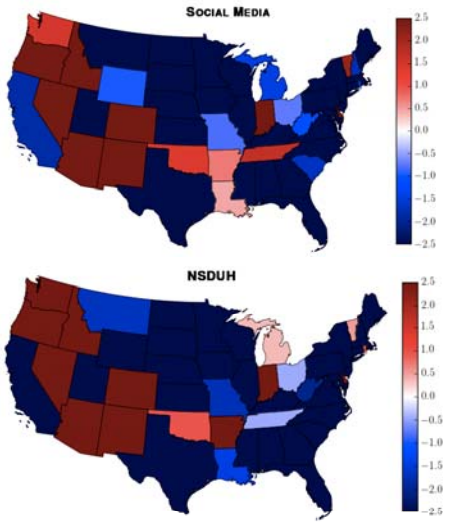
Medical Names	Street Names
Morphine	Dope
Methadone	Pain killers
Coedine	Oxy
Hydrocodone	OC
Oxycodone	Percs
Propoxyphene	Pancakes and Syrup
Fentanyl	Demmies
Tramadol	Captain Cody
Roxanol	Tango and Cash

Results

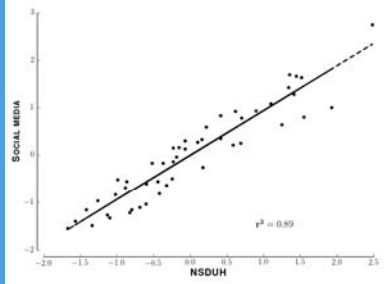


Separation of tweets into semantic clusters. Projection of semantic distances of tweets onto the first two semantic dimensions. Red cluster denotes opioids, blue cluster denotes other drugs, black unrelated or "noise" tweets.

Results



Top: Map of location quotients estimated from Twitter.
Bottom: Map of relative density of illicit use of pain relievers in the past year from Table 8 of 2011 NSDUH. Both panels use the same color scale.



Limitations

• **Biased Sampling**. Drug users who tweet may be a minority of drug users. Or, they may use differently than other drug users. Only 1% of tweets are geocoded. Using Carmen only increases that fraction to 20%.
• **Biased Comparison**. The NSDUH data and Twitter data are not from the same time period. This suggests that our analysis captured a slow dynamic, instead of the pattern of opioid abuse. We did control for variation in population density.

Conclusions

• Using this novel semantic distance, we were able to appropriately separate drug conversations on Twitter by topic

• Discussions of opioid use on Twitter and in national surveys follow the same prevalence distributions, which has widespread implications for utilization of social media data for epidemiologic toxicovigilance