TITLE: An Innovative Solution to Detect, Classify, and Report Illicit Online Marketing and Sales of Controlled Substances via Twitter: Project Finalist at HHS Opioid Code-a-Thon

Tim K. Mackey, MAS, PhD*1,2,3, Janani Kalyanam, MS, PhD*4, Josh Klugman4, Ella Kuzmenko4, Rashmi Gupta5

* Authors contributed equally

1 Department of Anesthesiology, University of California, San Diego School of Medicine, San Diego, CA USA
2 Division of Global Public Health, University of California, San Diego School of Medicine, Department of Medicine, San Diego, CA USA
3 Global Health Policy Institute, San Diego, CA USA
4 Department of Electrical and Computer Engineering, University of California, San Diego, San Diego, CA USA
5 IBM GBS (Watson AI and Data Platforms), Washington DC
6 U.S. Centers for Disease Control, Atlanta, GA

Word count (main text): 5,104

Tables/Figures: 3 Tables, 3 Figures

Address for Correspondence:

Timothy Ken Mackey, MAS, PhD

Global Health Policy Institute

8950 Villa La Jolla Drive, A124

La Jolla, CA 92037

Email: tmackey@ucsd.edu

Phone: (951) 491-4161

Fax: (858) 534-9168

Key Words: Non-medical use of prescription drugs; prescription opioid abuse; infoveillance; substance abuse; Twitter; social media; internet pharmacies; online pharmacies
ABSTRACT (384 words):

**Background:** On December 6-7, 2017, the U.S. Department of Health and Human Services hosted its first Code-a-Thon event aimed at leveraging technology and data-driven solutions to help combat the opioid epidemic. Authors, comprised of an interdisciplinary team from academia, the private sector, and the U.S. Centers for Disease Control and Prevention, participated in the Code-a-Thon as part of the “Prevention” track.

**Objective:** To develop and deploy a methodology using machine learning to accurately detect the marketing and sale of opioids by illicit online sellers via Twitter as part of participation at the HHS Opioid Code-a-Thon event.

**Methods:** Tweets were collected from the Twitter public API stream filtered for common prescription opioid keywords in conjunction with participation in the Code-a-Thon from November 15 – December 5, 2017. An unsupervised machine learning-based approach was developed and used during the Code-a-Thon competition (24 hours) to obtain a summary of the content of the Tweets to isolate those clusters associated with illegal online marketing and sale using a Biterm Topic Model (BTM). After isolating relevant tweets, hyperlinks associated with these tweets were reviewed to assess characteristics of illegal online sellers.

**Results:** We collected and analyzed 213K tweets over the course of the Code-a-Thon containing keywords codeine, percocet, vicodin, oxycontin, oxycodone, fentanyl and hydrocodone. Using BTM, 692 (0.3%) tweets were identified as being associated with illegal online marketing and sale of prescription opioids. After removing duplicates and dead links, we identified 34 unique “live” tweets with 15 (44.1%) directing consumers to illicit online pharmacies, 11 (32.4%) linked to individual drug sellers, and 7 (20.6%) used by marketing affiliates. In addition to offering the “no prescription” sale of opioids, many of these vendors also sold other controlled substances and illicit drugs.

**Conclusions:** Results of this study confirm prior studies that have identified social media platforms, including Twitter, as a potential conduit for supply and sale of illicit opioids. To
translate these results into action, authors also developed a prototype wireframe for the purposes of detecting, classifying, and reporting illicit online pharmacy tweets selling controlled substances illegally to the U.S. Food and Drug Administration and U.S. Drug Enforcement Agency. Further development of solutions based on these methods has the potential to proactively alert regulators and law enforcement agencies of illegal opioid sales, while also making the online environment safer for the public.
INTRODUCTION:

It is estimated that 90 Americans die daily by overdosing on opioids, a staggering figure highlighting the human toll of this public health crisis that continues to escalate [1]. Since the year 2000, an estimated 300,000 lives have been claimed by the opioid epidemic, which has expanded beyond non-medical use of prescription opioids into transition of use to heroin addiction and deaths occurring from illicitly manufactured synthetic opioids (such as fentanyls and their analogues [2-5].) Additionally, the U.S. Centers for Disease Control and Prevention (CDC) estimates that the annual economic losses from this crisis equate to $78.5 billion due to the costs of healthcare, addiction treatment, the criminal justice system, and lost productivity [6,7]. Rising death tolls and growing economic burden (with CDC reporting a quadrupling of deaths attributable to prescription opioids since 1999) have prompted certain state jurisdictions and recently President Donald Trump to declare the opioid crisis a public health emergency [8,9].

Responses to tackle the opioid epidemic have occurred at both the state and Federal level, largely focused on actions aimed at reducing inappropriate prescribing, expanding opioid treatment and prevention programs (including access to naloxone), establishing prescription drug monitoring programs, preventing drug diversion, and even the use of litigation against pharmaceutical companies [5,8,10-13]. These approaches largely fit into the five major priority areas outlined by the U.S. Department of Health and Human Services (HHS) to combat the opioid crisis, which include improving access to treatment and recovery services, promoting use of overdose-reversing drugs, strengthening public health surveillance, enhancing research on pain and addiction, and supporting better practices for pain management. These strategic goals are being carried out through investments in science, training, various mitigation strategies, community-based activities, efforts to change prescribing and management practices, and policymaking [13-15].
Seeking to further catalyze efforts around HHS’ five-part opioid strategy through technology and innovation, on December 6-7\textsuperscript{th} 2017, the agency hosted an Opioid Code-a-Thon event that brought together over 300 participants to develop data-driven solutions to combat the opioid epidemic in three challenge tracks (treatment, usage, and prevention tracks, see Table 1) [16]. The Code-a-Thon, the first of its kind for HHS, also involved partnership with several data science providers, organizations, and platform sponsors, including Socrata, Tableau, IEEE Standards Association, and Google. The event provided participants access to a hosted data portal that included de-identified datasets from HHS, federal, state, and local governments, and the private industry (e.g. data sets such as CDC WONDER – Multiple Causes of Death, Medicare Part D Prescribing Data – Centers for Medicare & Medicaid Services, National Survey on Drug Use and Health – Substance Abuse and Mental Health Services Administration, and Medical Expenditure Panel Survey – Agency for Healthcare Research and Quality), which were used by participants to build data visualizations, interpret data in new ways, build analysis tools, and propose broader solutions [16]. In total, over 50 teams competed and pitched their ideas through two rounds of judging, with nine selected for a final round of presentations. In the end, a total of three winners were awarded prizes of $10,000 each to further develop and implement their solutions [16].

Table 1: Summary of HHS Opioid Code-a-Thon Challenge Tracks

<table>
<thead>
<tr>
<th>Track Name</th>
<th>Description</th>
<th>Winning Team</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prevention Track</strong></td>
<td>Solutions used to better predict and analyze the supply and movement of legal and illicit opioids</td>
<td>Visionist Inc., solution to assess unmet need for takeback programs at pharmacies</td>
</tr>
<tr>
<td><strong>Treatment Track</strong></td>
<td>Solutions to improve access to effective treatment and recovery services</td>
<td>Origigami Innovations, developed model for real-time tracking of overdoses</td>
</tr>
<tr>
<td><strong>Usage Track</strong></td>
<td>Solutions to help identify at-risk populations and their underlying risk characteristics of opioid misuse or abuse</td>
<td>Visionist Inc., solution to assess unmet need for takeback programs at pharmacies</td>
</tr>
</tbody>
</table>
The co-authors of this paper, comprised of an interdisciplinary public-private team with members from UC San Diego – School of Medicine and Jacobs School of Engineering, the CDC, and IBM, participated as Team “Ryan Haight” as part of the Prevention Track (a track which asked for solutions designed to predict and analyze the supply and movement of legal and illicit opioids) and were selected as one of nine finalists (for recorded video of Code-a-Thon presentation visit the HHS YouTube video: https://www.youtube.com/watch?v=7B7jWbsktC0.)

In this article, we describe the opioid challenge we addressed, the methods and solution we used to address the challenge, the results of our analysis, and provide a discussion on our approach to move this innovation forward in an attempt to address the digital dangers of opioid abuse via social media and illicit online sellers.

Opioid Challenge: Illicit Online Marketing of Sale of Opioids Direct-to-Consumer via Twitter

We named our Code-a-Thon team after an 18-year old adolescent from San Diego, California, who overdosed and died after purchasing the prescription opioid Vicodin® from a “no prescription” online seller in February 2001 [17,18]. The death of Ryan Haight eventually led to passage of Federal legislation, the 2008 Ryan Haight Online Pharmacy Consumer Protection Act (RHA), aimed at curbing illegal diversion of controlled substances by making it a federal crime to purchase, sell, or import controlled substances online without a valid prescription as currently enforced by the U.S. Drug Enforcement Agency (DEA) [17].

However, since the enactment of RHA, Internet technologies have experienced rapid growth, with an estimated 84% and 65% of adult Americans using the Internet and social media respectfully [19]. Further, a survey conducted by the U.S. Food and Drug Administration (FDA) of adults who had made purchases online found 23% had purchased a prescription medication online [20]. Reflecting this trend of increasing Internet adoption and use of Internet for health information seeking and e-commerce, online pharmacies have also rapidly proliferated with an estimated 30,000 now in existence [21-23]. Of these tens-of-thousands of cyber-pharmacies, the Internet security firm LegitScript estimates that approximately 9% sell controlled substances,
with a separate report by the National Association of Boards of Pharmacy (NABP), which reviewed more than 11,000 online pharmacy websites, estimating that 13% illegally dispensed controlled substances [21,24,25]. Results from these studies also confirm findings from U.S. government investigations, including a 2004 study by the U.S. Government Accountability Office (GAO) where investigators were able to purchase OxyContin®, Percocet®, and Vicodin® from no prescription online pharmacies and a recent bipartisan report detailing how Senate investigators were able to easily purchase fentanyl online and have product shipped through the mail [26-28].

Supplementing reports from the FDA, NABP, the GAO, and Senate investigations, a number of published research studies have also found that prescription opioids are readily marketed and sold online, both pre and post-RHA [29-34]. This includes recent studies establishing a link between illicit online pharmacies that use social media channels (primarily Twitter) to market and directly sell opioids and other controlled substances (including fentanyls) [33-36]. Growing social media popularity among consumers or the fact that social media platforms are generally less regulated compared to other parts of the web, such as indexed search engine results, may be driving these trends (e.g. in 2011 Google was fined $500 million by the U.S. Department of Justice for knowingly allowing illegal ads for fraudulent pharmacies, including those selling controlled substances, and has since instituted policies for prevention [37].) The connection between illicit access and social media is particularly concerning given that many social media platforms are popular among young adults (a recent 2017 study reported that 30.85% of young adults used Twitter), a population at specific and increasing risk to prescription opioid addiction [38,39].

Hence, recognizing the need for innovative solutions leveraging advances in infoveillance, big data, machine learning, and web forensic analysis, our team developed a method to detect marketing and sale of controlled substances via twitter by online sellers. The
main component of our solution was use and validation of an unsupervised machine learning
algorithm that detected illicit online opioid seller tweets. We also created a prototype wireframe
of a web application to detect, classify, and report results for potential use by stakeholders such
as the DEA, FDA, pharmaceutical manufacturers, and consumer patient safety groups.

METHODS:
The analysis for this study was conducted in two distinct phases including: (1) Code-a-Thon
challenge assessment; and (2) “big data” analysis using machine learning of a Twitter dataset.
We describe the design of our prototype web application solution with a wireframe demo later in
the discussion section. The first phase involved coordinating with one of the co-authors of this
manuscript who is currently a CDC Entrepreneur in Residence to scope out an appropriate
challenge problem that fit the specific objectives and appropriate track of the Code-a-Thon (i.e.
addressing an under recognized threat in illegal opioid supply and access), identify the relevant
data sets needed to address the challenge (i.e. collecting a Twitter dataset associated with
opioids pre-competition), and organize team registration and logistics associated with Code-a-
Thon participation. The second phase comprised of data analysis conducted during the Code-a-
Thon as described below. As this study involved the collection and analysis of existing
publicly-available data, it did not require Institutional Review Board approval, nor was that
required for participation in the Code-a-Thon.

Data Collection and Analysis:
After determining the challenge problem in partnership with CDC, we proceeded to collect
messages (i.e. tweets) published on Twitter over a period of approximately 20 days from
November 15 and ending on December 5, 2017 (the day prior to the start of the Code-a-Thon.)
The public streaming API available from Twitter was used with certain preselected keywords
that were a combination of International Nonproprietary Names (INN) and brand names of
commonly abused opioids. Our final keyword list contained the terms codeine, fentanyl, hydrocodone, oxycodone, Oxycontin®, Percocet®, and Vicodin®.

Upon commencement of the competition, we used a machine learning protocol to isolate word groupings associated with tweets that mentioned marketing and purported sale of prescription opioid drugs as has been carried out in prior studies by first and second author [33,34,36]. To identify relevant tweets related to our challenge problem (i.e. “signal” data) in large volumes of Twitter data (in the hundreds of thousands) versus non-relevant data (i.e. “noise” that contains opioid-related keywords but do not relate to online promotion or sale), the application of machine learning is critical in order to achieve scale and comprehensive analysis in a reasonable time frame compared to approaches solely using manual annotation by human coders.

Specifically, unsupervised methods like topic models prove to be useful in obtaining a summary of the underlying themes present in large text corpora. We used a model called the Biterm Topic Model (BTM) designed to detect themes and patterns in corpora of short-texts (such as tweets), which we have previously used to examine prescription drug abuse behavior and online marketing and access (see Figure 1 for summary of this methodology) [34,40]. BTM (in its “learning” phase) first detected a preconfigured number of themes from the filtered dataset of tweets containing prescription opioid keywords. This produced a set of topics (or word groupings) that are thematic summaries of the contents of the entire set of tweets. The resulting word groupings are used to inform the next steps in the methodology – which are either identification of themes/signal associated with illegal marketing and sales of prescription opioids or elimination of noise (discarding tweets that are deemed as “noisy” to isolate signal tweets) and re-applying BTM on the smaller subset that has been filtered for noise.

*Figure 1: Summary of Study Methodology*
Once the learning phase of BTM produced a set of relevant themes (or word groupings), they were coded using human annotation to manually identify “signal” tweets/data clearly associated with prescription opioid marketing, distribution and/or sale. For example, a theme with a combination of words including “[prescription opioid drug name]”, “buy”, “cheap”, “price”, “discount” (all adjectives and “selling arguments” identified as used by online sellers) were then
extracted for further analysis to identify specific characteristics of the seller/marketplace [41].

The tweets that were highly correlated with the word-groupings that contained words like “buy”, “cheap”, “price” and “discount” were obtained using what is called the “inference” phase of BTM. Within this subset of tweets, any false positives were first manually eliminated by two of our co-authors (TM and EK) by discarding those tweets whose content did not have a clear indication that the tweet was about the sale or promotion of a prescription opioid.

For example, since we narrowly focus on tweets purportedly offering online sale of opioids, only those tweets with hyperlinks contained in the message of the tweet or other contact information were considered. This produced a narrower group of tweets that contained hyperlinks to external websites or contract information to directly purchase controlled substances for prospective consumers. Hyperlinks were further manually coded to determine if the link was still active (i.e. still redirecting to an external website versus a “dead” link which failed to redirect to a working website or produced an error code), whether it marketed or purported to directly sell an opioid product, and classified according to the nature and type of seller (e.g. illicit online pharmacy, individual sale, and marketing affiliate) [23]. TM and EK coded for false positive tweets and the characteristics of links/websites independently and achieved a high Inter-coder reliability for results (k=0.94). Discrepancies were resolved through re-evaluation and consensus.

For all “signal tweets” that were specifically categorized as illegal online pharmacies, we also cross-referenced the URLs of these websites with LegitScript's external database that includes a legal classification. LegitScript legal classification is based on its own assessment of whether the website is: (a) “rogue”: vendor engaged in illegal, unsafe, or misleading activity; (b) “unapproved”: vendor with a problem of regulatory compliance or risk in one or more jurisdictions; (c) “unverified”: not subject to LegitScript review or monitoring; or (d) “legitimate”: passed LegitScript certification criteria.[33,42] LegitScript classification queries offer another layer of verification regarding an online pharmacy's legal status and can help to confirm that the
sites present high risk for consumers. We also reviewed WHOIS data to determine the IP address and registered owner location for links classified as online pharmacies.

RESULTS:

We collected data for approximately 20 days resulting in a total of 213,041 tweets used for our analysis. The number of preconfigured themes was set to 100 during our BTM learning phase. During our first run of BTM on the entire dataset, we observed a significant number of the word groupings were related to fentanyl, all of which were news-related themes (see examples in Table 2). This can be attributed to the high volume of news events surrounding illicit fentanyl prior to the time of the competition during the data collection phase. News essentially adds noise to the dataset (i.e. it does not contain conversations related to online sales) – hence we discarded fentanyl from the data and were left with ~117K tweets. Applying BTM on this smaller dataset gave clear signals of certain selling argument word groupings with words like “buy”, “online”, “cheap”, “free”, “shipping” and the name of a prescription opioid. This demonstrates that fentanyl-related tweets added noise to the dataset by suppressing signal data (see Examples in Table 3.)

Table 2: Some example themes that were obtained from the first round of BTM

<table>
<thead>
<tr>
<th>Example theme – 1</th>
<th>fentanyl, deaths, dangers, nations, drugs, china, learn, surge, deals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example theme – 2</td>
<td>fentanyl, man, police, charges, faces, arrest, brockton, bust, Calgary</td>
</tr>
<tr>
<td>Example theme – 3</td>
<td>fentanyl, overdose, suspected, teen, life, drugs, dead, death, vancouver</td>
</tr>
</tbody>
</table>

Table 3: Some example themes that were obtained from the data after fentanyl-related tweets were removed (bold denotes relevant “signal” theme)

<table>
<thead>
<tr>
<th>Example theme – 1</th>
<th>cannabis, legalization, respect,</th>
</tr>
</thead>
</table>

12
<table>
<thead>
<tr>
<th>Example theme – 2</th>
<th>drug, war, opposition, benches, friends, conservative, political, dead, props</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example theme – 3</td>
<td>buy, online, pill, free, cheap, find, shipping, generic</td>
</tr>
</tbody>
</table>

**Signal Twitter Data Analysis:**

Using the inference phase of BTM, we retrieved tweets that were most correlated and second most correlated to selling argument word groupings (0.3%, n= 692). Of the total 692 tweets retrieved, 476 (68.79%) contained hyperlinks. After manually coding tweets associated with these hyperlinks, we removed all duplicates (e.g. retweets, identical tweets) and also any tweets with dead links. This resulted in a total of 34 live tweets with hyperlinks (followed by an estimated 1,800 users during the duration of the data collection period) that belonged to one of the following three online seller categories: (1) an online pharmacy (defined as a website that purports to be an online pharmacy storefront, operates an e-commerce shopping cart where products can be “checked out” and paid for directly by a consumer); (2) an individual seller (defined as a seller representing themselves as a single individual or entity offering the sale of prescription opioids via email, phone, or other direct contact solicitation); and (3) a marketing affiliates (defined as a website that hosts links to other websites that directly sell controlled substances) (see Figure 2). Of the live links coded at the Code-a-Thon, 15 (44.11%) were online pharmacies, 11 (32.35%) were individual sellers, 7 (20.59%) were marketing affiliates, and 1 was purportedly from the Twitter handle of the darknet site AlphaBay (operated via the Tor network) linking to a reddit community page. The authenticity of the final result is questionable given that AlphaBay was shut down by law enforcement officials in July 2017.
The first category of 15 coded tweets consisted of 10 distinct live hyperlinks to illicit online pharmacies (see select examples in Figure 2, Example A). All of these websites sold opioids with “no prescription” in combination with other drug products including other controlled
substances (e.g. barbiturates, Xanax, Ketamine, fentanyl patch, codeine cough syrup), non-opioid prescription drugs therapeutic classes (e.g. injectable steroids, antidepressants, weight loss drugs including Sibutramine which has been removed from several markets, hormones, contraception and erectile dysfunction drugs), recreational drugs (e.g. cannabis/marijuana), and some sites that also sold illicit drugs (e.g. bath salts, MDMA, cocaine, heroin, methamphetamines.) One site purportedly only accepted payment via cryptocurrencies bitcoin and ether (the cryptocurrency for the blockchain platform Ethereum.) When cross-referencing for LegitScript status, six (40%) were identified as “rogue” with the remaining having no information available (i.e. likely not detected or included in LegitScript’s database.) Some online pharmacy tweets also used interesting selling arguments, including advertising “Black Friday” sales in reference to the day after Thanksgiving holiday in the United States, when retail discount sales are heavily marketed. Interestingly, when examining purported geographic location of IP addresses and website registered owners, the majority reported addresses in the United States.

The second category of 11 tweets appeared to originate from individual sellers using Twitter accounts that advertised the direct sale of opioids using hashtags (e.g. #buy, #sell, #buypainmeds, #drugsforsale, #opioids, #painmeds, #controlled substance names] and generally included contact information for the seller including an email address (in most cases a Google gmail account) and a phone number to call, text or contact via the messaging application WhatsApp (phone numbers were mostly U.S. 10-digit phone numbers, though phone numbers for UK country code +44 were also observed) (see Figure 2, Example B.) These tweets generally included pictures of prescription opioids and other drugs shown in an individual's hand or displayed next to a pill bottle. Some twitter individual accounts also linked to an external webpage (including wordpress domains), blogs, or online classified ad services. The sellers generally included in their tweets or external webpages a list of all drugs they
offered to sell, often comprised of a mix of controlled substances, other prescription drugs, and illicit drugs.

The third category consisted of 7 tweets that included links to marketing affiliate websites and networks that were hosted on their own distinct domains or used other blog sites (e.g. blogspot.) These sites did not directly sell opioids but included information on how to purchase illegally from other sites and hosted hyperlinks on their webpages that redirected traffic directly to an online pharmacy engaged in that activity (see Figure 2, Example C.) One of these sites included a FAQ page that stated: “No, we are not sell any pills or medication. We are only provide medical information” and included several web banners/advertisements to illicit online sellers, including one that used a fake banner claiming it was “FDA approved.”

Limitations:
There are limitations associated with the results generated from Twitter data as were reported in the Code-a-Thon and as part of this study. Specifically, we first cleaned the dataset prior to analysis to exclude non-English language tweets, which may further limit the generalizability of our sample to tweets containing opioid keywords generated on Twitter. For signal data that either included a direct hyperlink to an online pharmacy or marketing affiliates that redirected to live websites, content analysis was reviewed at a specific point of time after the tweets were collected and analyzed using BTM. Though tweets were coded right after the data collection was completed, it is possible that the content residing on hyperlinked content and/or the online pharmacy’s website/domain may have changed from the exact date of data collection and website content coding as websites often change content. Additionally, our exclusion of “fentanyl” from our analysis may have removed content related to the illicit online sale of fentanyls via online pharmacies and individual sellers but was necessary to further refine our results in order to detect “signal” tweets with clear selling argument word topic groups. A recent study by first and second author focused on detection of illicit online sales of fentanyls and can
inform analysis of this sub-set of data [36]. The validity of WHOIS geographic data for online pharmacies reviewed is also unclear. Though many of online pharmacies listed an IP address or registered owner address in the USA, their actual server location and/or physical business location/registration could be falsely entered or masked by a privacy Internet service provider company. Finally, we note that the RHA and DEA explicitly have rules and regulations that make it illegal to purchase controlled substances online and have no clarifying guidance or exemptions in relation to conducting test purchases for research purposes. Since researchers reside in U.S. jurisdiction, we were unable to actually purchase controlled substances from online pharmacies and test them for authenticity.

DISCUSSION:

Similar to findings in our prior studies that have examined the use of Twitter by illicit online sellers to market and sell prescription opioids, the overall volume of tweets directly engaged in this illegal activity was relatively low compared to the entire corpus of tweets collected [33,34,36,40]. This indicates that though this activity presents clear risk to the consumer and could contribute to diversion of opioids and convenient access via online pharmacies or individual sellers, the occurrence of these tweets are not widespread compared to other more commonly occurring conversations including news reports about the opioid crisis, tweets about opioid abuse behavior self-reported by users, and other content that includes opioid keywords but which we would classify as “noise” per the aims of this study. However, some interesting results emanating from this study also provide us with key information on emerging trends regarding social media-enabled opioid abuse and methodological considerations when attempting to identify this type of content.

First, our prior studies did not detect our second category of tweets characterized as individuals who market and sell prescription opioids via open solicitations and offers via direct contact with users. These “digital drug dealers” openly tweet that they can sell prescription
opioids and other illicit substances directly to the public and that they can be contacted through a simple email or phone number. Oftentimes, they purport to validate availability of drugs offered by also including a picture of their products included in the tweet and also use hashtags (#) to curate and target their marketing messages. These digital sellers sometimes represented themselves as individuals, and in other cases as a company (primarily a name that represented itself as a pharmacy or pharmaceutical company.) These results are alarming as they represent a potential new strategy where traditional “street” dealers may simply use Twitter in an effort to broadly market and extend their services to a wider and more diverse customer base that they would otherwise not have access to. The range of followers for these accounts varied, with observed ranges as few as 24 followers to as high as 989 followers, though generally these accounts were observed to have less that 100 followers. At the time of this writing, these accounts were still active on Twitter.

Second, compared to previous studies where the data collection and analysis process was separated by more than 1 year, in this study we immediately analyzed collected data as it was part of our Code-a-Thon competition [34,36,40]. This resulted in far fewer dead links where URLs were no longer active or website landing pages were no longer being used by owners. Hence, our study demonstrates that a near real-time infoveillance approach, which collects data over a short duration (e.g. less than 30 days) and conducts data analysis immediately after collection, can result in better detection of illegal activities of online pharmacies and sellers. This approach could inure benefits to regulators and law enforcement officials, as Twitter accounts, illegal websites, and marketing affiliates could have their content taken down more quickly in order to mitigate potential exposure to consumers and accompanying patient safety risk. This would allow for proactive detection, particularly important given that illicit online pharmacies do not maintain a consistent presence on the Internet and often change URLs frequently [23,43,44]. In fact, the first author of this paper has shared preliminary results of this study directly to the FDA Center for Drug Evaluation and Research, Office of Compliance for

18
further action, as part of a separate invited visit to discuss and present on drug safety and surveillance at FDA.

**Web Application Wireframe Prototype Solution:**

In parallel with our Twitter data collection and analysis and to better demonstrate the potential real-world application of our BTM methodology to detect, classify, and report online pharmacies marketing the sale of controlled substances, we developed a wire frame solution using the prototyping tool for web and mobile apps, Justinmind ([https://www.justinmind.com](https://www.justinmind.com)). We based our conceptual model of our webapp solution on simplicity, streamlined navigation, and user interaction with a task focus. The webapp wireframe was designed to implement three primary functions: (1) data collection; (2) detection of tweets marketing the sale of opioid online using BTM assisted by human interpretation; (3) html classification of websites to determine if they are illicit online pharmacies; and (4) automating a script to report to the FDA and DEA about detection of illicit online pharmacy for further regulatory action (see **Figure 3** with screenshots of the justinmind wireframe solution developed and as presented during the Code-a-Thon.)

**Figure 3: Design Elements and Screenshots of Prototype Wireframe Solution**

- A streamlined user-friendly design to collect, detect, report and respond to the public health challenge of illicit opioid sales online
The first component of this solution included a webpage with a query function that included a date range for data collection and the ability to enter prescription opioid keywords for filtering of tweets from the public streaming API. The second page exports tweets filtered for selected keywords and then specifically outputs “signal” tweets highly correlated with selling arguments associated with online pharmacies. This step would also include an output of Twitter handles/accounts that have interacted with this high-risk content (e.g. Twitter users that are followers, have retweeted, or favorited identified “signal” tweets) for possible countermarketing, health communication regarding potential risk, and also generate data for potential social network analysis. The third page would then output screenshots of hyperlinks associated with “signal tweets” for human inspection, while also reporting the IP geographic location of the website (using WHOIS data) and the LegitScript legal classification of the URL if available (LegitScript operates a fee-based API for their information.) This page would also include a “report” button under each website screen capture. User selection of violating websites would generate a final page detailing what URLs had been reported to the FDA and DEA using an automated script that would fill in the necessary information fields on the FDA’s “Reporting Unlawful Sales of Medical Products on the Internet” webpage (https://www.fda.gov/Safety/ReportaProblem/ucm059315.htm) and the DEA’s “Report Submission Form for Suspected Unlawful Sales of Pharmaceutical Drugs on the Internet (https://apps.deadiversion.usdoj.gov/webforms/jsp/umpire/umpireForm.jsp). This prototype was presented during our finalist presentation at the Code-a-Thon, but due to time considerations, was only a “click-thru” but not fully functioning demo. However, illicit online pharmacy website results were reported to both the DEA and FDA by manually filling out the online reporting tools.

CONCLUSION:
This study and our participation in the HHS Code-a-Thon validates an important methodology to detect illicit online sales of opioids that are marketed via Twitter. Importantly, the machine learning approach, which represents the core technology for our proposed solution, is scalable and can be done relatively quickly following infoveillance-related data collection. This allows us to more rapidly detect illicit online sellers and classify their marketing characteristics before they remove their web presence. In fact, the vast majority of online sellers detected in this study remain active on social media and the web at the time of this writing. Though the machine learning component of this study is relatively mature, with validation of application to this use case now published in four separate studies, the translation of this approach to an easy-to-use, accessible, and largely automated solution is still at a very early-stage [33,34,36]. Though our prototype wireframe demonstrates the potential extension of the BTM machine learning algorithm into a web application that could be used by stakeholders such as the FDA, DEA, and pharmaceutical manufacturers, it does not have the functionality or integration of different data sources to be considered a minimally viable product (MVP). To take this next step, funding with the primary aim of translating this research into MVP phase and eventual production and scale-up is needed (such as our recent application to the NIH NIDA 2017 “Start a SUD Startup” Challenge, which provides small awards for startups related to substance abuse disorders and that could potentially lead to a successful NIDA Small Business Innovation Research grant.) Additionally, we would need to automate data collection and backend analysis of Twitter data with integration via a web application, while also developing solutions using natural processing language to automate classification of hyperlinks suspected as engaged in the sale of prescription opioids, techniques that have been explored in prior studies [22,45,46]. Finally, automated scripts that generate information needed for standardized reporting of results to the FDA and DEA via their online web forms would also need to be developed. Despite these challenges, results from this study are useful and can inform regulators, law enforcement, public health officials, and the public about current and changing trends regarding supply, access, and
distribution of illicit opioids. Technology, such as the big data and machine learning approaches used in this study, will be critical components of any strategy to combat the opioid epidemic, an approach that HHS through its Code-a-Thon has catalyzed.
Multimedia Appendix of Supplementary Files:

Recorded video of Code-a-Thon finalist presentations HHS YouTube video:
https://www.youtube.com/watch?v=7B7jWbskIC0

Recorded video discussing Code-a-Thon solution submitted as part of NIH NIDA 2017 “Start a SUD Startup” Challenge grant:
https://www.youtube.com/watch?v=XZaOggHVn3Q
Author Disclosures:

**Role of Funding Sources:** TM received funding for a pilot grant unrelated to this study from the Alliance for Safe Online Pharmacies (ASOP), a 501(c)(4) social welfare organization engaged on the issue of illicit online pharmacies. TM also received travel support from the FDA for an invited accredited educational activity on drug safety and surveillance, which included discussion regarding the results of this study.

**Contributors:** JK conducted the data collection for the study. All authors contributed to the formulation, analysis, drafting, completion, and approval of the final manuscript.

**Conflict of Interest Statement:** TM is a non-compensated member of the ASOP academic advisory panel of ASOP. ASOP had no role or input in the study. There was no involvement of anyone other than the authors in the conception, design, collection, planning, conduct, analysis, interpretation, writing, and discussion to submit this work. Authors report no other financial relationships with any organizations that might have an interest in the submitted work.

**Acknowledgment:** Authors would like to thank ASOP for their support of prior work associated with developing a methodology to detect and characterize illicit online pharmacies using social media platforms. Authors also would like to thank the Robert Woods Johnson Foundation New Connections Program for financial support associated with publication fees (article processing fees) to publish this study.
References:


38. Abbasi J. Emergency Department Opioid Misuse Diagnoses Increasing in Adolescents and Young Adults. JAMA American Medical Association; 2017 Dec 26;318(24):2416–2417.


**Abbreviations:**

BTM: Biterm Topic Model

CDC: U.S. Centers for Disease Control and Prevention

DEA: U.S. Drug Enforcement Agency

FDA: U.S. Food and Drug Administration

GAO: U.S. Government Accountability Office

HHS: U.S. Department of Health and Human Services

MVP: Minimally viable product

NABP: National Association of Boards of Pharmacy

RHA: 2008 Ryan Haight Online Pharmacy Consumer Protection Act