STEM Fellowship Big Data Challenge 2018-2019

An Ontario-Based Prediction Platform for Opioid-Related Emergency Department Admissions

February 1, 2019

Abstract

The lack of real-time data on drug overdose rates has made it difficult for governments to plan for and establish the infrastructure necessary to combat cases of substance abuse. The age of digitization, however, has created an environment where the internet serves as the powerhouse of information. Thanks to its prevalence, the internet ecosystem can be used as a cost-effective and accurate data source for observing and predicting behavioural patterns of public health phenomenon. By leveraging data from Google Trends, we aimed to examine whether search interest in opioid-related keywords can serve as as a predictor of the number of opioid-related emergency department visits in Ontario.

A polynomial regression model was utilized to evaluate the relationship between a opioid-related keyword's interest in Ontario and the province's count of opioid-related ED visits in the following month. The final model contained 10 keywords.

We found that the keyword "Opioid Ontario" was the most accurate predictor of ED visits, as its best fitting model predicted 99 percent of the variance in opioid-related ED visits. Using the methods outlined, Google Trends, and internet search data in general, can be leveraged as a cost-effective source for predicting opioid-related emergency department visits and optimizing government budget for opioid related programs.

Keywords

Drug Overdose, Emergency Department Visits, Google Search Interests, Ontario

1 Introduction

Over 2000 people nationwide died due to opioid overdose in the first half of 2018 [1]. Indubitably, overdose cases resulting from illegal drug use are rising and prominent issues within our national environment. Due to the national increase in overdose cases, the government has decided to make extra funding for treatment facilities in order to lower the number of overdose cases. The urgency of this issue can be evidenced by the province of Ontario's pledge of 222 million dollars over 2.5 years to combat its opioid crisis by aiming to improve treatment measures [2]. The difficulties involved in collecting overdose data further exacerbates the crisis; one study revealed that opioid-related deaths are undercounted by as much as 20 to 35 percent, making it difficult for governments to predict spending on drug overdose cases [3]. Evidently, the need for new methods for monitoring prescription and non-prescription opioid use is vital.

Fortunately, technological advancements in the 21st century have created a social environment that has made access to information and communication far easier.

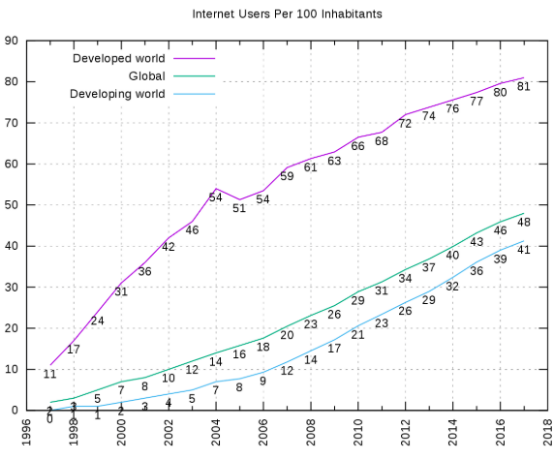

Figure 1: Internet Users Per 100 inhabitants from 1996 to 2018

This phenomena, coupled with the fact that Internet users as a percentage of the population has risen 20 percent from 2008 to 2018 (shown in figure above), has not only changed the social environment, but also has become the new social landscape in both developed and developing world. Therefore, internet search data, such as those provided by Google Trends, can be used as a data source for monitoring and predicting public health outcomes, including influenza, skin cancer, and depression.

In fact, internet data are important sources of abundant information regarding diseases. A number of case studies found an association between internet searches and outbreaks of infectious diseases, including HIV [4][5][6]. However, no known research has examined the relationship between Google Trends data and the recent opioid outbreak in Canada and Ontario.

Therefore, based on the hypothesis that opioid-related emergency department admissions might be preceded by people searching online for information relevant to opioids, we sought to identify whether Google Trends data could be used to predict the following month's emergency department opioid admissions in the province of Ontario. Because both prescription and non-prescription opioid use has been associated with future heroin use [7], we compiled a list of search terms for common descriptors of prescription and non-prescription related opioids, and evaluated whether a model built upon these search terms can predict the following month's emergency department heroin-related visits.

If identified as a strong predictor of the count of ED visits, Google Trends data can serve as a valuable asset to health organizations and localized neighborhoods. Firstly, the data can allow health organizations to better allocate resources. Health Quality Ontario has commented that the province's emergency department could be strained beyond its capacity to provide quality care to all its patients. The number of annual visits to Ontario's emergency departments increased 13.4% – more than double the 6.2% increase in the province's population [8]. Furthermore, the World Health Organization found that drug markets and associated violence in a neighborhood cause neighborhood deprivation, such as a lack of employment opportunities and vacant housing [9]. Evidently, the effects of drug use are localized. As drug usage is proportional to the amount ED visits, constructing a model to predict ED visits will allow municipal governments to better monitor drug usage and thus mitigate the effects of drug usage in specific locations.

2 Materials and Methods

To obtain opioid-related keywords, a list of common descriptors for prescription and non-prescription opioids from the Drug Enforcement Agency and Government Canada's controlled drug list was identified. To increase the variability and searchability of the keywords, each keyword was coined with the words "Overdose" and "Ontario" to obtain new keywords. Due to the large number of opioid-related terms in this list, the list was refined by manually entering each opioid-related descriptor into the Google Trends search to determine the descriptors that have high search volume in Ontario. Based on this screening, 10 opioid-related keyword were identified for use in the final model: Brown Sugar, Hydromorphone, Methadone, Opioid, Opioid Overdose, Fentanyl, Fentanyl Overdose, Opioid Ontario, Codeine, Pregabalin.

Datasets obtained include the count of opioid-related emergency department visits per 10,000,000 population [10] and the Google Trends interest of the 10 keywords in Ontario from January 2008 to March 2018. Google Trends computes the relative interest of a keyword by taking into account the total searches of the keyword in the geographical area and time range given. A value of 100 in search interest is the largest numerical popularity value for a given query. A value of 50 means that the query is half as popular, while a value of 0 represents negligible interest for the given query.

The CSV files for Google Trends Interest for the 10 keywords and the count of ED visits per 10,000,000 population were imported into Python and merged by month, whereby the Google Trends interest and the count of ED visits in Ontario over time can be formalized as the following:

Given a dataset $D_N = (X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$, where X_i is an element from the dataset of Google Trends interest after N months and Y_i is the set of ED visit after N month.

To match the scale of Google Trends datasets, the ED visit dataset was normalized. This normalized dataset can be formalized as the following:

Given the original dataset of ED visits, $G_N = \{X_1, X_2, \ldots, X_N\}$, let the maximum element of the data set be X_j, and the minimum element of the dataset be X_i, the normalized dataset would be: $G_N = \{(\mathrm{X}_1 - X_i)/(X_j - X_i), (X_2 - X_j)/(X_j - X_i), \ldots, (X_N - X_j)/(X_j - X_i)\}$.

To ensure the validity of the model, 85 percent of the Google Trends dataset was selected to train each keyword's best-fitting curve, while the remaining 15 percent of the dataset was set aside as the testing set. Then, Python libraries, including pandas, matplotlib, numpy, and seaborn, were used to plot and analyze the relationship between the Google Trends dataset and the normalized number of ED visits.

Firstly, polynomial regression was used to plot the best-fitting curve between each keyword's Google Trends interest and the number of months after January 2008. To determine the most accurate and unbiased degree of the curve for each keyword, best-fitting curves from degrees of 1 to 20 were tested, and the curve with the degree that yields smallest the root mean square error (RMSE) was selected.

Then, the r-squared value between the normalized count of opioid-related ED visits and each keyword's best fitting curve was determined.

To assess how well the model was trained, the testing dataset was plotted against the best-fitting curve for the most accurately predicting keyword. The r-squared value was obtained.

3 Results

Polynomial models were created for the selected 10 keywords. We found that 5 keywords shown below best-fitted the trend for ED visits. The best keyword was "Opioid Overdose", which yielded an r-squared value of 0.99.

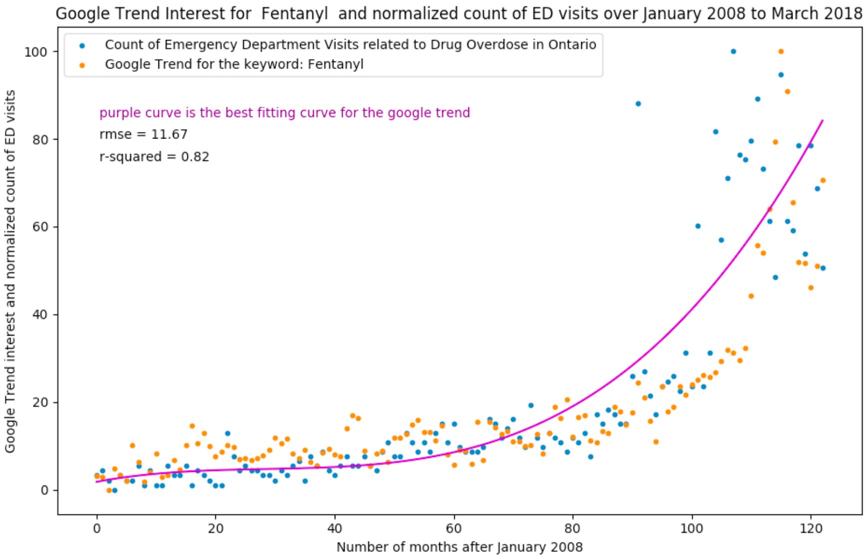

Figure 2: Google Trend Interest for Fentanyl and normalized count of ED visits from January 2008 to March 2018

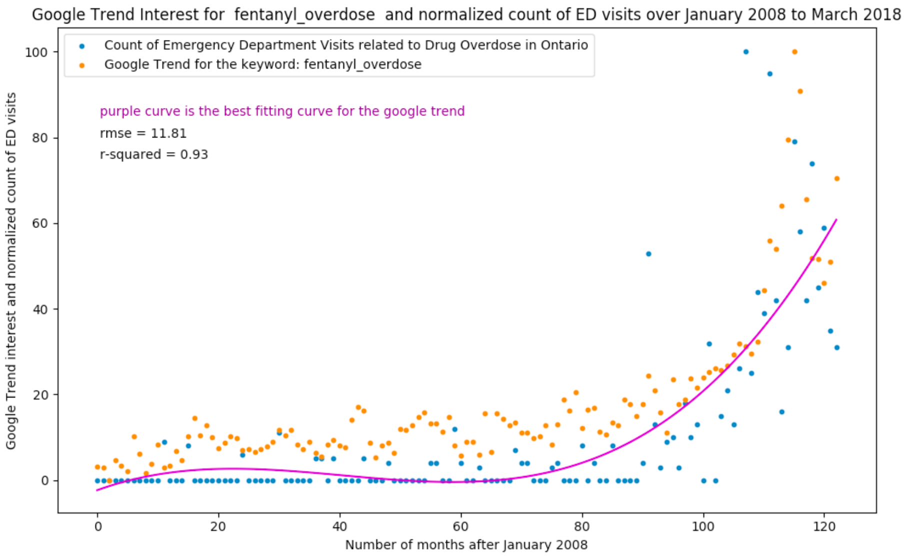

Figure 3: Google Trends Interest for Fentanyl Overdose and normalized count of ED visits from January 2008 to March 2018

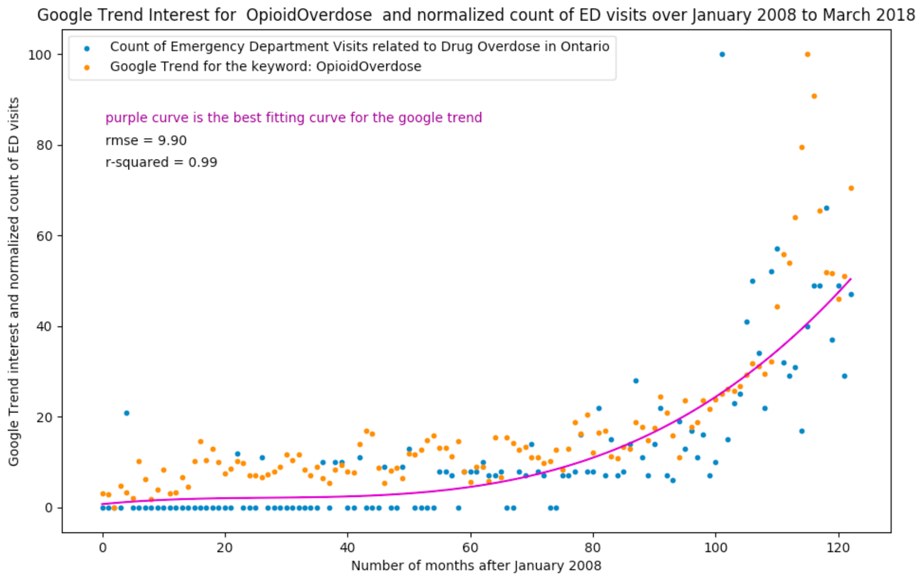

Figure 4: Google Trend Interest for Opioid Overdose and normalized count of ED visits from January 2008 to March 2018

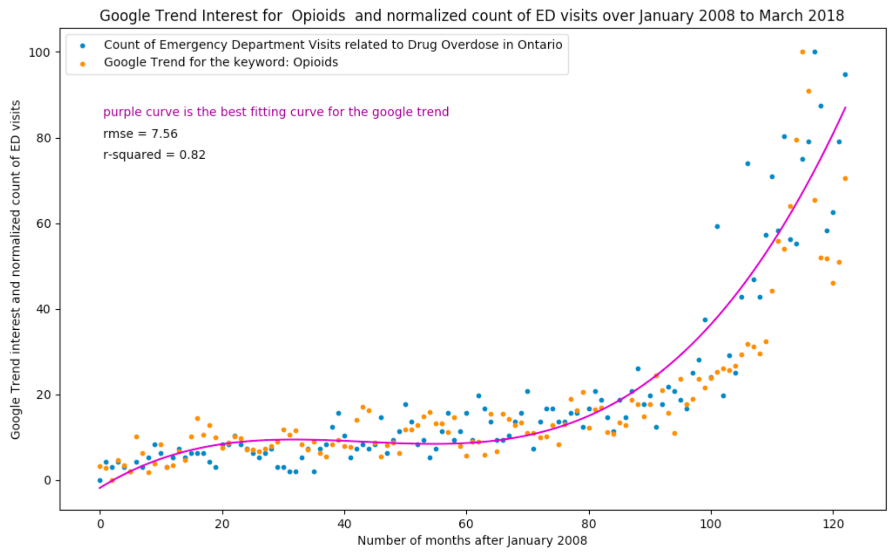

Figure 5: Google Trends Interest for Opioid and normalized count of ED visits from January 2008 to March 2018

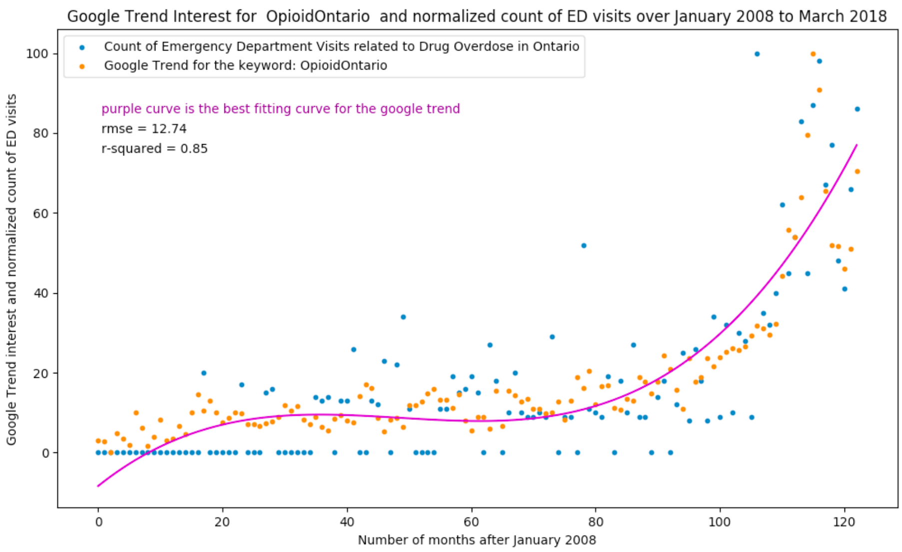

Figure 6: Google Trends Interest for Opioid Ontario and normalized count of ED visits from January 2008 to March 2018

To assess how well our program was trained, testing datasets were plotted against the best-fitting line for the keyword "Opioid Overdose".

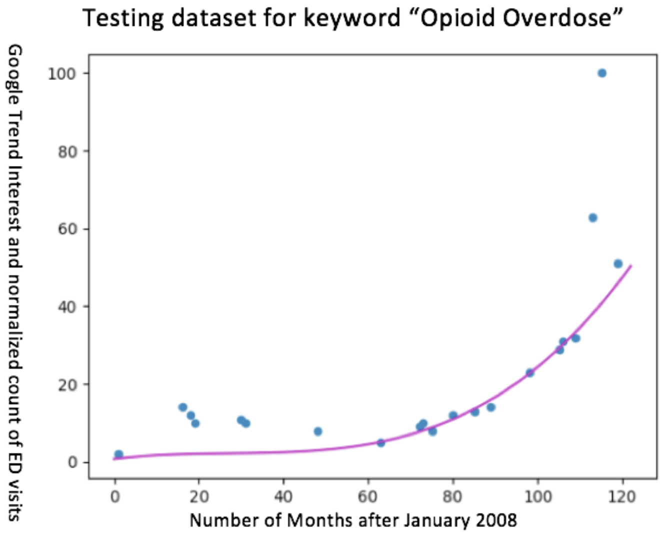

Figure 7: Testing dataset for keyword "Opioid Overdose" from January 2008 to March 2018.

The model has a r-squared value of 0.76. The purple curve represents the polynomial regression best-fit line for the keyword, and the bluedot represents the testing data points of the keyword.

The 6 keywords that did not fit the trend of ED visits as well and their respective RMSE and r-squared values are:

	Brown Sugar	Hydromorphone	Methadone	Codeine	Pregabalin
RMSE	16.45	7.74	12.73	11.65	10.79
r-squared	-0.91	-3.34	-7.11	-3.46	-1.41

Table 1: lowest RMSE and respective r-squared values for the 5 poorly-fitting keywords

4 Discussion

In Ontario, opioid-related admissions increased from about 162 per 10,000,000 population in January 2008 to about 752 per 10,000,000 population in March 2018. We found that the keyword "Opioid Overdose" best explains the variance in the number of opioid-related ED visits. With a r-squared value of 0.99, the model explained 99% of the variance in opioid-related ED visits. The keyword "Fentanyl Overdose" was the second-best predictor, explaining 93% of the variance in the number of opioid-related ED visits. The strong relationship between Google Trends results and the number of opioid-related ED admissions confirms our hypothesis of using Google Trends interests for certain keywords as a model for predicting the number of opioid-related ED visits. With a r-squared value of 0.76, the best-fitting line fitted for the training dataset of the keyword, "Opioid Overdose" provides further evidence of the accuracy of this model in predicting opioid-related admissions.

Due to the unpredictability of drug overdose cases, search interest for opioid-related keywords will rarely occur on on a linear basis, so we believed it to be more practical to use polynomial regression, instead of linear regression, to find the best-fitting curve. To find the degree of the best-fitting curve that minimizes overfitting or underfitting, we used the Root Mean Squared Error (RMSE) value, as RMSE is derived from the root of variance plus bias-squared, measuring both variance and bias to reduce the total error.

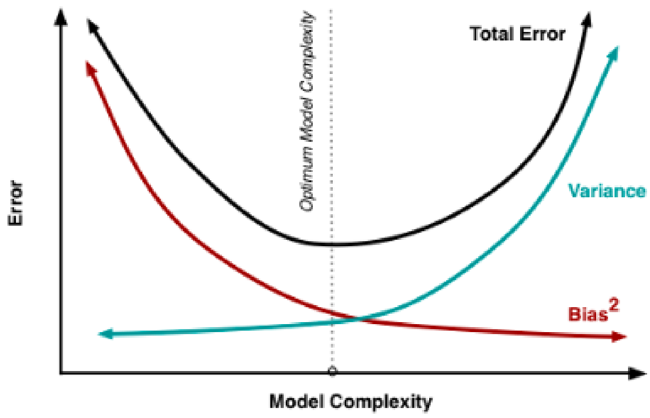

Figure 8: Optimal Model Complexity in Regressions

As shown in the figure, the lowest possible RMSE value should achieve the optimal model complexity, where there's no overfitting nor underfitting.

Furthermore, we used the r-squared value between each keyword's best-fitting curve and the count of ED visits to evaluate whether Google Trends interest can predict following month's count of ED visits. Keywords yielding r-squared values larger than 0.70 were seen as accurate predictors of the number of opioid-related ED visits.

Thanks to its real-time data, Google Trends is a relevant tool in predicting the count of opioid-related ED visits, especially in comparison to using prediction metrics such as opioid-related keywords detected in journals. In fact, data from Google Trends can become a cost-effective source for insights as they are freely and publicly available for research use and provide information on the opioid-related search from a demographic. In geographic regions or countries where no current heroin-related data exist, internet searches can be valuable in providing an estimate of opioid-related trends.

The geolocation of search behaviors that this method provides also enables targeted community-based opioid overdose prediction and drug distribution interventions. This allows high-need neighborhoods to be proactively monitored and targeted. Furthermore, analyzing opioid-related search terms could reveal changing trends in the keywords that people use for prescription and non-prescription opioids This trend can public health officials with the necessary information so that prevention efforts can be quickly assessed and rapidly refined at regional and national levels. As the effects of drug are localized [9], utilizing Internet could lead to more sustainable living environments in neighborhoods and communities.

Internet-related searches can also allow the Canadian Federal Government to accurately allocate its budgets to combating opioid use. As prediction models become more developed, historical Google Trends results for the 5 keywords in each province can provide an estimate on how much each province spends per year within the healthcare system. For example, if a model or software program were able to identify an increase in opioid-related search terms associated with future opioid-related admissions, public health officials could be better prepared for rapid education and medical interventions.

This study has several limitations. Firstly, our analysis relied solely on opioid-related keyword searches on Google, which does not provide context around people's thoughts and behaviors for the search. Secondly, this model did not take into account policy changes across the examined time period. For instance, the legalization of domestic use of marijuana in October 2018 in Ontario could have impact the Google Trends interest if the time period we chose to study contains data from October 2018.

5 Conclusion

The strong correlation between Google Trends interest for certain opioid-related keywords and the number of ED visits approach suggests that Internet search-based modelling should be explored as an avenue for predicting and managing opioid-related admissions in real-time and in selected geographical locations. Furthermore, due to the scarce amount of data on opioid-related behavior, and the dire need to combat growing overdose cases, researchers should seek ways of utilizing the Internet and social media search data as valuable tools for estimating changing opioid use trends.

The ability to connect the internet ecosystem with health-related outcomes also demonstrates the beauty of data science, a tool that can be trained and tested on real-life scenarios. This ability allows users to implement and apply data science models to specific locations, events, and times, including space travel. As space travel becomes more prevalent, the feasibility of utilizing Google Trends as a predictor of ED visits suggests that similar techniques can be applied when the human race begins to colonize another planet. Potentially, authorities can leverage the internet and online communication platforms as powerful tools for monitoring human behavior when colonizing a new planet and in supporting the planet's health infrastructure.

While the study is prone to per said limitations, it is not to provide a definitive model for predicting future drug overdoses, but to present a case study to illustrate how the internet and the social environment at large can be leveraged to address drug-related and public health problems. Therefore, this study encourages future studies to build upon this current model by incorporating symptoms and behaviors for opioid use, such as searches related to needle sharing, and through integrating more contextual data from other popular search engines and social media platforms.

6 Acknowledgements

We would like to extend our sincerest thanks to STEM Fellowship for running the Big Data Challenge and to all those who helped run learning sessions. First, we want to thank our mentor Curtis for finalizing our plan. We also want to acknowledge the Ontario Health Database for providing us with the dataset and Google Trends for compiling the historic keyword search interest. Lastly, we thank Dr. Sacha Noukhovitch and Ms. Qaiser for their ongoing support.

References

[1] https://globalnews.ca/news/4754876/opioid-overdose-canada-deaths-2018/

[2] https://www.theglobeandmail.com/news/national/ontario-invests-222-million-to-combat-opioid-crisis/article36113584/

[3] https://www.theglobeandmail.com/news/national/ontario-invests-222-million-to-combat-opioid-crisis/article36113584/

[4] https://bmjopen.bmj.com/content/8/10/e018335

[5] https://qz.com/1189730/google-is-using-46-billion-data-points-to-predict-the-medical-outcomes-of-hospital-patients/

[6] https://www.ncbi.nlm.nih.gov/pubmed/29864105

[7] https://www.nejm.org/doi/full/10.1056/NEJMra1508490

[8] http://underpressure.hqontario.ca

[9] https://www.who.int/violenceprevention/interpersonal_violence_and_illicit_drug_use.pdf

[10] https://www.publichealthontario.ca/en/dataandanalytics/pages/opioid.aspx