MedCrave
Step into the World of Research

International Clinical Pathology Journal

Research Article

Open Access

CrossMark

A data science approach to identify previously unknown indicators that could lead to the prevention of suicide in USA

Abstract

The goal of this research paper is to propose an approach to uncovering important and previously unknown indicators that could prevent suicide by analyzing data released by the Centers for Disease Control and Prevention (CDC). The CDC data includes 75 variables that identify characteristics and demographics of the deceased. This paper describes the process of exploring data science methods to build a predictive model that could lead to prevention of suicide rate in USA. The data was explored with the Python programming language, R scripting language, RStudio and in the Tableau environment. Preliminary analysis of the CDC 2013 deaths by suicide data indicate that education was one of the important factors that lead to suicide; further analysis of this and other data may provide a new set of objective risk factors from one or more of the variables in this data set.

Keywords: CDC, WHO, data science, demographics, education, suicide, suicide risk factors, 2013 death data

Volume 2 Issue 4 - 2016

Toni Brandt,[1] Gilberto Diaz,[1] Jacob Jones,[1] Todd Gary,[1] Ashwini Yenamandra[1,2]

[1]College of Computing and Technology, Lipscomb University, USA

[2]Vanderbilt University Medical Center, USA

Correspondence: Ashwini Yenamandra, Medical Director, Dept. of Pathology, Immunology and Microbiology, Vanderbilt University Medical Center, 719 Thompson Lane, Nashville, USA, Email ashwini.yenamandra@vanderbilt.edu

Received: June 13, 2016 | **Published:** June 29, 2016

Abbreviations: CDC, disease control and prevention; WHO, world health organization; SSI, scale for suicide ideation; MSSI, modified scale for suicide ideation; SIS, suicide intent scale; SABCS, suicidal affect behavior cognition scale; SBQ, suicide behaviors questionnaire; LOI, life orientation inventory; RFL, reasons for living inventory; NGASR, nurses global assessment of suicide risk

Introduction

Problem space

A better understanding of the risk factors leading individuals to commit suicide may help reverse the increase in rate and lead to better prevention strategies. Referencing the Suicide rates rising in US, CDC report, Wikipedia states, "suicide is a leading cause of death in the United States and this rate has risen 24% from 1999 to 2014.[1]" The World Health Organization (WHO) published, "Unlike other causes of death such as cancer, accidents, etc., suicide may be preventable, especially if the risk factors are known and appropriate interventions employed (2016)." There exists a large body of research literature focused on different known factors that contribute to suicide; however the rate of suicide continues to rise. This would suggest that the current identified risk factors are not enough to help with prevention? Many of the risk factors are gathered from survivors and can be subjective. Are there objective factors that can predict suicide and help in prevention? The CDC has identified suicide as a major threat to American lives and has taken efforts to observe and identify suicidal risk factors. The risk factors listed below[2] for suicide are combination factors of individual, relational, community, and society.

i. Family history of suicide
ii. Family history of child maltreatment
iii. Previous suicide attempt(s)
iv. History of mental disorders, particularly clinical depression
v. History of alcohol and substance abuse
vi. Feelings of hopelessness
vii. Impulsive or aggressive tendencies
viii. Cultural and religious beliefs (e.g., belief that suicide is noble resolution of a personal dilemma)
ix. Local epidemics of suicide
x. Isolation, a feeling of being cut off from other people
xi. Barriers to accessing mental health treatment
xii. Loss (relational, social, work, or financial)
xiii. Physical illness
xiv. Easy access to lethal methods
xv. Unwillingness to seek help because of the stigma attached to mental health and substance abuse disorders or to suicidal thoughts.

The issue is that the identified risk factors are not based on demographics, rather self or professional evaluation. Since many of the identified risk factors are based on a subjective assessment that is difficult to gather or identify, a demographic suicide risk factor assessment could be conducted based on a demographic data to determine high-risk suicide patients. This study proposes analysis of the data gather by the CDC about the 2013 US suicide victims, to find if a correlation exists between suicide and varying demographic information. If such a correlation exists, data modeling should help in predicting individuals at risk of suicide.

Motivation

The CDC recognizes suicide is one of leading cause of death worldwide, and the United States is not an exception, accounting for 42,773 deaths in just 2013. As of 2015,[3] suicide was the 10th highest causes of death (2015). For years, healthcare professionals have been fighting relentlessly regarding this issue. According to Dinah Miller, M.D. of Psychology *today* suicide rates are increasing every year.[4]

In Simon's[5] article *Suicide risk assessment: is clinical experience enough?* He states, "Accurate and defensible risk assessment requires a clinician to integrate a clinical judgment with the latest evidence-based practice, although accurate prediction of low base rate events, such as suicide, is inherently difficult and prone to false positives (2006)." According to contributors to the *Assessment of Suicide Risk* Wikipedia page,[6] effective suicide risk assessment, "...should distinguish between acute and chronic risk. Acute risk might be raised because of recent changes in the person's circumstances or mental state, while chronic risk is determined by a diagnosis of a mental illness, and social and demographic factors. Suicide risk assessments are currently conducted with the following assessments:"

i. The Scale for Suicide Ideation (SSI)
ii. The Modified Scale for Suicide Ideation (MSSI)
iii. The Suicide Intent Scale (SIS)
iv. The Suicidal Affect Behavior Cognition Scale (SABCS)
v. The Suicide Behaviors Questionnaire (SBQ)
vi. The Life Orientation Inventory (LOI)
vii. The Reasons for Living Inventory (RFL)
viii. The Nurses Global Assessment of Suicide Risk (NGASR)

As noted by Bryan et al.[7] in Advances in the assessment of suicide risk, "There are risks and disadvantages to both overestimation and under-estimation of suicide risk. Over-sensitivity to risk can have undesirable consequences, including inappropriate deprivation of patients' rights and squandering of scarce clinical resources. On the other hand, underestimating suicidality as a result of a dismissive attitude or lack of clinical skill jeopardizes patient safety and risks clinician liability (2006)." Given that suicide rates continue to rise, there is reason to believe the assessments that are currently in use are not comprehensively capturing the motivations and severity of being able to properly classify the suicide risk level of patients.

The purpose of this paper is to explore social and demographic factors that are currently not being utilized in assessments that could help identify the risk of suicide in people. By uncovering these indicators, we speculate insights can be gained to aid in the improvement of suicide risk assessments being utilized by healthcare professionals to positively impact the efficacy of the efforts to decrease suicide rates in the United States. If the correlation is discovered between suicide and previously ignored risk factors,[8] actionable programs could be developed and targeted for high risk groups. These programs could be designed to help patients seeking clinical care and also, as well as those at high-risk who are not actively seeking clinical help.

Related work

While the CDC has gathered an extraordinarily large data set on suicides in the United States, the variables have not revealed clear motivators as to why suicides rates continue to rise. Per Miller (2013), the CDC reports revealed the following: "The news from the Centers for Disease Control shows a striking increase in suicide rates. Among those ages 35 to 64years old (the baby boomers), there is a 28% increase in suicide rates from 1999 to 2010. It holds for males (up 27%), females (up 31%), and across different regions of the country. The peaks were seen in men in their 50's and women in their early 60's. The gender difference continues to show that men die of suicide at three times the rate of women, and suicide is now the 4th cause of death for that age group. More people die of suicide than car accidents. The rise is most striking in non-Hispanic whites and Native American Alaskan Indians, groups that have always had the highest rates. The suicide rate is now 17 per 100,000, up from 13 per 100,000. And while we worry more about homicide, *suicide rates are twice the homicide rates*. Marriage is protective, as is a college education, and in fact the suicide rate in college-educated women went down."Further, according to Milner et al.[9] in a study published in the British Journal of Psychology, not all variables are known that are strong indicators of suicide, (2013). "This study confirms that certain occupational groups are at elevated risk of suicide compared with the general employed population, or compared with other occupational groups. At greatest risk were laborers, cleaners and elementary occupations (ISCO major category 9), followed by machine operators and ship's deck crew (ISCO major group 8) ...The greater risk of suicide in lower skilled occupational groups may be symptomatic of wider social and economic disadvantages, including lower education, income and access to health services."

In addition to the non-monetary impacts of suicide, according to the CDC's website (2016), there are significant financial impacts to society:

i. Suicide costs society over $44.6 billion a year in combined medical and work loss costs.
ii. The average suicide costs $1,164,499.

The majority of the individual suicide cost is a result of the work-loss amounting in $1,160,655 and the and the total amount increases once the average medical cost of $3,646 is calculated in.[10] The monetary impact to society was explored in *Suicide and Suicidal Attempts in the United States: Costs and Policy Implications.* Sheppard calculated the cost to society in millions by component and age range in the following Table 1.[11]

Outline of paper

This paper is organized as follows. Section 1 is the introduction of the paper, containing the problem space, motivation and related work. Section 2 describes the materials and data preparation involved in analyzing and drawing meaningful conclusions. Section 3 is a discussion of the results of the various data science techniques. Section 4 discusses potential challenges. Section 5 presents the possible future work for this project. Section 6 offers a conclusion based on the results. Section 7 lists the references used in the research paper. Finally, section 8 contains Tables, code, and charts of interest.

Materials and methods

Data collection

The data was collected from The Centers for Disease Control and Prevention (CDC), http://www.cdc.gov/nchs/data_access/vitalstatsonline.htm#Mortality_Multiple. The file contains all the death records of people who die in the United State during 2013. The file format is DUSMCPUB. The data includes 75 variables that identify characteristics of the death as well as some characteristics of the deceased. For example, where did it happen, what day of the week, race, education level, etc.

Description of CDC data: The CDC death data is based on detailed mortality files. This data is recorded for every death in the country and each row of the data file represents a single death. All data comes from the CDC's National Vital Statistics Systems, with the exception of the Icd10Code, which are sourced from the World Health Organization.

Data Preparation and processing

i. Step 1: The 2013 death data file was downloaded in a DUSMCPUB format.

ii. Step 2: The file required conversion into a CSV format to be imported into RStudio. To accomplish the conversion, a modified python parser (Appendix A) available on GitHub was utilized to parse the DUSMCPUB data into a CSV file. Step 3: After parsing the data set into a CSV file, the data was imported into RStudio.

iii. Step 4: As suicide risk could be correlated to objective demographic variables, a subset of 17 variables from the original 75 variables were identified to create a subset of data needed for correlation analysis of unknown variables to suicide risk. The variables identified as potentially significant are listed in Appendix A.

iv. Step 5: To aid in correlation analysis and, ultimately, a linear regression model of the most significantly correlated variables, variables were assigned factor levels, as described in the R script in Appendix B.

v. Step 6: The original data contained all causes of death in the Manner_Of_Death variable; however, the research is focused only on variables correlated with the Suicide value. As described, in the R script in Appendix A subset of data was created to contain only suicide related death.

vi. Step 7: According to aforementioned studies, the Education variable was likely to be highly correlated to suicide risk. As described in the R script in Appendix C, the records without Education data were removed to remove noise from the data set.

vii. Step 8: To conduct preliminary analysis of the data subset described in steps 4-7 in Tableau, the data subset was exported from RStudio in a CSV format.

viii. Step 9: The exported CSV file containing subset of data was imported into Tableau for data visualization and analysis Figure 1.

Results

Description of data set found and created for analysis

As can be seen in the Tables 1 & Table 2 below, the original death data file contained 2,601,452 rows and 75 variables (i.e., columns). This data was reduced to contain only death labeled as suicided, bringing the record count to 41,509. Some of these rows of data were missing education levels; these records were removed, leaving a sample size of 6,336 and 17 variables that represent demographic data that could be valuable to the research.

Potential data science approach

Potential data science approaches being explored are clustering, regression and hypothesis testing to identify any significance the 17 selected variables will have to predict the probability a subject would commit suicide. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is the main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics (Wikipedia). Regression is defined as a technique in which a straight line is fitted to a set of data points to measure the effect of a_single independent variable. The slope of the line is the measured impact of that_variable.[12] Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true. The usual process of hypothesis testing consists of four steps namely null hypothesis, test statistic, p value and comparison of p value to an acceptable significance value called alpha value to see if the effect is statistically significant, then the null hypothesis is ruled out, and the alternative hypothesis is valid.[13]

The rationale for applying a data science approach to the 2013 CDC death dataset is the success achieved in genome sequencing using data science. In the Center for Disease Control and Prevention blog, Khoury states, "Genome sequencing of humans and other organisms has been a leading contributor to Big Data, but other types of data are increasingly larger, more diverse, and more complex, exceeding the abilities of currently used approaches to store, manage, share, analyze, and interpret it effectively. We have all heard claims that Big Data will revolutionize everything, including health and healthcare." Khoury MJ[14] by discovering associations and understanding patterns and trends within the data, big data analytics has the potential to improve care, save lives and lower societal impact.

Preliminary findings

The data was placed into the Tableau environment. Initial analysis, shown in Figure 1, including variables, such as: gender, marital status, location of death, and education attainment levels. This initial observation seems to indicate lower education level appears to be an indicator and is at least one of the potential risk factors that the authors feel deserves more research. This finding supports the idea that risk factors can be found in this data set.

Rationale for using data set

The CDC's website,[15] defines suicide as, "Death caused by self-directed injurious behavior with intent to die as a result of the behavior." It is believed the appropriate data to find unknown suicide risk indicators from the large number variables available and data published by the CDC. In addition to the large number of variables and records, their data is reliable. According to the CDC's website,[16] their suicide data is gathered through the resources:

i. National Electronic Injury Surveillance System-All Injury Program (NEISS-AIP)

ii. National Hospital Ambulatory Medical Care Survey

iii. National Inpatient Sample (NIS)

National Violent Death Reporting System

The National Vital Statistics System

WISQARS

Youth Risk Behavior Surveillance System (YRBSS)

Other federal data sources

i. Drug Abuse Warning Network

ii. National Survey on Drug Use and Health (NSDUH)

Non-federal data sources

i. Pan American Health Association, Regional Core Health Data Initiative

ii. The American Association of Suicidology

iii. WHO Statistical Information System (WHOSIS).[17]

Table 1 Suicide breakdown cost

Components	Males	Females	Total	%
Medical Cost				
Suicides	$121	$26	$146	0.3
Nonfatal Suicide Attempts	$1,149	$388	$1,537	2.6
Total (all self-inflicted injuries)	$1,270	$413	$1,684	2.9
Indirect Economic Cost				
Suicides	$43,589	$9,458	$53,047	90.8
Nonfatal Suicide Attempts	$3,196	$518	$3,714	6.4
Total (all self-inflicted injuries)	$46,785	$9,976	$56,761	97.1
Total Economic Cost				
Suicides	$43,710	$9,483	$53,193	91
Nonfatal Suicide Attempts	$4,346	$906	$5,251	9
Total (all self-inflicted injuries)	$48,056	$10,392	$58,445	100

Source: Author's calculation.

[a]Items may not sum to totals due to rounding.

Table 2 Observation and variable counts

Name	Observations	Variables
Original death data set	2,601,452	75
Original by suicide	41,509	17
Suicide by education	6,336	17

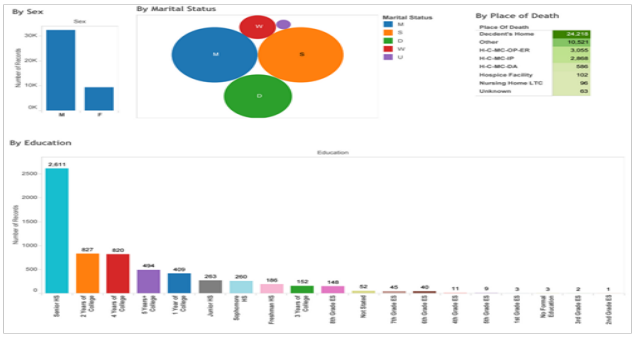

Figure 1 Distribution of 4 variables among victims of suicide. This Tableau dashboard view contains the data distribution by (a) gender, (b) marital status, (c) location of death, and (d) education attainment levels.

Potential challenges

There has been significant research conducted in the rise of suicide in the United States, however, the published research has been unable to produce a solution to the rise of suicide. Within the scope of this research, the potential challenges have been identified by the research team:

i. The original data set from the 2013 death data made available by the CDC is incredibly large compared to the final sample size of 6,336 records from 2,601,452.

ii. The majority of the research team is new to data science and is considered non-experts in the domain. The ability to find and use the correct data science method could prove difficult.

iii. The data set made available only accounts for the deceased and does not include any living patients. Additional data sets may be needed for a conclusive study, which may be protected information.

iv. Managing false positives and mis-classifying someone with a high risk or low risk of committing suicide.

v. If an unknown variable is identified, how would information is provided to the proper people in a timely manner to help with prevention?

Future work

This project will be expanded in the following way. First the best data science approach will be defined in detail in practicum I portion of the Lipscomb University Data Science Master's degree program. Then this approach will be implemented and previously unknown risk factor variables for suicide identified. Once this is completed, this information and findings will be shared at a research conference and possible published in the scientific literature.

Conclusion

Due to the increasing rise of suicide in the United States, research was initiated with a data science approach to identify previously unknown indicators that could lead to the prevention of suicide in the US. Previous research has been conducted to determine indicators of suicidal deaths; however, the research was based on subjective analysis of a suicidal subject's likelihood to commit suicide. This research sought to focus on indicators that were objective characteristics so the risk assessments conducted on suspected suicidal patients could

potentially increase the accuracy of the risk assessment study. The studies reviewed prior to forming the research question did not utilize data science approaches to reach their conclusions that lower education levels and labor intensive occupations lead to a higher suicidal risk. It is believed that a linear regression model can be formed to fit the variables identified in the Center for Disease Control's death dataset of 2013 that are the most significantly correlated with reported suicidal death. If the model proves accurate, subjects of the populations fitting the criteria of high risk characteristics could be introduced to potentially life-saving preventative actions to reduce the probability the subject's cause of death would be suicide.

Acknowledgements

The Authors acknowledge the CDC and the WHO for the making the data available to public and for analysis and Sarah Goteluschen MS (Data Science) for her critical comments and suggestions.

Conflict of interest

The author declares no conflict of interest.

References

1. https://en.wikipedia.org/wiki/Assessment_of_suicide_risk
2. Paddock C. *Suicide rates rising in US, CDC report*. Medical News; 2016.
3. Suicide: Consequences. CDC; 2015.
4. Dihna Miller. Rising suicide rates: have we simply failed? *Psychology Today*. 2013.
5. Simon R. Suicide risk assessment: is clinical experience enough? *J Am Acad Psychiatry Law*. 2006;34(3):276–278.
6. http://www.who.int/mediacentre/factsheets/fs398/en/
7. Bryan CJ, Rudd MD. Advances in the assessment of suicide risk. *Journal of Clinical Psychology*. 2006;62(2):185–200.
8. Suicide: Risk and Protective Factors. CDC; 2015.
9. Milner A, Spittal MJ, Pirkis J, et al. Suicide by occupation: Systematic review and meta-analysis. *The British Journal of Psychiatry*. 2013;203(6):409–416.
10. *The double bottom line impact*. Working Minds; 2009.
11. Sheppard DS, Geruwich D, Lwin AK, et al. Suicide and suicidal attempts in the united states: costs and policy implications. *Suicide and Life-Threatening Behavior*. 2015;46(3):352–362.
12. Kutner MH. *Applied linear statistical models*. 4th ed. Chicago, USA: McGraw-HiII Companiea; 1996. 318 p.
13. Weisstein, Eric W. Hypothesis Testing. Wolfram Math World; 2016.
14. Khoury MJ. Public Health Approach to Big Data in the Age of Genomics: How Can We Separate Signal from Noise? CDC; 2014.
15. Definitions: Self-directed Violence. CDC; 2015.
16. Suicide: Data Sources. CDC; 2015.
17. https://en.wikipedia.org/wiki/Cluster_analysis