Predicting deaths by suicide in the US general population in 2017, using sociodemographic variables: a machine learning study.

Elena Benini [1], Stefano Boldrini [2], Davide Dametti [3], Ennio Antonio Guzman Culmenares[4] , Corrado Montoro [5]

[1] Università degli studi di Milano-Bicocca, APPLIED EXPERIMENTAL PSYCHOLOGICAL SCIENCES
[2] Università degli studi di Milano-Bicocca, corso di laurea magistrale DATA SCIENCE
[3] Università degli studi di Milano-Bicocca, corso di laurea magistrale DATA SCIENCE
[4] Università degli studi di Milano-Bicocca, corso di laurea magistrale DATA SCIENCE
[5] Università degli studi di Milano-Bicocca, corso di laurea magistrale DATA SCIENCE

Abstract

Objective: We wanted to exploit ML potentiality to predict suicides in the general population, without having available psychological and health indicators. If this would prove feasible, national social health care system may benefit from these models in order to provide aimed intervention to citizens at major risk.

Methods: We trained four ML algorithms, a Random Forest, a J48, a Naive Bayes Tree and a Logistic Regression Models on a wide dataset (N= 1048575), reporting information about all the deaths in the USA in 2017.

Results: Our models reached a decent recall (around 0.75), which is a crucial index when such imbalanced-class classifications are performed, but a lower accuracy (around 0.62), probably because of the lack of psychological and health variables.

Conclusions: National social health system may benefit of ML models in order to predict citizens at higher risk of suicide, since it is possible to predict this phenomenon far above chance, just using sociodemographic variables that such system has easily available.

1. Introduction

Suicide is one of the major causes of deaths all around the world (WHO, 2014), which could be assessed and treated, reducing its enormous impact. Hence, it is a major concern for public health to be able to predict citizens' suicide risk, in order to devise and implement effective interventions. Machine learning techniques are being increasingly employed in psychological research and have proved effective in improving prevention of suicidal intentions and behaviour and thus, may contribute to inform and support intervention efforts. ML has proved useful in overcoming limitations of classical research aimed at predicting suicidal behaviour and intentions, thanks to the opportunity to consider a large bunch of variables and their interactions, and allowing non-linear relations among them, which are crucial factors when trying to predict such a complex and multifaceted behaviour as suicide (Barros et al., 2017; Burke, Ammerman, & Jacobucci, 2019). Furthermore, a tool which allows to predict the rarer suicide deaths phenomenon, and not just suicidal thoughts and intentions, may be of extreme social interest, but few studies have focused on this aim, according to the literature review carried over by Burke and colleagues (Burke, Ammerman & Jacobucci, 2019). Therefore, our research aimed at predicting suicide death in the general population, leveraging the wide availability of public health data. We selected a dataset from the National Center for Health Statistics (part of the United States Department of Health and Human Services), reporting all deaths occurrences in the U.S. in 2017.

Alongside cause of death, place and date of death, some sociodemographic and sociocultural data are available, which have proved useful in order to classify people as suicide-prone (Hettige et al., 2017, Burke et al., 2019). In facts, we had data about sex, marital status, place of death occurrence with respect to place of residence, education level, age of death, if the subject underwent some injury and where did it happen, race, Hispanic origin, what they were doing when death occurred (a complete list of variables and their description can be found in Appendix).

However, our data lacked variables such as suicide ideation, suicidal attempts and psychological/psychiatric conditions, which are well established predictors of suicide commitment, together with history of alcohol or drug abuse and stress, that we also lack (Hettige et al., 2017, Passos et al., 2016, Burke, Ammerman & Jacobucci, 2019). On the other hand, our research benefited from a wide publicly available community dataset (N = 1.048.575) as Burke and colleagues recommended to do in their review. Since it collects information regarding the general population, the suicide rate is very low in our dataset, as compared to clinical samples, (e.g. schizophrenia patients). However, it is clear that such a wide sample would not be easily drawn from a specific population (e.g. psychiatric, abused, veteran population), thus, in spite of the lack of some crucial variables we mentioned, we trust our results to be easily generalizable. The present research is thus aimed at developing a ML model which is able to predict the rare occurrences of suicide deaths in the general population starting from sociodemographic and sociocultural variables, that have been extensively used to this scope in the literature and to which social health care has easy access.

2.Methods

2.1 Preprocessing

Before the training phase, we adopted some preprocessing techniques to adapt data to our scope. First of all, we filtered out some observations according to the age of death. Particularly, we took the decision to exclude from analyses all the deaths occured to children being less than 9 years old. Despite children under that age who think about, or commit, suicide are, sadly, much more than one may expect (Tishler et al., 2017), figures were extremely small for deaths in USA in 2017. Hence, through this filtering we reduced the size of the dataset by 10646 units and we only lost 2 suicide occurrences. Proportion of positive class of the variable of interest is now 1,74% of the whole dataset.

2.2 Feature selection

As it visible in the Appendix, many features measured the same sociodemographic variable. For example, we have filtered out educ1989 variable, because of the extensive presence of missing values. Furthermore, it would have been uneffective to use a system that was changed in 2003 for a dataset from 2017. Instead, we use educ2003. The variable educflag does not refer directly to the subjects, but to the system used for recoding (see appendix).

Weekday, Monthdth, placedth and age have been excluded. The aim of the research was to create a model that could be used, for instance, in a Center for mental health that wants to predict if the subject will die by taking his or her own life. It is crucial to notice that the variable age was actually the age of death (how old the subject was when he or she died). Now, it is impossible for a Center for mental health to have available the information of the age of death as an independent variable when the subject is still alive. In fact, the value of the age of death will remain unknown until the subject actually die (which could happen in the near future as well as in 50 years). The same line of reasoning can be applied to the month of death and the week of death. It would be impossible to say that a subject will eventually die, for instance, the first of March or the 13th of November. Thus, it is impossible to use the information of the month and of the day of the week in which the subject die as an independent variables. Again, we didn’t use placdth (place of death) because the place where the subject will die cannot be known before the subject actually die.

Since the variable age has been sorted out all the other variables related to age (age flag, ager57, ager27, ager12, ager22) have been excluded (see appendix for details).

The variable injury stands for the place of injury. Of course, it is not possible to know when the subject will die. Thus, this variable cannot be an independent variable. Injwork instead, stands for injury at work and it have been excluded because of a massive number of missing values (‘U’ which stands for unknown). Moreover, here we faced the same theoretical problem we had for age of death. The same happened for activity, mandeath (which furthermore is perfectly correlated with suicide, since rows equal to 2 corresponded to suicides).

Ucod, ucr358, ucr113, ucr130, ucr39 are the underlying cause of death (and various different categorization of this variable. See appendix). While initially we thought they could be very useful to predict cases of suicide, it turned out that they are not really the 'underlying' cause of death, but the cause of death itself. Again, we do not know the cause of death before the subject is dead. Thus, we excluded these variables.

Several variables (such as eanum, econdp_1, …, record_20) have nothing to due with the subjects themselves but rather with the way the data were originally stored (see appendix). Thus, they have been sorted out.

Marital status is an important socio-demographic variable and it can be known. Of course it can be that the value of the variable will change in the time. For instance, it can be that a single person will eventually get married. This is a limitation of the model.

Race is a variable that has been used, since its value can be known before the subject dies. The variables brace and raceimp have been excluded since they just tell the criteria for evaluating the values of race. Racer3 and racer5 are different recording of race, but, since they contain the same information, we excluded them.

The feature hispanic and hispanicr contain the same information in 2 different recordings. We used hispanicr since it is a data that it is available before the subject dies.

2.3 Definition of suicide class variable

We manually created the feature suicidal from the feature mandeath. In facts, as we stated above, the value 2 of the feature mandeath indicated "suicide" as a manner of death. The feature suicidal takes the values 0 (not suicidal) and 1 (suicidal).

2.4 Descriptive statistics

After the removal of irrelevant variables, we assessed the classes frequency of the variables of interest: restatus, edu2003, sex, marstat, suicidal, race and hspanicr. As anticipated in the previous paragraph, we dealt with a seriously imbalanced dataset with respect to the target variable, suicides being a 1,74% of the total deaths. Sex variable was equally distributed (51,9% males). The vast majority are US residents, for a total of 84%, while intrastate non-residents covers 13% and the remainder is composed of foreign residents. Variable marstat have five different possibilities that are Married (37%), Divorced (17%), Single (13%) , Widowed (32%) and Unknown (1%). In a multiracial nation such as the United States is quite usual that the variable race could have lots of different races as confirmed in our dataset, where however the 82% is composed by "white", followed by "black" with 10% and the remaining 9% involve all the others possible races. Hspanicr, the hispanic origin variable, is composed for 68% by Non-Hispanic white, 11% Mexican, 9% Non-Hispanic black, 8% others non-hispanic races while the other variables have quite insignificant percentages. Edu2003 is the variable with the largest number of possible levels, one may just want to know that the most frequent value was 3, which corresponded to the high school diploma. 9 instead, meant unknown education history, that will be assessed in "missing replacement" section. It may be of interest that only 7,23% of deaths regarded people with a master degree (7-8).

2.5 Missing replacement

Selected variables did not comprise missing values, but marstat and educ2003 showed some unknown values (U) which were a small percentage among both the general sample (0.9% having unknown marital status and 2.1% unknown education level) and the suicide victims (1.2% having unknown marital status and 2.1% having unknown education level). Such unknown values have been replaced with the most frequent value, conditioned on the suicide class attribute. In particular, most frequent marital status value was single among suicide victims and married among natural deaths, whereas most frequent education level was 3 (high school diploma) among both groups.

2.6 Models

In order to find the best model to predict suicides in the general population, we incorporated four differents algorithms within the workflow. The chosen algorithms were:

1. Random Forests

2. J48 Decision Tree
3. Naïve Bayes Tree
4. Logistic regression

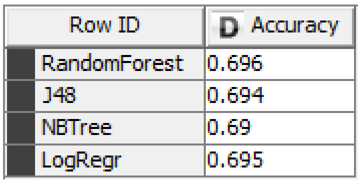
Through this choice, we could assess the accuracy of different classes of models, namely Heuristic Models with Random Forest and Decision Tree J48, Probabilistic Models with Naive Bayes and Regression Based Models with Logistic regression. We preferred J48 Decision Tree provided by Weka environment over Decision Tree by Knime, since the latter performed better on a equal-size-sampled testset (50% suicides, 50% natural deaths) but it performed worse on a testset with a realistic proportion of suicides.

2.7 Cross validation and classifiers comparison

After preprocessing, the dataset has been partitioned with a stratified sampling technique, where 90% of the observations would be used for the consecutive partition, while the remaining 10% would be used as testset for the final evaluation of the chosen models. Such partition is crucial in order to save part of the dataset that will be used as "unknown data" to evaluate models performance in case of new sets of data, namely its generalizability.

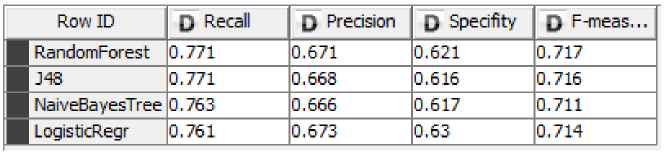
We further divided the 90% of the initial dataset into two additional partitions, where the 70% was submitted to an equal size sampling, which resulted in 22814 observation: 50% of suicide case and 50% of natural causes of deaths. Having exactly the same number of observations for each class of an attribute was considered the best solution due to the huge imbalance of cases of suicides. This strategy would not lead to a precise evaluation of our models' goodness on real data, this notwithstanding, it was useful to assess if the algorithms we chose were able to predict the phenomenon at all, given the available set of variables . Indeed, classifiers proper evaluation would be carried over using the 10% of the observations we cut out in advance.

After the equal size sampling, in order to limit overfitting and underfitting phenomena and increase accuracy of models performance estimates, we adopted a cross-validation technique with random sampling. Data were divided into 3 subsets, with each fold containing approximately the same percentage of suicides and natural deaths. The decision to use stratified 3 k-cross validation is justified by CV better performance with respect to other techniques like holdout and iterated holdout and by the necessity to have a model which is faster in terms of computation time than 5 or 10 k-cross validation considering all the models used. Models accuracy are reported in table 1 and in the box plot of figure 1.



Row ID	D Accuracy
RandomForest	0.696
J48	0.694
NBTree	0.69
LogRegr	0.695

Tabella 1:Accuracies of the 4 models averaged across cross validations



Row ID	D Recall	D Precision	D Specifity	D F-meas...
RandomForest	0.771	0.671	0.621	0.717
J48	0.771	0.668	0.616	0.716
NaiveBayesTree	0.763	0.666	0.617	0.711
LogisticRegr	0.761	0.673	0.63	0.714

Tabella 2:Performance measures of the classifiers.

When averaged accuracy was take into account, the 4 classifiers showed rather similar performances, with Random forest being slightly better than the other and the worst being Naive Bayes Tree. The same pattern was found for F-measure index. What is more interesting for our scope is Recall, since it is better to erroneously identify a natural death as a suicide than failing to identify a suicide as such. In the first scenario, some public resources are wasted, since the person was not at risk of committing suicide, whereas in the second a person who might have benefited from a psychological/psychiatric intervention would have been left without. Here again, Random forest algorithm showed higher recall with respect to other models. The lack of relevant differences between classifiers performances is clearly visible through ROC curve depiction (figure 3), where the 4 different-coloured curves are mostly overlapping. Area under ROC curves was equal to .77 for logistic regression, 0.75 for Naive Bayes Tree and J48 and 0.74 for Random Forest. Despite we cannot claim that classifiers achieved an excellent performance, these models are still able to detect suicides using sociodemographic and sociocultural variables far above chance level.

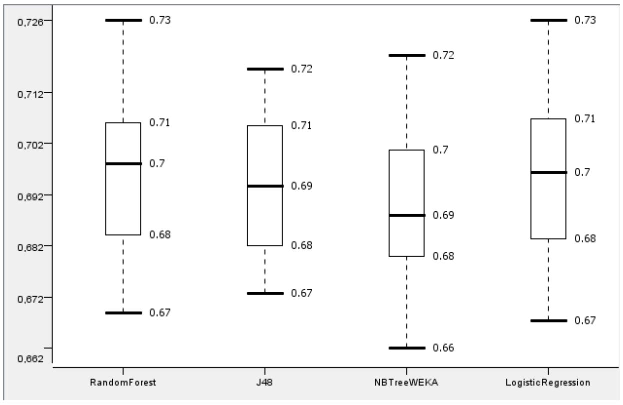


Figura 1: Box Plot of the accuracies of the 4 models in 20 cross validations on a sample with half suicides and half natural deaths.

2.8 Backward features elimination

As we extensively commented above, we selected a restricted set of predictors (7), so that we did not need to run a features selection to improve interpretability, nor to discard less relevant variables. This notwithstanding, we decided to perform a feature selection on the random forest model, which was the only “black box” method we used, in order to gain insight on the contribution of each variable to the model predictive accuracy. To do so, we implemented a backward features elimination loop in knime: Random Forest algorithm was fed with the same portion of equal-size-sampled data, whereas we used an independent partition of the data as a test set. We saved out the latter partitioning as 30% of the 90% of the original dataset through stratified sampling. What emerged, was that sex was the first variable to be eliminated, then education, residence status, hispanic race and, finally, race.

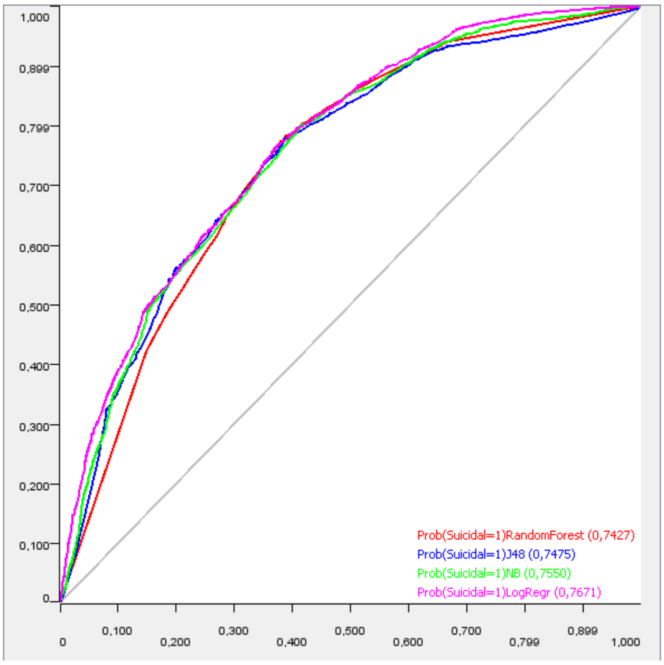


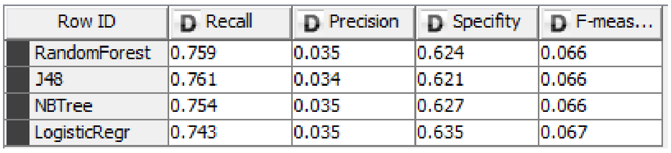
Figura 2: ROC curve of the 4 models performances

3.Evaluation

All of the 4 models performed similarly in predicting target variable in a sample with half of suicides and half of the natural deaths. Particularly, error rates of the best model (Random Forest) were not significantly different from error rates of the worst (Logistic Regression): this meant that also all the other models, the performance of which fell within these 2, were not significantly different between themselves and thus the 4 models were not distinguishable in terms if error rates. Thus, we carried over all of them to the evaluation phase. In this phase, we trained the 4 algorithms on the same equal-size-sampled training set we used for cross-validation, but we used an independent 10% of the original data as test set. Such 10% had been obtained through a stratified sampling of the original dataset, so that it showed the same proportion of suicides. Clearly, we expected performance to decrease in this strongly imbalanced testset. Nonetheless, if the aim of the present research was to build models which were able to predict suicides in the general population, then we needed to test their performance on data which presented the same proportion of positive class as the real phenomenon. While during cross-validation phase we wanted to test the feasibility of our project, in the evaluation phase we focused on the efficacy of it (the effectiveness will be briefly assessed in the discussion). Results are visible in Table 3. As it can be seen in the table below (Table 4) recall was about 75%, but precision was quite low. This means that the number of false positives was high in comparison to the number of true positives. In spite of the presence of a high number of false positives, we had a higher specificity because of the huge amount of true negatives. Thus, this pattern is clearly explained by suicide variable extreme imbalance. This means that a lot of records might have been mistakenly classified as positive, resulting in false positives. Conversely, there were few positive cases of suicide that could be mistakenly classified as negative, becoming false negative (this explain the high recall). The overall low precision is also the cause of the overall low level of the F-measure. Since we have a high number of false positives, when trying to predict cases of suicide, it must be taken into consideration the low level of precision of the models. If a subject is predicted to be at risk of suicide, it can be the case of a false positive. However, since the recall was high, the models could be useful to predict if a subject will take his or her life.

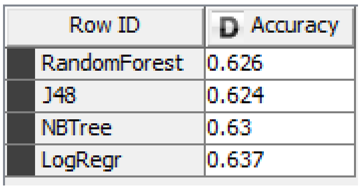
Accuracy was around 62% in each model, as visible in Table 4.

It might have been possible to reach a higher accuracy, if the models were also trained (and not only tested) on a stratified sample, thus with a very low suicides rate. Nonetheless, given such low rate, a high accuracy might have been reached classifying all of the observations as natural deaths, and none as a suicide. It is clear that this would have not been of much help, given our scope.



Row ID	D Recall	D Precision	D Specifity	D F-meas...
RandomForest	0.759	0.035	0.624	0.066
J48	0.761	0.034	0.621	0.066
NBTree	0.754	0.035	0.627	0.066
LogisticRegr	0.743	0.035	0.635	0.067

Tabella 3: Some indeces of performance of our models



Row ID	D Accuracy
RandomForest	0.626
J48	0.624
NBTree	0.63
LogRegr	0.637

Tabella 4: Overall accuracy of the models in the evaluation phase

3.1 Discussion

The aim of the present research was to employ the increasingly used machine learning methods in order to predict deaths by suicide. Many others trained machine learning algorithms for this scope, showing that sociodemographic variables are good predictors of suicide commitment, despite psychological and health indicators revealed much more relevant (Burke et al., 2019). In our dataset, we could benefit of a huge sample (N = 1048575), but we lacked those crucial health variables. As a matter of facts, our models reached a recall ranging from 0.74 to 0.76, which is just slightly less than what Barros et al. (2017) achieved (sensitivity=0.77) using patients diagnosed with mental health conditions. Their specificity was a bit higher though, being 0.78. In another ML study which collected questionnaire-based data from mental health patients, (Morales et al., 2017), their models showed a recall of 0.63 and a specificity of 0.79. These are just some examples, others can be found in Burke et al. 2019. Since our major concern is to gain a high recall (which implies not to ignore suicides), our models performed decently also when compared to what has been done previously in the literature. Moreover, those authors could benefit of samples which were much less imbalanced than ours, since, unfortunately, people with mental conditions have much higher suicides rates than the general population. In addition, they had available some crucial psychological variables that we lacked. This notwithstanding, our models' performances were far above chance. Hence, it may be possible for a country social health care system to predict which citizens present a higher suicide risk, using only sociodemographic information, which the mentioned social health system has easily available.

3.2 Limitations and further improvement

We often mentioned throughout the document the fact that our dataset lacked some crucial variables as psychological and health indicators. This was a strong limitation that strongly undermined our models' predictive accuracy. In the future, it may be of help to collect some more mind/body health indicators in the general population. Moreover, some of our variables are not stable across life, marital status for example. Thus it may be difficult to assess with certainty a people marital status in the future, given her actual one.

Given the relevance of suicide attempts as a predictor for suicide commitment, it may be sensible to train a model on a sample of suicide attempters versus non-attempters, who are a bigger slice of the population when compared to suicide victims (Burke et al., 2019) and are worth of the same deal of attention by the health care system. A classification may be performed on a community sample, leveraging on the higher occurrence rate which should increase models' performance, so that suicide attempters may be reached by public interventions which, in turn, may reduce their suicidal intention.

A further improvement may come from collecting data from written text on social media, which has proved to be a good predictor of suicide risk (Burke et al., 2019). However, this may be viewed as an unbearable violation of citizens privacy. It is true that many users publish their thoughts on public profiles, but this does not automatically implies for a country health system to be authorized to collect such information and perform analyses on it.

Conclusions

Suicide is a leading cause of death in the world. In this paper, we tried to address suicide prevention using Machine Learning. We found out that statistical models can indeed predict suicide cases beyond the level of randomness using sociodemographic variables alone. In particular, in our analyses, accuracy was around 62%. Furthermore, the recall was even higher (up to 0.76), meaning that our algorithms can classify correctly a high percentage of the future suicides. Thus, there is no doubt that Machine Learning can be useful in suicide prevention. One problem our models face is the huge amount of false positives. This is a problem that must be taken into consideration when employing these models in the actual practice of suicide prediction.

References

Barros, J., Morales, S., Echávarri, O., García, A., Ortega, J., Asahi, T., … Núñez, C. (2017). Suicide detection in Chile : proposing a predictive model for suicide risk in a clinical sample of patients with mood disorders. *Revista Brasileira de Psiquiatria*, *39*, 1–11. https://doi.org/10.1590/1516-4446-2015-1877

Burke, T. A., Ammerman, B. A., & Jacobucci, R. (2019). The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of Affective Disorders*, *245*, 869–884. https://doi.org/10.1016/j.jad.2018.11.073

Hettige, N. C., Nguyen, T. B., Yuan, C., Rajakulendran, T., Baddour, J., Bhagwat, N., … De Luca, V. (2017). Classification of suicide attempters in schizophrenia using sociocultural and clinical features: A machine learning approach. *General Hospital Psychiatry*, *47*, 20–28. https://doi.org/10.1016/j.genhosppsych.2017.03.001

Morales, S., Barros, J., Echávarri, O., García, F., Osses, A., Moya, C., … Tomicic, A. (2017). Acute mental discomfort associated with suicide behavior in a clinical sample of patients with affective disorders: Ascertaining critical variables using artificial intelligence tools. *Frontiers in Psychiatry*, *8*(2), 1–16. https://doi.org/10.3389/fpsyt.2017.00007

Passos, I. C., Mwangi, B., Cao, B., Hamilton, J. E., Wu, M. J., Zhang, X. Y., … Soares, J. C. (2016). Identifying a clinical signature of suicidality among patients with mood disorders: A pilot study using a machine learning approach. *Journal of Affective Disorders*, *193*, 109–116. https://doi.org/10.1016/j.jad.2015.12.066

Tishler, C. L., Reiss, N. S., & Rhodes, A. R. (2007). Suicidal Behavior in Children Younger than Twelve: A Diagnostic Challenge for Emergency Department Personnel. *Academic Emergency Medicine*, *14*(9), 810–818. https://doi.org/10.1197/j.aem.2007.05.014

World Health Organization. (2014). Preventing suicide: a global imperative. World Health Organization. http://www.who.int/iris/handle/10665/131056

Appendix: Values of the features

restatus: "resident status". The possible values are:

1. RESIDENTS (State and County of Occurrence and Residence are the same)
2. INTRASTATE NONRESIDENTS (State of Occurrence and Residence are the same, but County is different)
3. INTERSTATE NONRESIDENTS (State of Occurrence and Residence are different, but both are in the U.S.)
4. FOREIGN RESIDENTS (State of Occurrence is one of the 50 States or the District of Columbia, but Place of Residence is outside of the U.S.)

educ1989: the way the education level is recorded from 1989 until 2003.

- 00. No formal education
- 01-08.Years of elementary school
- 09. 1 year of high school
- 10. 2 years of high school
- 11. 3 years of high school
- 12. 4 years of high school
- 13. 1 year of college
- 14. 2 years of college
- 15. 3 years of college
- 16. 4 years of college
- 17. 5 or more years of college
- 99. Not stated

educ2003: the way 2003 on.

- 1. 8th grade or less
- 2. 9 - 12th grade, no diploma
- 3. high school graduate or GED completed
- 4. some college credit, but no degree
- 5. Associate degree
- 6. Bachelor's degree
- 7. Master's degree
- 8. Doctorate or professional degree
- 9. Unknown

educflag: which education recording system has been used (in our dataset, it has always been used only educ2003)

- 0. 1989 revision of education item on certificate
- 1. 2003 revision of education item on certificate
- 2. no education item on certificate

monthdth: the month in which the subject died. The value 1 corresponds to January, the value 2 to February and so on

sex: it can take the value F for female subjects or M for male subjects.

age: the actual age of the subject summed to the constant 1000.

ageflag: If reported age is unknown but a valid age is calculated using dates of birth and death, the calculated age is substituted for the unknown reported age.

- Blank. Calculated age is not substituted for reported age
- 1. Calculated age is substituted for reported age

ager52: the age of the subjects has been recorded in 51 intervals. The value 52 corresponds to missing values. The values that it can take are:

- 01. Under 1 hour (includes not stated hours and minutes)
- 02. 1 - 23 hours
- 03. 1 day (includes not stated days)
- 04. 2 days
- 05. 3 days
- 06. 4 days
- 07. 5 days
- 08. 6 days
- 09. 7 - 13 days (includes not stated weeks)
- 10. 14 - 20 days
- 11. 21 - 27 days
- 12. 1 month (includes not stated months)
- 13. 2 months
- 14. 3 months
- 15. 4 months
- 16. 5 months
- 17. 6 months
- 18. 7 months
- 19. 8 months
- 20. 9 months
- 21. 10 months
- 22. 11 months
- 23. 1 year
- 24. 2 years
- 25. 3 years
- 26. 4 years
- 27. 5 - 9 years
- 28. 10 - 14 years
- 29. 15 - 19 years
- 30. 20 - 24 years
- 31. 25 - 29 years
- 32. 30 - 34 years
- 33. 35 - 39 years
- 34. 40 - 44 years
- 35. 45 - 49 years
- 36. 50 - 54 years
- 37. 55 - 59 years
- 38. 60 - 64 years

- 39. 65 - 69 years
- 40. 70 - 74 years
- 41. 75 - 79 years
- 42. 80 - 84 years
- 43. 85 - 89 years
- 44. 90 - 94 years
- 45. 95 - 99 years
- 46. 100 - 104 years
- 47. 105 - 109 years
- 48. 110 - 114 years
- 49. 115 - 119 years
- 50. 120 - 124 years
- 51. 125 years and over
- 52. Age not stated

Ager27: the age of the subjects has been recorded in 26 intervals. The value 27 corresponds to missing values. The values that it can take are:

- 01. Under 1 month (includes not stated weeks, days, hours, and minutes)
- 02. 1 month - 11 months (includes not stated months)
- 03. 1 year
- 04. 2 years
- 05. 3 years
- 06. 4 years
- 07. 5 - 9 years
- 08. 10 - 14 years
- 09. 15 - 19 years
- 10. 20 - 24 years
- 11. 25 - 29 years
- 12. 30 - 34 years
- 13. 35 - 39 years
- 14. 40 - 44 years
- 15. 45 - 49 years
- 16. 50 - 54 years
- 17. 55 - 59 years
- 18. 60 - 64 years
- 19. 65 - 69 years
- 20. 70 - 74 years
- 21. 75 - 79 years
- 22. 80 - 84 years
- 23. 85 - 89 years
- 24. 90 - 94 years
- 25. 95 - 99 years
- 26. 100 years and over
- 27. Age not stated

ager12: the age of the subjects has been recorded in 11 intervals. The value 12 corresponds to missing values. The values it can take are:

- 01. Under 1 year (includes not stated infant ages)
- 02. 1 - 4 years
- 03. 5 - 14 years
- 04. 15 - 24 years
- 05. 25 - 34 years
- 06. 35 - 44 years
- 07. 45 - 54 years
- 08. 55 - 64 years
- 09. 65 - 74 years
- 10. 75 - 84 years
- 11. 85 years and over
- 12. Age not stated

ager22: the age of the subjects has been recorded in 21 intervals. The subjects are all babies. The value Blank corresponds to missing values or more thant 1 year in general. The values that it can take are:

- Blank. Age 1 year and over or not stated
- 01. Under 1 hour (includes not stated hours and minutes)
- 02. 1 - 23 hours
- 03. 1 day (includes not stated days)
- 04. 2 days
- 05. 3 days
- 06. 4 days
- 07. 5 days
- 08. 6 days
- 09. 7-13 days (includes not stated weeks)
- 10. 14 - 20 days
- 11. 21 - 27 days
- 12. 1 month (includes not stated months)
- 13. 2 months
- 14. 3 months
- 15. 4 months
- 16. 5 months
- 17. 6 months
- 18. 7 months
- 19. 8 months
- 20. 9 months
- 21. 10 months
- 22. 11 months

placdth: the variable is the place of death of the subject. It can take the values:

- 1. Hospital, clinic or Medical Center - Inpatient
- 2. Hospital, Clinic or Medical Center - Outpatient or admitted to Emergency Room
- 3. Hospital, Clinic or Medical Center - Dead on Arrival
- 4. Descendent’s home
- 5. Hospice facility
- 6. Nursing home/long term care
- 7. Other
- 9. Place of death unknown

marstat: it describes the marital status. It can take the values:

- S. Never married, single
- M. Married

- W. Widowed
- D. Divorced
- U. Marital Status unknown

weekday: this variable describes the day of the week of death. It can take the values:

- 1. Sunday
- 2. Monday
- 3. Tuesday
- 4. Wednesday
- 5. Thursday
- 6. Friday
- 7. Saturday
- 9 . Unknown

year: this variable describes the year when the death occured. It can take only one value, which is 2017.

injwork: the variable describes if the subject was injured at work. It can take the values:

- Y. Yes
- N. No
- U. Unknown

mandeath: a categorization of the possible manners in which subjects may die. It can take the values:

- 1. Accident
- 2. Suicide
- 3. Homicide
- 4. Pending investigation
- 5. Could not determine
- 6. Self-Inflicted
- 7. Natural
- Blank. Not specified

suicidal: this variable was not present in the original dataset It has been added manually. For every value "2" of the mandeath variable (see the previous variable)

methdisp: it stands for 'method of disposition'. It can take the values:

- B. Burial
- C. Cremation
- D. Other
- U. Unknown

autopsy: it says if any autopsy has been done on the corpse. It can take the values:

- Y. Yes
- N. No
- U. Unknown

activity: it describes the activity in which the subject was engaged when he passed away. It can take the values:

- 0. While engaged in sports activity
- 1. While engaged in leisure activity
- 2. While working for income
- 3. While engaged in other types of work
- 4. While resting, sleeping, eating (vital activities)
- 8. While engaged in other specified activities
- 9. During unspecified activity
- Blank. Not applicable

injury: place of Injury. It can take the values:

- 0. Home
- 1. Residential institution
- 2. School, other institution and public administrative area
- 3. Sports and athletics area
- 4. Street and highway
- 5 ... Trade and service area 6 ... Industrial and construction area 7 ... Farm 8 ... Other Specified Places 9 ... Unspecified place Blank ... Causes

ucod: it stands for 'underlying cause of death'. This variable has not been used in the present study. For further references see the International Classification of Diseases, 2004 Revision, Volume 1.

ucr358. A recode of the ICD cause code into 358 groups for NCHS publications