Short Paper

Identifying Rare Circumstances Preceding Female Firearm Suicides: Validating A Large Language Model Approach

Weipeng Zhou[1], BA; Laura C Prater[2,3], MPH, PhD; Evan V Goldstein[4], MPP, PhD; Stephen J Mooney[5], MS, PhD

[1]Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, WA, United States
[2]Department of Psychiatry and Behavioral Health, University of Washington, Seattle, WA, United States
[3]Harborview Medical Center, School of Medicine, University of Washington, Seattle, WA, United States
[4]Department of Population Health Sciences, University of Utah, Salt Lake City, UT, United States
[5]Department of Epidemiology, School of Public Health, University of Washington, Seattle, WA, United States

Corresponding Author:
Stephen J Mooney, MS, PhD
Department of Epidemiology
School of Public Health
University of Washington
Hans Rosling Center for Population Health
3980 15th Ave NE
Seattle, WA, 98195
United States
Phone: 1 206 685 1643
Email: sjm2186@uw.edu

Abstract

Background: Firearm suicide has been more prevalent among males, but age-adjusted female firearm suicide rates increased by 20% from 2010 to 2020, outpacing the rate increase among males by about 8 percentage points, and female firearm suicide may have different contributing circumstances. In the United States, the National Violent Death Reporting System (NVDRS) is a comprehensive source of data on violent deaths and includes unstructured incident narrative reports from coroners or medical examiners and law enforcement. Conventional natural language processing approaches have been used to identify common circumstances preceding female firearm suicide deaths but failed to identify rarer circumstances due to insufficient training data.

Objective: This study aimed to leverage a large language model approach to identify infrequent circumstances preceding female firearm suicide in the unstructured coroners or medical examiners and law enforcement narrative reports available in the NVDRS.

Methods: We used the narrative reports of 1462 female firearm suicide decedents in the NVDRS from 2014 to 2018. The reports were written in English. We coded 9 infrequent circumstances preceding female firearm suicides. We experimented with predicting those circumstances by leveraging a large language model approach in a yes/no question-answer format. We measured the prediction accuracy with F_1-score (ranging from 0 to 1). F_1-score is the harmonic mean of precision (positive predictive value) and recall (true positive rate or sensitivity).

Results: Our large language model outperformed a conventional support vector machine–supervised machine learning approach by a wide margin. Compared to the support vector machine model, which had F_1-scores less than 0.2 for most infrequent circumstances, our large language model approach achieved an F_1-score of over 0.6 for 4 circumstances and 0.8 for 2 circumstances.

Conclusions: The use of a large language model approach shows promise. Researchers interested in using natural language processing to identify infrequent circumstances in narrative report data may benefit from large language models.

(JMIR Ment Health 2023;10:e49359) doi: 10.2196/49359

KEYWORDS

female firearm suicide; large language model; document classification; suicide prevention; suicide; firearm suicide; machine learning; mental health for women; violent death; mental health; language models; women; female; depression; suicidal

Introduction

Suicide is a leading cause of death in the United States. Suicide risk factors include physical and mental health disorders, substance use disorders, prior exposure to violence, and having a firearm at home [1-4]. Firearm suicide has been more prevalent among men, but age-adjusted female firearm suicide rates have increased by 20% from 2010 to 2020, outpacing the rate increase among males by about 8 percentage points [5]. However, relatively few studies have focused specifically on female firearm suicide instead focusing on samples in which males are overrepresented (eg, military veterans) [6,7]. More data are needed to identify circumstances surrounding female firearm systems, and a primary source of these data derives from unprocessed narrative reports in the National Violent Death Reporting System (NVDRS).

In a previous study [1], we found that conventional natural language processing (NLP) algorithms could successfully identify some relatively common circumstances preceding female firearm suicide, using coroners' and medical examiners' (CMEs') and law enforcement (LE) narrative reports provided by the NVDRS. However, because reliably training a conventional NLP pipeline requires a sizeable training data set, the approach worked well only for the most common preceding circumstances.

Recently, large language models such as ChatGPT were found to perform well on tasks such as answering yes/no questions and document classification [8-10]. Large language models were developed on the basis of large corpora of data crawled from the web and can be used to solve machine learning tasks in a question-answer format. Moreover, these large language models do not rely on the task's data set size. In this study, we explored the value of a large language model approach by framing our coding task as a binary response for classification. Specifically, we tested a large language model approach to identify infrequent circumstances preceding female firearm suicide and compared this approach's performance to that of traditional machine learning models.

Methods

Overview

ChatGPT is the state-of-the-art large language model, but we could not use it directly in this study. Our data contain protected information; hence, we could not upload them directly to ChatGPT. Instead, we used open-source large language model alternatives. We ran these models locally to protect potentially sensitive information in the data. These models are developed similarly and are competitive in certain areas compared to ChatGPT. In a benchmark evaluation of large language models' ability in problem-solving [11], FLAN-T5 [12] and FLAN-UL2 [13] were found to be less accurate than ChatGPT for world knowledge understanding and programming but competitive in following complex instructions, comprehension and arithmetic, and causal reasoning. In preliminary studies, we experimented with multiple large language models (FLAN-T5 [12], FLAN-UL2 [13], and others) and found that FLAN-UL2 performed the best. We hence used FLAN-UL2 for our subsequent experiments. Developed by Google LLC, FLAN-UL2 is an open-source, 20 billion–parameter encoder-decoder model and is useful for zero-shot learning (ie, the model makes predictions directly without further training).

Data Sets

We used the NVDRS Restricted Access Database of female firearm suicides from 2014 to 2018 [1]. The data set contained unstructured CME and LE narrative reports describing the circumstances leading up to the suicide deaths of 1462 females. The reports were written in English. We manually coded 9 infrequent circumstances (ie, labels) preceding the firearm suicide deaths following the instructions specified by Goldstein et al [1]: sleep problems, abusive relationships, custody issues, sexual violence, isolation or loneliness, substance abuse, dementia, bullying, and caregiver issues. All infrequent labels occurred in <5% of the cases. We have provided details regarding the circumstance distribution in Table S1 in Multimedia Appendix 1. We split the data set into training and test sets with a 0.5:0.5 ratio.

Model Evaluation

A prompt is the input for the large language model and will guide it for generating outputs. For FLAN-UL2, we designed a prompt as a pair of a narrative report and a question. The narrative report is the text we input into a traditional machine learning model. The question varies depending on the circumstances we want to the model to code. For example, for the circumstance "bullying," "Answer the following yes/no question: was the decedent experiencing bullying in-person or online? Answer:" is the question. The model will yield an output of "Yes" if "bullying" is mentioned in the narrative report; if not, "No" will be the output. The question was adapted from the definition of each label developed through a previously reported manual review process with minimal changes [1]. A complete list of questions and definitions for each label is included in Table S2 in Multimedia Appendix 1. As a baseline, we used a series of conventional support vector machine (SVM) models [14] trained to identify each circumstance. FLAN-UL2 was only applied on the test set. SVM models were trained on the training set and applied on the test set. We repeated all experiments 5 times with resampling of the training and test sets. We reported the average F_1-score, which is the harmonic mean of the precision (positive predictive value) and recall (true positive rate or sensitivity). The F_1-score measures the model's accuracy considering the imbalance in the data set and ranges from 0 to 1.

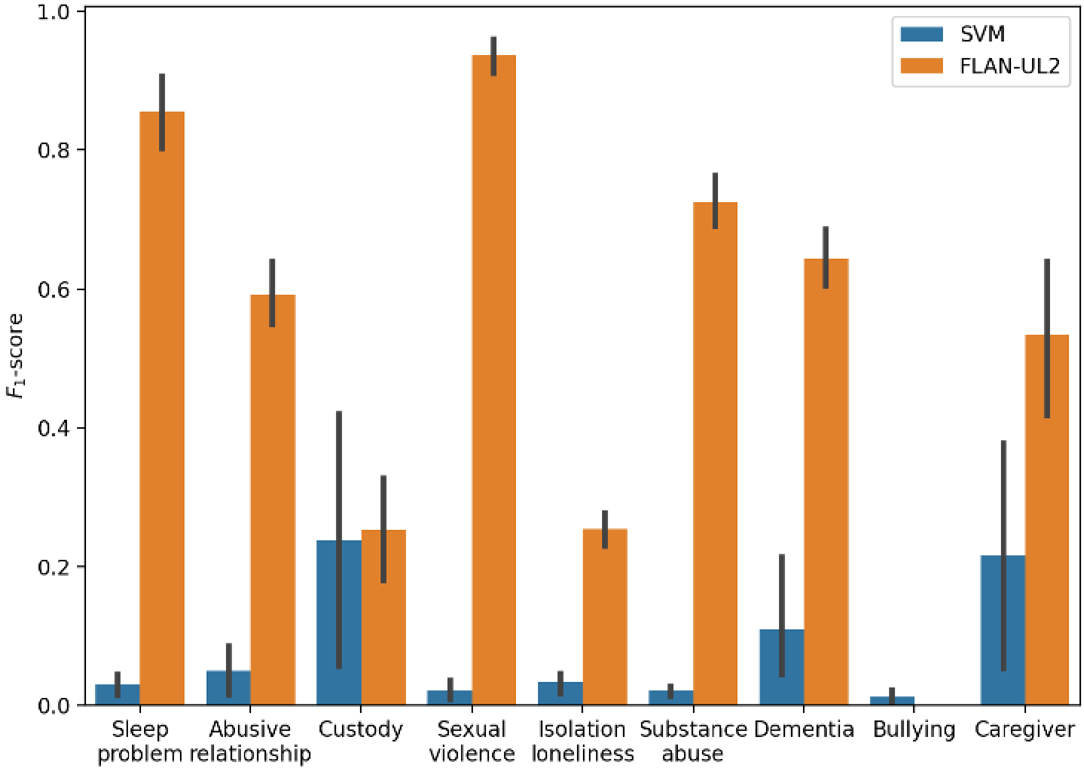
Ethical Considerations

This study did not require approval from the University of Washington institutional research board because these deidentified data on deceased persons were not considered human subjects research. We received ethical approval from the National Violent Death Reporting System Restricted Access Database review committee (request number #410).

Results

FLAN-UL2 performed better than the SVM for most of the female firearm suicide circumstances, sometimes substantially better (Figure 1). The F_1-score of FLAN-UL2 was greater than 0.8 for "sleep problem" (3.4% prevalence) and "sexual violence" (2.6% prevalence). "Bullying" (1.6% prevalence) was the only circumstance where the SVM outperformed FLAN-UL2, and F_1-scores were 0 or nearly 0 for all SVM and FLAN-UL2 runs. See Table S1 in Multimedia Appendix 1 for prevalence in other circumstances.

Figure 1. F_1-scores of the support vector machine (SVM) and FLAN-UL2 model for coding rare circumstances of female firearm suicide from suicide reports when averaged over 5 runs. The height of the bars represents the mean F_1-score, and the line at the tip of the bars represents the SD across 5 runs.



Discussion

Principal Findings

We found that the large language model (FLAN-UL2) outperformed the SVM for 8 of the 9 infrequent circumstances preceding female firearm suicide deaths. These findings suggest that a large language model approach can address a critical gap in identifying infrequent circumstances in unstructured text. Unlocking these data efficiently allows for subsequent analyses of female firearm suicide risk, including relationships among sexual violence, dementia, sleep problems, and firearm suicide.

The characterization of circumstances preceding female firearm suicides is an understudied area. In a previous study, Goldstein et al [1] used traditional NLP methods to predict 5 circumstances from suicide reports, with F_1-scores ranging from 0.6 to 0.8. However, all these circumstances had a prevalence of at least 15%. In our study, all 9 circumstances had a prevalence of less than 5%. We complemented the existing work by providing a method for automatically coding circumstances preceding female firearm suicides at a larger scope.

The failure in identifying the "bullying" circumstance may be due to the fact that bullying is one of the most infrequent circumstances preceding female firearm suicide in the NVDRS. The question we provided to the large language model, "was the decedent experiencing bullying in-person or online?" might not be sufficiently sensitive for the model to understand how to identify these circumstances in the narrative reports. More detailed explanation of "bullying," such as the victim was insulted or hurt at school or at the workplace, might be needed for the model to reason better. Large language models are known to be sensitive to prompt text, and designing an appropriate prompt (also known as prompt engineering) is an essential part of using large language models effectively [10]. Novel prompting techniques, such as few-shot learning (provide problem examples as part of the prompt) [10], have been proposed and may improve large language models' performance. In this study, we used simple and consistent prompts to provide a baseline for using large language models to code infrequent circumstances preceding female suicide. Large language models are also computationally expensive. The experiments in this study were carried out on 2 NVIDIA A100 40 GB graphics processing units. Large language models are also known to be sensitive to "hallucination" [15], meaning that they generate paragraphs of texts that look reasonable but are factually incorrect. In this study, we prompted the model to generate yes/no answers, bypassing the risks of hallucination.

Conclusions

Our large language model successfully identified infrequent circumstances preceding female firearm suicide deaths, having outperformed conventional NLP approaches by a wide margin. This finding suggests that large language models can be used to unlock textual analysis within public health research. More broadly, the success of our relatively simple queries at identifying infrequent circumstances suggests that large language models may be useful in public health surveillance, potentially allowing practitioners to track the prevalence of infrequent conditions that are never explicitly coded into surveillance systems. Future studies should explore the performance of different large language models and variations in the models' underlying mechanisms when applied to coding infrequent circumstances.

Acknowledgments

This work was funded by grants from the National Collaborative on Gun Violence Research (PI: Prater/ Goldstein, 2021) and the National Library of Medicine (4R00LM012868). We thank Dr Meliha Yetisgen for providing the computational resources.

Authors' Contributions

WZ conceptualized the study; carried out the formal analysis, investigation, and validation; designed the methodology; visualized the data; and drafted, edited, and reviewed the manuscript. EVG conceptualized the study, carried out the investigation, acquired funding, designed the methodology, supervised the study, and reviewed and edited the manuscript. LCP conceptualized the study, curated the data, acquired funding, designed the methodology, carried out the investigation, and edited and reviewed the manuscript. SJM conceptualized and supervised the study, acquired funding, designed the methodology, carried out the investigation, and edited and reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prevalence, definitions and derived questions of the rare female firearm suicide circumstances.
[DOCX File , 28 KB-Multimedia Appendix 1]

References

1. Goldstein EV, Mooney SJ, Takagi-Stewart J, Agnew BF, Morgan ER, Haviland MJ, et al. Characterizing Female Firearm Suicide Circumstances: A Natural Language Processing and Machine Learning Approach. Am J Prev Med 2023 Aug;65(2):278-285 [doi: 10.1016/j.amepre.2023.01.030] [Medline: 36931986]
2. Anestis MD. Prior suicide attempts are less common in suicide decedents who died by firearms relative to those who died by other means. J Affect Disord 2016 Jan 01;189:106-109 [doi: 10.1016/j.jad.2015.09.007] [Medline: 26432034]
3. Miller M, Zhang Y, Prince L, Swanson SA, Wintemute GJ, Holsinger EE, et al. Suicide Deaths Among Women in California Living With Handgun Owners vs Those Living With Other Adults in Handgun-Free Homes, 2004-2016. JAMA Psychiatry 2022 Jun 01;79(6):582-588 [FREE Full text] [doi: 10.1001/jamapsychiatry.2022.0793] [Medline: 35476016]
4. Kellermann AL, Rivara FP, Somes G, Reay DT, Francisco J, Banton JG, et al. Suicide in the home in relation to gun ownership. N Engl J Med 1992 Aug 13;327(7):467-472 [doi: 10.1056/NEJM199208133270705] [Medline: 1308093]
5. WISQARS™ — Web-based Injury Statistics Query and Reporting System. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/injury/wisqars/index.html [accessed 2023-02-03]
6. Spark TL, Cogan CM, Monteith LL, Simonetti JA. Firearm Lethal Means Counseling Among Women: Clinical and Research Considerations and a Call to Action. Curr Treat Options Psychiatry 2022;9(3):301-311 [FREE Full text] [doi: 10.1007/s40501-022-00273-3] [Medline: 35791313]
7. Rowhani-Rahbar A, Simonetti JA, Rivara FP. Effectiveness of Interventions to Promote Safe Firearm Storage. Epidemiol Rev 2016;38(1):111-124 [doi: 10.1093/epirev/mxv006] [Medline: 26769724]
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
9. Xu F, Alon U, Neubig G, Hellendoorn V. A systematic evaluation of large language models of code. 2022 Presented at: MAPS '22: 6th ACM SIGPLAN International Symposium on Machine Programming; June 13, 2022; San Diego, CA [doi: 10.1145/3520312.3534862]
10. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv Preprint posted online May 28, 2020. [FREE Full text] [doi: 10.5860/choice.189890]
11. Chia K, Hong P, Bing L, Poria S. INSTRUCTEVAL: towards holistic evaluation of instruction-tuned large language models. arXiv Preprint posted online June 7, 2023. [FREE Full text]

12. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. arXiv Preprint posted online October 20, 2022. [FREE Full text]
13. Tay Y. A New Open Source Flan 20B with UL2. Yi Tay. URL: https://www.yitay.net/blog/flan-ul2-20b [accessed 2023-04-03]
14. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, MA: MIT Press; 2016.
15. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 2023 Mar 19;11(6) [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]

Abbreviations

CME: coroners or medical examiners
LE: law enforcement
NVDRS: National Violent Death Reporting System
NLP: natural language processing
SVM: support vector machine

Edited by J Torous; submitted 30.05.23; peer-reviewed by R Panirselvam, L Souza; comments to author 26.08.23; revised version received 31.08.23; accepted 02.09.23; published 17.10.23

Please cite as:
Zhou W, Prater LC, Goldstein EV, Mooney SJ
Identifying Rare Circumstances Preceding Female Firearm Suicides: Validating A Large Language Model Approach
JMIR Ment Health 2023;10:e49359
URL: https://mental.jmir.org/2023/1/e49359
doi: 10.2196/49359
PMID: 37847549