HHS Public Access

Author manuscript

Inj Prev. Author manuscript; available in PMC 2023 February 01.

Published in final edited form as:

Inj Prev. 2022 February ; 28(1): 74–80. doi:10.1136/injuryprev-2021-044322.

Leveraging data science to enhance suicide prevention research: a literature review

Avital Rachelle Wulz[1], **Royal Law**[2], **Jing Wang**[2], **Amy Funk Wolkin**[2]

[1]Oak Ridge Associated Universities (ORAU), Division of Injury Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

[2]Division of Injury Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Abstract

Objective—The purpose of this research is to identify how data science is applied in suicide prevention literature, describe the current landscape of this literature and highlight areas where data science may be useful for future injury prevention research.

Design—We conducted a literature review of injury prevention and data science in April 2020 and January 2021 in three databases.

Methods—For the included 99 articles, we extracted the following: (1) author(s) and year; (2) title; (3) study approach (4) reason for applying data science method; (5) data science method type; (6) study description; (7) data source and (8) focus on a disproportionately affected population.

Results—Results showed the literature on data science and suicide more than doubled from 2019 to 2020, with articles with individual-level approaches more prevalent than population-level approaches. Most population-level articles applied data science methods to describe (n=10) outcomes, while most individual-level articles identified risk factors (n=27). Machine learning was the most common data science method applied in the studies (n=48). A wide array of data sources was used for suicide research, with most articles (n=45) using social media and

Correspondence to: Avital Rachelle Wulz, Division of Injury Prevention, Centers for Disease Control and Prevention, Atlanta, GA 30341, USA; AWulz@cdc.gov.

Contributors ARW was involved in all aspects of the study, including planning, analysing and reporting the results in the article. RL was involved in analysing, writing and reviewing the article. JW was involved in reviewing the article and providing subject matter expertise to the overall project. AFW was involved in the planning, writing and reviewing of the article.

Competing interests None declared.

Disclaimer The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Patient consent for publication Not required.

Ethics approval This study was exempt from institutional review board approval due to the nature of the literature review.

Provenance and peer review Not commissioned; externally peer reviewed. Data are available in a public, open access repository.

Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/injuryprev-2021-044322).

web-based behaviour data. Eleven studies demonstrated the value of applying data science to suicide prevention literature for disproportionately affected groups.

Conclusion—Data science techniques proved to be effective tools in describing suicidal thoughts or behaviour, identifying individual risk factors and predicting outcomes. Future research should focus on identifying how data science can be applied in other injury-related topics.

BACKGROUND

The increase in availability and type of data in recent years has given rise to expanding the epidemiologic methods used to improve the understanding of disease burden. Traditionally, epidemiology has been based on primary data collection (eg, surveys) or secondary use of data sources (eg, administrative billing records). In recent years, novel data sources have emerged as communication and information collection modernised at an exponential pace. This data modernisation also gave rise to many new analytic approaches to understand and harness data, primarily through advances in data science methodologies. Data science is a multidisciplinary field incorporating statistics, computer science, programming and subject matter expertise to extract knowledge and insights from structured and unstructured data.[1] Data science domains include static and interactive data visualisation (eg, maps), computer automation, machine learning and prediction analytics.[2] Expanding traditional epidemiologic methods to include data science methodologies provides a unique approach to advancing the field of injury prevention.

Ballesteros *et al*[2] recently outlined a public health strategy for applying data science to injury prevention research and surveillance across different injury topics. Their paper highlighted the growing application and added value of data science methods and techniques to injury prevention research. To date, data science has applied techniques and methods to several injury topics, including motor vehicle accidents,[3] falls,[4 5] poisoning,[6 7] violence[8 9] and suicide.[10] Previous reviews have investigated suicide and specific data science methodologies, such as machine learning,[11] and novel data sources.[12 13] However, no study to our knowledge has broadly described how data science as a field has been applied to suicide research.

Suicide is the tenth leading cause of death overall and second leading cause of death for people ages 10–34 years in the USA.[14] About 47 500 lives were lost to suicide in 2019 in the USA[14] and many more people think about (12 million), plan (3.5 million) or attempt (1.4 million) suicide.[15] Suicide and suicide attempts have profound impacts on the society. Suicide attempts may lead to physical injuries[16] and emotional trauma[17] for individuals. Additionally, families, friends and communities may experience immeasurable emotional suffering[18] and economic consequence in terms of medical care and loss of productivity due to suicide and suicide attempts.[14] However, suicide is preventable through evidence-based strategies.[19] Data science may be uniquely positioned to advance our understanding of disproportionately impacted communities and aid in the clinical and public health decision-making process for suicide prevention.

The purpose of this research is to identify in the literature how data science is applied in suicide prevention, describe the current landscape of this literature and highlight areas where data science may be useful for future injury prevention research.

METHODS

Design and search strategy

We conducted a literature review in April 2020 and January 2021 in Medline, PsycInfo and Scopus using the search terms listed in table 1. Due to the nature of this literature review, it was not appropriate or possible to involve patients or the public in the design, conduct, reporting or dissemination plans of this review. This search yielded 2436 records. We removed 622 duplicates. The original literature search involved all topics related to injury prevention and data science. Out of all injury topics, suicide emerged as having the most published research available. For this manuscript, we limited our review to suicide-related articles.

Screening and eligibility assessment

To narrow our review to suicide-related manuscripts, we screened the 1814 records based on titles, abstracts and keywords. We excluded articles that were not original reports and not grey literature (ie, review articles, program or policy evaluations and editorials were excluded), unrelated to suicide (ie, did not contain suicide terms in the abstract, title or keywords), not available in English or did not apply data science methodologies. Data science and suicide were determined using the following criteria.

Data science—Only records that included data science terms such as, ‘machine learning’, ‘natural language processing’, ‘data mining’, ‘data linkage’ and ‘predictive analytics’ were included (table 1). Studies about social media were only included if they used social media data.

Suicide—Only records that included the terms ‘suicide’, ‘suicidality’, ‘suicidal ideation (or thinking about suicide)’, ‘suicidal behaviours’, ‘suicide attempt’ or ‘die by suicide’ were included. This broad range of suicide-related terms was referred to as suicidal thoughts or behaviour in general in this review. Suicidal behaviour was used to refer to acts including suicide plans, non-fatal suicide attempts or suicides (deaths by suicide). We excluded articles where suicide was not central to the paper. Non-suicidal self-injury was not included in the suicide terminology unless it also stated an interest in suicide specifically. Of note, the suicide terms used during the manual review were not applied during the keyword search of the databases.

This screening phase resulted in 225 articles. Two authors individually reviewed the full articles applying the same prior criteria, additionally screening out records not published from 2017 to 2020 to capture the most recent articles for this review. This assessment resulted in 99 eligible articles. Figure 1 illustrates the literature review screening process.

DATA EXTRACTION

For the included 99 articles, we extracted the following: (1) author(s) and year; (2) title; (3) study approach (4) reason for applying data science method; (5) data science method type; (6) study description; (7) data source and (8) disproportionately affected population.

We identified each article's study approach by categorising the article as focusing on individual-level risk factors (eg, Agne *et al*[20] identified predictors of suicide attempts in patients diagnosed with obsessive-compulsive disorder using a machine-learning algorithm) or population-level risk factors (eg, van Mens *et al*[21] used machine learning to predict future suicidal behaviour with population-based longitudinal data).

We categorised five main reasons a study applied data science methods in the suicide literature. We categorised these reasons as follows: (1) identify, (2) predict, (3) classify, (4) describe and (5) link. Studies that aimed to identify suicidal cases were labelled identify. For example, one study used natural-language processing (NLP) of electronic health records (EHRs) to identify suicidal behaviour among psychiatrically hospitalised adolescents.[22] Studies predicting whether individuals or groups are at high risk for suicidal thoughts or behaviour were labelled as predict. Shen *et al*[23] developed a machine-learning algorithm that predicted the probability of suicide attempts in medical college students using self-report data. Records classifying individuals or groups into subgroups of risk for suicidal thoughts or behaviour were categorised as classify, such as the study by Burke *et al*[24] that used machine learning to classify suicide attempt history among youth presenting to medical settings. Articles describing suicidal thoughts or behaviour within a sample population were labelled as describe. Dagar and Falcone[25] analysed content from YouTube videos to describe information on teenage suicide and viewer's engagement with these videos. Records linking datasets together were labelled as link, such as the study by Borschmann *et al*[26] linking ambulance service data to baseline interview data to identify factors for self-harm, including suicide attempts, following release from prison in Australia. Most studies fell into one of the five data analysis types; however, some were classified into two groups when the study used data science methods for two purposes. For instance, Low *et al*[27] illustrated the benefits of NLP to identify and describe changes in language, such as posts related to suicidal thoughts, from online mental health support groups during the initial months of the COVID-19 pandemic.

Type of data science method included:

- Application of novel data—application of novel data in this context involves the use of non-traditional data sources such as social media data or forum-based data for answering suicide-related research questions. Tasks within this category generally require a data science approach as the datasets are difficult to obtain, wrangle and clean and oftentimes require NLP to extract meaningful information from unstructured data.
- Machine learning—machine learning is defined in this context as the capability of a machine to improve its own performance using statistical models to make decisions and incorporate results of each new trial into that model.

- NLP or linguistic analysis—NLP or linguistic analysis involves any method where computers or analytic methods process complex and often unstructured human language.
- Data mining—data mining involves scoping large datasets to identify and extract data and is frequently used as an exploratory tool.
- Data linkage—data linkage is a method of connecting two or more large, independent datasets together to provide new insight into a topic.
- Data sources were categorised into six main groups, multiple categories were selected if applicable:
- Medical record data—medical record data were defined in this context as any data source that directly referred to the health information of an individual or community, including clinical notes and EHR data.
- Survey/questionnaire data—survey/questionnaire data refer to any data collected from individual interviews, self-reported scales and measures and other questionnaires.
- Administrative data— administrative data refers to data collected for administrative purposes (eg, ambulance or correctional facility data).
- Social media and web-based behaviour data—social media data include data extracted from social media platforms, blogs, search trends or other open-sourced websites. (eg, Twitter and Facebook posts).
- Population-based/aggregated data—population-based/aggregated data in this context refer to surveillance data or other systems level data with information about populations or communities (eg, registry data).
- Internet of things (IOT) data—IOT data refer to data collected from technologies people use daily, such as wearable technologies (eg, smart watches) or cellphone data.

Lastly, to identify trends related to health equity, we noted any articles where the population of interest involved disproportionately affected groups. Disproportionately affected groups in this review included veterans/military personnel, incarcerated or formerly incarcerated persons, older adults and Native American/Alaskan Native people. While people experiencing mental illness are also disproportionately affected by suicide,[28 29] we did not include them in this category since most studies in this review identified this group as their population of interest.

RESULTS

Study approach

Characteristics, main findings and the main variables used in the study from all articles focusing on individual-level risk factors and population-level risk factors are found in online supplemental files 1 and 2. The majority of the articles in the review were focused on individual-level risk factors (n=77). The remaining articles focused on population-level risk

factors (n=22). Figure 2 shows the counts of individual-level, population-level and overall articles published between 2017 and 2020. The literature on data science and suicide more than doubled overall from 2019 to 2020. Both individual-level and population-level articles increased each year, with individual-level articles substantially increasing from 2019 to 2020. Articles with individual-level approaches are more than four times as prevalent than articles with population-level approaches (figure 2).

Reason for applying data science method

Most population-level articles applied data science methods to describe (n=10) suicidal thoughts or behaviour or to predict cases of suicidal thoughts or behaviour within a population (n=7) (figure 3). Individual-level articles also tended to use data science methods to predict (n=19) suicidal thoughts or behaviour, as well as to identify (n=27) individual risk factors for suicidal thoughts or behaviour (figure 3). Eleven individual-level articles used data science to describe suicidal thoughts or behaviour in a population or sample. In addition, 11 articles had multiple reasons for applying data science methods to suicide research (figure 3). Less common reasons for applying data science methods in individual-level and population-level articles included linking one or more datasets together and classifying individuals or cases into subgroups for risk of suicidal thoughts or behaviour (figure 3).

Type of data science method

The type of data science method the study employed varied between articles. Table 2 shows the number of articles by article dimension and by individual-level and population-level articles. The most common types of data science methods applied in this review were machine learning (n=46) and application of novel data (n=22). Articles such as by Liu *et al*,[30] Tadesse *et al*[31] and Hettige *et al*[32] used machine learning to identify individuals at risk for suicide. Several studies applied novel data, such as social media data or Google Trends data,[33 34] to their study design to identify and predict suicidal thoughts or behaviour. Less common data science methods used included NLP (n=9), data linkage (n=7) and data mining (n=4). Studies such as by O'Dea *et al*[35] applied NLP or linguistic analysis to identify social media posts to predict suicidal behaviour. Other articles such as by Young *et al*[36] applied data linkage methods to combine different types of data (ie, administrative data, EHR data and baseline survey data) together to predict suicidal thoughts or behaviour.

Both individual-level (n=36) and population-level (n=11) articles tended to use machine learning over other methods, with application of novel data as the second most applied method for individual-level (n=13) and population-level (n=9) studies.

Several articles applied two or more data science methods in one study. For example, Kessler *et al*[37] used data linkage to join two independent datasets together and then applied machine learning to identify inpatient veterans who may be at increased risk of suicide after discharge from a hospital. Additionally, Arendt *et al*[38] analysed Instagram data for suicidal subliminal messages using frame-by-frame coding procedures illustrating the value of applying novel data and NLP together.

Data source

Table 2 also shows the number of individual-level and population-level articles by data source. The majority of data sources used in articles included in this review were social media-based/web-based behaviour (n=45), survey/questionnaire (n=30) and medical record (n=26) data. Not surprising, population-level articles tended to use population-based/aggregated data more than individual-level articles (n=14 vs n=6). Only two studies incorporated IOT data.

Disproportionately affected populations

Eleven studies assessed suicidal thoughts or behaviour on disproportionally affected populations. In particular, six studies focused on veterans or military personnel[34 39–43] (table 2). Borschmann *et al*,[26] Borschmann *et al*[44] and Young *et al*[36] conducted suicide prevention research with people released from prison. Additionally, one study conducted research on the application of data science to suicide data for Native American people[45] and another applied data science to suicide data from older adults living in nursing homes.[46]

DISCUSSION

The present study described the current landscape of the literature that applies data science to suicide prevention to illustrate how data science can advance injury prevention. Our results show that articles with individual-level approaches are more prevalent than articles with population-level approaches. One reason for this difference may be the complexity and availability of data to conduct studies at the population-level. However, applying data science methods to population-level surveillance and epidemiology research in public health may increase efficiency of data collection, improve the development of predictive models and quickly analyse large quantities of data. For example, Choi *et al*[47] illustrated ways to apply machine-learning methods using multiple novel data sources to predict national suicide rates in near real time. This study shows how data science addresses the need to improve timeliness of suicide surveillance to inform public health response.

Most studies in this review described suicidal thoughts and behaviour, identified individual risk factors for suicidal thoughts and behaviour or predicted trends or risks for suicidal thoughts or behaviour. These results highlight ways various data science methodologies can be applied to achieve different study goals.

Novel data are often used to improve timeliness compared with traditional public health data systems, such as annual mortality data from U.S. death certificates that are typically released 11 months after the end of the calendar year.[48] These data are often publicly available increasing access and timeliness. Novel data also have the benefit of capturing insights that are traditionally difficult or otherwise expensive to measure, such as contextual information. For example, Vioules *et al*[49] examined posts in Twitter data streams to identify sudden changes in a user's online behaviour to detect individuals who are at risk with suicidal thoughts. As individuals interact and communicate in new ways, their digital trails left behind, or novel data, may provide added insight into their physical, mental, emotional and social well-being.

Machine learning can identify trends not identified through traditional epidemiologic methods, which often requires a priori knowledge of causation and relationships. NLP is a field of machine learning with the ability of a computer to understand, analyse and manipulate human language. NLP removes the human dependency to preidentify words; the computer determines relationships and similarities to better capture and translate language to a numerical format for ease of computation and analysis. While the use of NLP can save significant time and human resources dedicated to data processing such as categorising data, removing identifying information and extracting summary information from large volumes of data, extracting meaning from text data is complex and may require sophisticated methods or nuanced solutions to avoid errors such as misclassification. Nonetheless, NLP remains a useful tool for suicide prevention research. For instance, Carson *et al*[22] developed a machine-learning algorithm using NLP of EHRs to identify suicidal behaviour among adolescents who were psychiatrically hospitalised.

This review illustrates the wide array of data sources that researchers used for suicide research, such as social media data and Google search data. Public health professionals may be interested in connecting two or more data sources together creating an innovative dataset to identify trends or analyse information in a new way. The data-linkage technique results in a different dataset from the original sources, illustrating data science's added value to research methods and developing new data sources for research. This review showed only two studies incorporating IOT data for suicide prevention.[50 51] Smartphone application data may be an untapped data source for suicide-related research, since smartphone applications can be used for real-time data collection and efficient data sharing. For example, Cohen *et al*[50] used a smartphone application to record the entire therapy session and collect therapist's notes and impressions from clinical sessions with adolescents who indicated suicide risk. Incorporating smartphone technology into the study enabled fast data collection and more efficient data analysis. In addition, as wearable technology becomes less expensive and more accessible, more people will have access to these products resulting in larger datasets for IOT data. For example, a feasibility study by Coppersmith *et al*[51] analysed multiple data sources, including social media and wearable technology (eg, Fitbit, Jawbone), using NLP and machine learning to design an automated system for estimating suicide risk based on the detection of quantifiable signals around suicide attempts.

This review demonstrates the value of applying data science to suicide prevention literature for disproportionately affected groups. Although, only 11 of the included studies prioritised disproportionately affected groups, these articles highlight ways to creatively apply data science to advance health equity-related research. For example, Haroz *et al*[45] applied machine-learning approaches to community-based suicide surveillance system data to better understand factors that impact Native American individuals who attempt suicide.[45] These findings pave the way for future injury-related topics to use data science methods to focus on disproportionately affected groups.

As more individual behaviours are shared online, data science can be used for epidemiologic discovery. Not only does online data tend to have a larger volume but also the data are often contextualised offering additional insight into behaviour, risk factors and circumstance. Data

science methods, such as machine learning, are well positioned to identify and contextualise data for suicide prevention efforts.

While data science proves to bring added value to suicide prevention research, there are important ethical considerations that should be acknowledged such as transparency and interpretability of machine-learning models, user privacy issues in large-scale online data, data security and risk of identification with linked data and bias in some data sources.[2] Other areas of consideration include the applicability of prediction models to real-world clinical practice. For instance, a model may be able to accurately predict half of all future suicides given that the model classified case into 'high risk' and 'low risk' groups. While half of the cases could be identified for intervention, the other half of missed cases would occur in the 'low risk' group.[52] Clinical applications of classification based on predictive risk assessments can impact resource allocation or unequal access to treatment.[53] Future research on improving translation of predictive models to clinical practices of assessing suicide risk may be useful. As data science is used more, advances in ethical practices should be considered in all stages of the research process.[54]

There are several limitations to this literature review. First, we may have missed data science articles. Data science represents a multidisciplinary field that includes many different types of methodologies. Second, due to the cross-cutting nature of data science, we may have oversimplified categorisations for analysis. For instance, the reasons for applying data science categories, such as identify, predict and classify, may oversimplify the reasons researchers harnessed data science methods for their study. In addition, many studies employ multiple data science methods and for some studies it was difficult to tease out methodologies; this analysis may represent under-representation in some of the categories.

CONCLUSION

Suicide prevention researchers have applied data science methodologies to advance the field. Data science techniques proved to be effective tools in describing suicidal thoughts or behaviour, identifying risk factors or other variables and predicting outcomes which may contribute to suicide prevention efforts. Additional research may focus on applying data science techniques to identify protective factors for suicide, alongside risk factors. Focus on how data science can be applied in other injury-related topics should also be considered.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Joseph Russell for his tremendous help initiating this project and organising the available literature within this review. The authors would also like to thank Ruby Samim for her helpful additions to the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

REFERENCES

1. Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. Data science and public health, 2021. Available: https://www.cdc.gov/injury/data/data-science/index.html
2. Ballesteros MF, Sumner SA, Law R, et al. Advancing injury and violence prevention through data science. J Safety Res 2020;73:189–93. [PubMed: 32563392]
3. Behbahani H, Amiri AM, Imaninasab R, et al. Forecasting accident frequency of an urban road network: a comparison of four artificial neural network techniques. J Forecast 2018;37:767–80.
4. Camps J, Samà A, Martín M, et al. Deep learning for freezing of gait detection in Parkinson's disease patients in their homes using a waist-worn inertial measurement unit. Knowledge-Based Systems 2018;139:119–31.
5. Mauldin TR, Canby ME, Metsis V, et al. SmartFall: a smartwatch-based fall detection system using deep learning. Sensors 2018;18:3363.
6. Neill DB, Herlands W. Machine learning for drug overdose surveillance. J Technol Hum Serv 2018;36:8–14.
7. Hu Z, Jing Y, Xue Y, et al. Analysis of substance use and its outcomes by machine learning: II. derivation and prediction of the trajectory of substance use severity. Drug Alcohol Depend 2020;206:107604. [PubMed: 31615693]
8. Jay J Alcohol outlets and firearm violence: a place-based case-control study using satellite imagery and machine learning. Inj Prev 2020;26:61–6. [PubMed: 31467144]
9. Pelham WE, Petras H, Pardini DA. Can machine learning improve screening for targeted delinquency prevention programs? Prev Sci 2020;21:158–70. [PubMed: 31696355]
10. Ryu S, Lee H, Lee D-K, et al. Detection of suicide attempters among suicide ideators using machine learning. Psychiatry Investig 2019;16:588–93.
11. Burke TA, Ammerman BA, Jacobucci R. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: a systematic review. J Affect Disord 2019;245:869–84. [PubMed: 30699872]
12. Robinson J, Cox G, Bailey E, et al. Social media and suicide prevention: a systematic review. Early Interv Psychiatry 2016;10:103–21. [PubMed: 25702826]
13. Picardo J, McKenzie SK, Collings S, et al. Suicide and self-harm content on Instagram: a systematic scoping review. PLoS One 2020;15:e0238603. [PubMed: 32877433]
14. CDC, US Department of Health and Human Services. Web-Based injury statistics query and reporting system (WISQARS), 2021. Available: https://www.cdc.gov/injury/wisqars/index.html
15. Substance Abuse and Mental Health Services Administration. Key substance use and mental health indicators in the United States: results from the 2016 national survey on drug use and health center for behavioral health statistics and quality; 2018.
16. World Health Organization. Suicide prevention: a global imperative. Geneva, Switzerland WHO Press; 2014.
17. Stanley IH, Boffa JW, Joiner TE. PTSD from a suicide attempt: phenomenological and diagnostic considerations. Psychiatry 2019;82:57–71. [PubMed: 30183554]
18. Chapman AL, Dixon-Gordon KL. Emotional antecedents and consequences of deliberate self-harm and suicide attempts. Suicide Life Threat Behav 2007;37:543–52. [PubMed: 17967121]
19. Stone DM, Holland KM, Bartholow BN. Preventing suicide: a technical package of policies, programs, and practice 2017.
20. Agne NA, Tisott CG, Ballester P, et al. Predictors of suicide attempt in patients with obsessive-compulsive disorder: an exploratory study with machine learning analysis. Psychol Med 2020:1–11.
21. van Mens K, Elzinga E, Nielen M, et al. Applying machine learning on health record data from general practitioners to predict suicidality. Internet Interv 2020;21:100337. [PubMed: 32944503]
22. Carson NJ, Mullin B, Sanchez MJ, et al. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. PLoS One 2019;14:e0211116. [PubMed: 30779800]

23. Shen Y, Zhang W, Chan BSM, et al. Detecting risk of suicide attempts among Chinese medical college students using a machine learning algorithm. J Affect Disord 2020;273:18–23. [PubMed: 32421600]
24. Burke TA, Jacobucci R, Ammerman BA, et al. Using machine learning to classify suicide attempt history among youth in medical care settings. J Affect Disord 2020;268:206–14. [PubMed: 32174479]
25. Dagar A, Falcone T. High viewership of videos about teenage suicide on YouTube. J Am Acad Child Adolesc Psychiatry 2020;59:1–3. [PubMed: 31678555]
26. Borschmann R, Thomas E, Moran P, et al. Self-Harm following release from prison: a prospective data linkage study. Aust N Z J Psychiatry 2017;51:250–9. [PubMed: 27012967]
27. Low DM, Rumker L, Talkar T, et al. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: observational study. J Med Internet Res 2020;22:e22635. [PubMed: 32936777]
28. Hjorthøj CR, Madsen T, Agerbo E, et al. Risk of suicide according to level of psychiatric treatment: a nationwide nested case-control study. Soc Psychiatry Psychiatr Epidemiol 2014;49:1357–65. [PubMed: 24647741]
29. Office of the Surgeon General, National Action Alliance for Suicide Prevention. 2012 National strategy for suicide prevention: goals and objectives for action: a report of the US surgeon general and of the National action alliance for suicide prevention; 2012.
30. Liu X, Liu X, Sun J, et al. Proactive suicide prevention online (PSPO): machine identification and crisis management for Chinese social media users with suicidal thoughts and behaviors. J Med Internet Res 2019;21:e11705. [PubMed: 31344675]
31. Tadesse MM, Lin H, Xu B, et al. Detection of suicide ideation in social media forums using deep learning. Algorithms 2020;13:7.
32. Hettige NC, Nguyen TB, Yuan C, et al. Classification of suicide attempters in schizophrenia using sociocultural and clinical features: a machine learning approach. Gen Hosp Psychiatry 2017;47:20–8. [PubMed: 28807134]
33. Barros JM, Melia R, Francis K, et al. The validity of Google trends search volumes for behavioral forecasting of national suicide rates in Ireland. Int J Environ Res Public Health 2019;16:3201.
34. Bryan CJ, Butner JE, Sinclair S, et al. Predictors of emerging suicide death among military personnel on social media networks. Suicide Life Threat Behav 2018;48:413–30. [PubMed: 28752655]
35. O'Dea B, Larsen ME, Batterham PJ, et al. A linguistic analysis of suicide-related Twitter posts. Crisis 2017;38:319–29. [PubMed: 28228065]
36. Young JT, Borschmann R, Heffernan E, et al. Contact with mental health services after acute care for self-harm among adults released from prison: a prospective data linkage study. Suicide Life Threat Behav 2020;50:990–1006. [PubMed: 32359122]
37. Kessler RC, Bauer MS, Bishop TM, et al. Using administrative data to predict suicide after psychiatric hospitalization in the Veterans health administration system. Front Psychiatry 2020;11:390. [PubMed: 32435212]
38. Arendt F, Markiewitz A, Scherr S. Investigating suicide-related Subliminal messages on Instagram. Crisis 2021;42:1–7. [PubMed: 32781896]
39. Gradus JL, King MW, Galatzer-Levy I, et al. Gender differences in machine learning models of trauma and suicidal ideation in veterans of the Iraq and Afghanistan wars. J Trauma Stress 2017;30:362–71. [PubMed: 28741810]
40. Levis M, Leonard Westgate C, Gui J, et al. Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models. Psychol Med 2020:1382–91.
41. Kessler RC, Bernecker SL, Bossarte RM. The role of big data analytics in predicting suicide. In: Personalized psychiatry. Springer, 2019: 77–98.
42. Lin G-M, Nagamine M, Yang S-N, et al. Machine learning based suicide ideation prediction for military personnel. IEEE J Biomed Health Inform 2020;24:1907–16. [PubMed: 32324581]
43. Rozek DC, Andres WC, Smith NB, et al. Using machine learning to predict suicide attempts in military personnel. Psychiatry Res 2020;294:113515. [PubMed: 33113452]

44. Borschmann R, Young JT, Moran P, et al. Accuracy and predictive value of incarcerated adults' accounts of their self-harm histories: findings froman Australian prospective data linkage study. CMAJ Open 2017;5:E694–701.
45. Haroz EE, Walsh CG, Goklish N, et al. Reaching those at highest risk for suicide: development of a model using machine learning methods for use with Native American communities. Suicide Life Threat Behav 2020;50:422–36. [PubMed: 31692064]
46. Murphy B, Kennedy B, Martin C, et al. Health and care related risk factors for suicide among nursing home residents: a data linkage study. Suicide Life Threat Behav 2019;49:695–706. [PubMed: 29665103]
47. Choi D, Sumner SA, Holland KM, et al. Development of a machine learning model using multiple, heterogeneous data sources to estimate weekly US suicide fatalities. JAMA Netw Open 2020;3:e2030932. [PubMed: 33355678]
48. Ahmad FB, Cisewski JA, Miniño A, et al. Provisional Mortality Data - United States, 2020. MMWR Morb Mortal Wkly Rep 2021;70:519–22. [PubMed: 33830988]
49. Vioules MJ, Moulahi B, Azé J, et al. Detection of suicide-related posts in Twitter data streams. IBM J Res Dev 2018;62:7:1–7:12.
50. Cohen J, Wright-Berryman J, Rohlfs L, et al. A feasibility study using a machine learning suicide risk prediction model based on open-ended interview language in adolescent therapy sessions. Int J Environ Res Public Health 2020;17:8187.
51. Coppersmith G, Leary R, Crutchley P, et al. Natural language processing of social media as screening for suicide risk. Biomed Inform Insights 2018;10:1178222618792860.
52. Large M, Kaneson M, Myles N, et al. Meta-Analysis of longitudinal cohort studies of suicide risk assessment among psychiatric patients: heterogeneity in results and lack of improvement over time. PLoS One 2016;11:e0156322. [PubMed: 27285387]
53. Berman AL, Carter G. Technological advances and the future of suicide prevention: ethical, legal, and empirical challenges. Suicide Life Threat Behav 2020;50:643–51. [PubMed: 31803971]
54. Moreno MA, Goniu N, Moreno PS, et al. Ethics of social media research: common concerns and practical considerations. Cyberpsychol Behav Soc Netw 2013;16:708–13. [PubMed: 23679571]

What is already known on this subject

- Data science is a useful tool for efficiently analysing data and predicting outcome.
- Data science has been applied to injury prevention research.

What this study adds

- This study highlights areas where data science may be useful for future injury prevention research including:
  - Reasons for applying data science.
  - Common data science methods used.
  - Types of data sources harnessed for applying data science to injury prevention research.

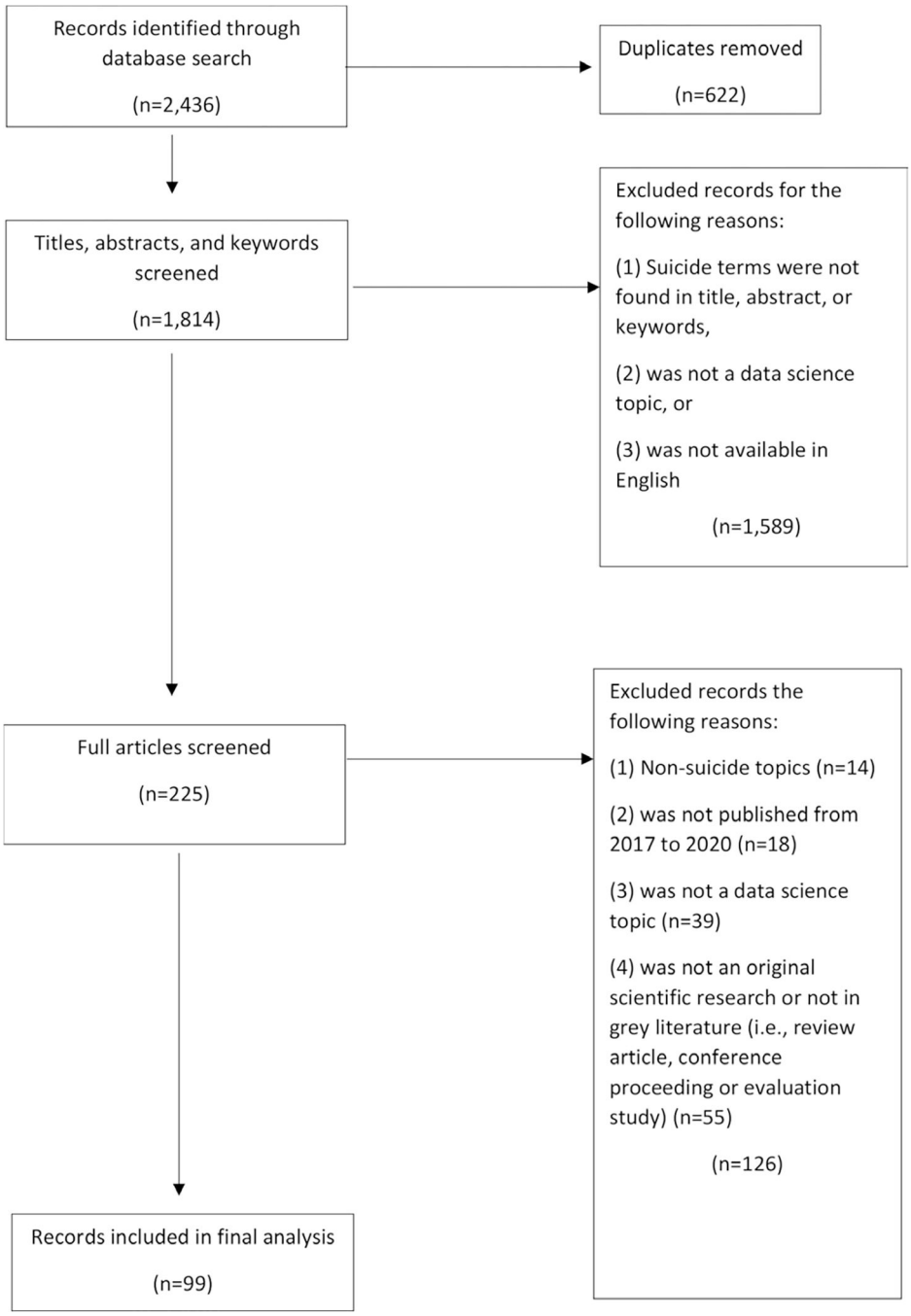

Figure 1.
Literature review screening process.

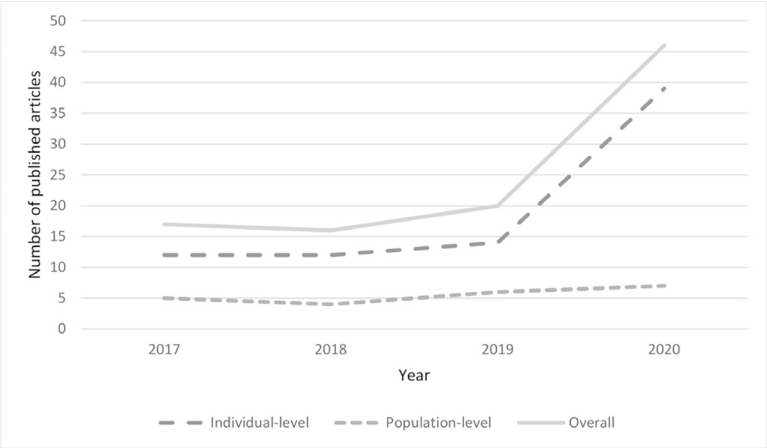

Figure 2.
Counts of individual-level, population-level and overall articles published in each year.

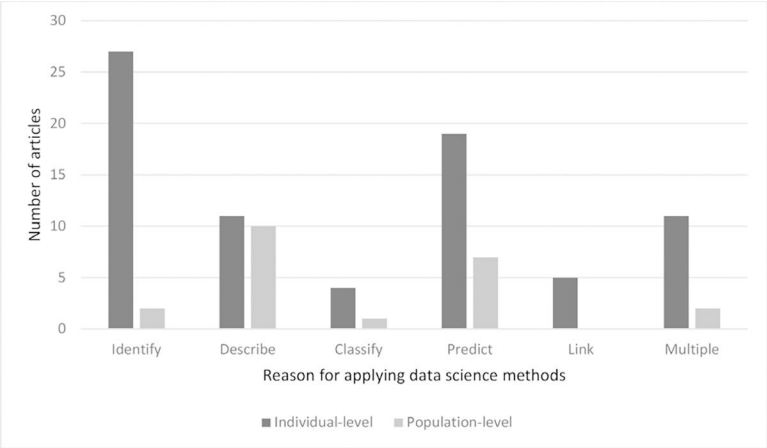

Figure 3.

Counts of individual-level and population-level articles categorised by reason for applying data science methods. ‘Multiple’ refers to any article with more than one reason for applying data science methods. These counts are not included in the other groups.

Table 1

Keyword search terms and results by database

Keyword search terms	Database	Unique records
Big Data/OR data science/OR (data science* OR data analy* OR data driven OR data mining OR data sharing OR administrative data OR machine learning OR artificial intelligence OR natural language processing OR text mining OR data visualization OR deep learning OR computer vision OR data linkage OR predictive modeling OR forecast* OR nowcast* OR agent-based modeling OR data engineering OR advanced computing OR (statistic* ADJ3 model*) OR domain expertise OR social media OR facebook OR twitter OR instagram OR reddit OR youtube OR tumblr) AND Accident Prevention/OR Suicide/OR homicide/OR alcohol drinking/OR Alcoholism/OR Substance-Related Disorders/OR Alcoholism/OR (Accident* OR injury OR injuries OR violence OR suicide* OR homicide* OR TBI OR falls OR drowning* OR seat belt* OR seatbelt* OR firearm* OR guns OR poison* OR binge drinking OR drunk OR substance abuse* OR substance use* OR drug abuse* OR drug use* OR overdose*) AND (Prevent* OR intervention OR mitigation OR communication OR messaging OR message* OR education OR health promotion OR control OR warning* OR public service announcement* OR PSA OR detect* OR risk* OR awareness).	Medline (OVID) PsycInfo (OVID) Scopus	1,042 124 648

Table 2

Individual-level and population-level articles by dimensions of articles

Dimensions of articles	No. of individual-level articles	No. of population-level articles	Total no. of articles
Type of data science method*			
Application of novel data	13	9	22
Machine learning	35	11	46
NLP/linguistic analysis	8	1	9
Data mining	2	2	4
Data linkage	7	0	7
Data source*			
Medical record	24	2	26
Survey/questionnaire	28	2	30
Administrative	10	1	11
Social media and web-based behaviour	28	17	45
Population based/aggregated	6	14	20
Internet of Things	2	0	2
Focus on disproportionately affected populations			
Veterans/military personnel	6	0	6
Formerly incarcerated people	3	0	3
Native American people	1	0	1
Older adults living in nursing homes	1	0	1

*Some articles applied multiple types of data science methods or data source and are included in multiple categories.

NLP, natural-language processing.