Can the Integrated and Selected Pattern Approach Effectively Predict Suicide Rates? A Study Using Internet Search Queries

Author

Institution

November 24, 2024

Abstract

Background: Suicide prevention has become a global public health and political issue. There have been active attempts to predict suicides (suicide rates and suicide numbers) using search queries from search engines.
Methods: This study used spike-and-slab regression which is one of the sparse variable selection techniques, and we called this approach the ‘integrated and selected pattern approach.’ we used monthly national suicide data and relative search volume (RSV) of 51 search queries from Google Trends for the United States (US) and Japan from 2004 to 2019, and compared the accuracy of models using this approach against those using other approaches.
Results: The model employing the integrated and selected pattern approach demonstrated the highest accuracy and stability of predicted values in predicting suicide rates and the number of suicides in the US. However, in the case of Japan, the accuracy and stability of the predicted values for this approach were lower than other approaches. Furthermore, even in the US, where the integrated and selected pattern approach outperforms other approaches at the country-level, it does not consistently outperform other models across all data subsets.
Limitations: The integrated and selected pattern approach is not universally the most efficient method for predicting suicides across all linguistic, cultural contexts, and demographic groups. It remains crucial to emphasize the importance of comparing its efficacy with several other approaches in practical implementations.
Conclusion: The integrated and selected pattern approach is one of the potentially effective methods for predicting suicides.

Keywords

Suicide, Public mental health, Suicide prevention, Internet search queries, Google trends, Sparse variable selection, Spike and slab regression

1 Background

Over 700,000 people worldwide attempt suicide annually (World Health Organization, 2023). Suicide prevention has become a global public health and policy issue (Arensman et al., 2020; World Health Organization, 2014, 2012). Governments and various stakeholders (e.g., police, hospitals, and NGOs) in different countries are tackling this issue by implementing programs to improve public mental health for suicide prevention (World Health Organization, 2014, 2012).

As a first step in addressing this issue and in assessing the impact of implementation, monitoring suicides (suicide rates and suicide numbers) in real time is essential for stakeholders. In the public health model for suicide prevention developed by WHO (2014), monitoring the current situation is essential for identifying risks, developing interventions, and evaluating and improving implementation.

Although monitoring the current situation is essential for suicide prevention, this is difficult in practice. Public reporting of suicides is often delayed by one to two years (Kandula et al., 2023; Kristoufek et al., 2016), and suicides are often underreported due to political or procedural reasons (World Health Organization, 2014). Therefore, predicting the current status of public mental health and suicides in the population is a challenge for suicide prevention policy and research.

Several previous studies have proposed solutions to this problem. Many of these solutions utilize currently available data to predict the status of public mental health and suicides. Among these approaches, since the 2010s, with the expansion of the internet user base and the increased ease of data collection, there have been active attempts to predict the status of public mental health and suicides utilizing search queries from search engines such as Google Trends (Adam-Troian and Arciszewski, 2020; Barros et al., 2019; Burnett et al., 2020; Chang et al., 2011; Ekinci et al., 2023; Gunn and Lester, 2013; Halford et al., 2020; Jimenez et al., 2020; Knipe et al., 2021; Kristoufek et al., 2016; McCarthy, 2010; Page et al., 2011; Sueki, 2011; Taira et al., 2021; Tran et al., 2017; Yang et al., 2011). These studies demonstrated the effectiveness and potential of utilizing search query data as a valuable tool for understanding and predicting the status of public mental health and suicides in real time.

Based on the use of search queries, these studies can be categorized into two main pattern approaches. The first pattern is the ‘selected pattern approach.’ Studies that follow this pattern approach utilize a small number of search queries that directly represent the phenomenon in which the author is trying to predict the outcomes (Chang et al., 2011; Gunn and Lester, 2013; Knipe et al., 2021; Kristoufek et al., 2016; McCarthy, 2010; Page et al., 2011; Sueki, 2011; Taira et al., 2021). An early study showed that the correlation between Google searches for ‘depression’ and suicidal deaths was statistically significant (Sueki, 2011). Another study using statistical models showed that a model using Google search queries for ‘suicide’ and ‘depression' outperformed a basic structural time-series model using only trends and seasonality to predict suicides (Kristoufek et al., 2016). Some studies identified patterns between search queries representing specific mental states, suicide rates, and mental health indicators. These search queries include terms like ‘ how to suicide' and ‘ hydrogen sulfide,’ which refer to suicide methods, as well as ‘ suicide ideation,’ ‘depression,’ or ‘loneliness,’ suggesting their potential utility in prediction (Chang et al., 2011; Gunn and Lester, 2013; Knipe et al., 2021).

The second pattern is the ‘ integrated pattern approach.’ Studies that follow this pattern approach use a large number of search queries that have been utilized in previous studies trying to predict the outcomes (Adam-Troian and Arciszewski, 2020; Barros et al., 2019; Ekinci et al., 2023; Halford et al., 2020; Jimenez et al., 2020; Tran et al., 2017; Yang et al., 2011). Some studies have focused on exploring and identifying search queries that exhibit high correlation coefficients with suicides (Ekinci et al., 2023; Jimenez et al., 2020; Yang et al., 2011). Meanwhile, studies like Barros et al. (2019) adopted a more focused approach. They analyzed the correlations between numerous individual search queries, their combinations, and suicide rates, then explored the inclusion or exclusion of these queries in their model. By selecting only those queries that demonstrated high predictive power for suicide rates, they effectively predicted suicide rates. There are also studies, such as Adam-Troian and Arciszewski (2020), which show a correlation between a large number of words related to a particular domain (e.g., absolutist thinking) and suicides.

While studies following these two pattern approaches have contributed to the development of solutions for monitoring the status of public mental health and suicides in real time, these two approaches pose challenges of their own. Studies that employ the selected pattern approach tend to utilize search queries that directly represent the phenomenon that the author is attempting to predict (e.g., ‘ suicide’ and ‘ depression’). Consequently, there is a risk of overlooking search queries that do not explicitly represent the phenomenon, but are nonetheless valuable for generating predictions, such as ‘allergy’ in Jimenez et al. (2020)). However, studies that employ the integrated pattern approach also face the risk of correlations between search queries (McCarthy, 2010), which may lead to a decrease in the accuracy of predictions. To address this problem, several studies have examined the correlations between individual search queries and outcomes (Ekinci et al., 2023; Jimenez et al., 2020; Yang et al., 2011), while others have explored manually including or excluding search queries in the model and selecting only those queries that demonstrated high predictive power for outcomes (Barros et al., 2019). Nevertheless, both approaches encounter practical challenges. As the number of search queries increases, the combinations to be verified grow exponentially, making the process increasingly inefficient and limiting its practical utility. Therefore, it is necessary to develop an approach that combines the advantages of both selected and integrated pattern approaches while addressing their respective limitations.

In this study, we aimed to introduce the ‘integrated and selected pattern approach’ for studies that utilize search queries from search engines to predict suicides. This approach aims to bridge the selected and the integrated pattern approaches, addressing the limitations of each while complementing their respective strengths, to enable more accurate predictions of suicides. This study utilized monthly national suicide data and the relative search volume (RSV) of 51 search queries from Google Trends for the United States (US) and Japan from 2004 to 2019. Utilizing these data, this study verifies the effectiveness of the model constructed using a sparse variable selection technique called spike-and-slab regression (Ishwaran et al., 2010; Ishwaran and Rao, 2005; Scott and Varian, 2014) as the integrated and selected pattern approach.

2 Materials and Methods

2.1 Data collection

This study utilizes two distinct datasets originating from two separate countries, the US and Japan, to investigate and validate the effectiveness of the integrated and selected pattern approach across different linguistic and cultural contexts.

Previous studies have widely used suicide rates and the number of suicides as proxies for public mental health status (Arensman et al., 2020; Barros et al., 2019; Ekinci et al., 2023; Jimenez et al., 2020; Kristoufek et al., 2016; Sueki, 2011; Taira et al., 2021; Tran et al., 2017; World Health Organization, 2014, 2012; Yang et al., 2011). Furthermore, there is a considerable societal demand for accurate predictions of these metrics, as there is typically a one to two year lag before the final fixed values become available (Kandula et al., 2023; Kristoufek et al., 2016). Therefore, this study predicted the monthly suicide rates and the number of suicides at the country-level in both countries.

In the case of the US, this study counted the monthly deaths caused by ICD-10 codes U03, X60-X84, and Y87.0 (Garnett et al., 2022) from the Mortality Multiple Cause Files (Centers for Disease Control and Prevention), excluding those whose place of residence was outside the U.S., as the number of suicides. The study then utilized national monthly population data from the Population Estimates Program (United States Census Bureau) to calculate the monthly national suicide rates (the number of suicides per 100,000 people).

In the case of Japan, this study counted the monthly deaths confirmed as suicides by the National Police Agency (Ministry of Health, Labour and Welfare, a) [1]. The study then utilized the national monthly population data from the population estimates[2] (Ministry of Internal Affairs and Communications) to calculate the monthly national suicide rates.

For search query data from search engines, this study utilized the relative search volume (RSV) of 51 search queries in Google Trends (Google). The RSV scales (from 0 to 100) the volume of each search query (e.g., ‘suicide’ and ‘depression’) during the target period. Google has the largest share of Internet searches in both the US and Japan (StatCounter); therefore, utilizing RSV in Google Trends as a proxy for Internet searches in the population is suitable for both countries.

Table 1 lists the 51 search queries used in this study. These queries have been utilized in two or more previous studies (Adam-Troian and Arciszewski, 2020; Barros et al., 2019; Burnett et al., 2020; Chang et al., 2011; Ekinci et al., 2023; Gunn and Lester, 2013; Halford et al., 2020; Jimenez et al., 2020; Knipe et al., 2021; Kristoufek et al., 2016; McCarthy, 2010; Page et al., 2011; Sueki, 2011; Taira et al., 2021; Yang et al., 2011).

[1] There are two statistics related to monthly suicides in Japan. One is Suicide Statistics. In this statistic, a suicide is determined when the National Police Agency confirms that a death was caused by suicidal behavior, and it includes foreigner suicides in Japan (Ministry of Health, Labour and Welfare, b). The proportion of foreigner deaths by suicide out of the total suicides was approximately 1% from 2008 to 2019 in Japan (Ministry of Health, Labour and Welfare, c). While Suicide Statistics are published monthly on a monthly basis in a timely manner, it should be noted that these figures are provisional. The final fixed values are not determined until the end of the year, which creates a time lag in accurately assessing suicide rates. The other statistic is Vital Statistics, which counts the monthly deaths caused by ICD-10 codes X60-X84 as suicides (Ministry of Health, Labour and Welfare, d), and it does not include foreigner suicides in Japan (Ministry of Health, Labour and Welfare, b). Because the monthly suicide numbers in Suicide Statistics eventually become fixed values (while Vital Statistics are not), and the proportion of foreigner deaths by suicide out of the total suicides is small, this study utilizes Suicide Statistics in the Japanese case.

[2] Because suicides in Suicide Statistics include foreigner suicides in Japan, this study utilizes population data that also includes the foreign population in Japan.

For the Japanese case, we translated them from English into Japanese and extracted the RSV. When search queries resulted in duplicates after translation (e.g., ‘suicide’ and ‘kill yourself’ translated to the same word ‘jisatsu’), they were treated as single search queries. Consequently, in the case of Japan, we used 46 search queries.

Finally, regarding the data period for the US cases, this study utilized data from January 2004, the earliest period for which RSV was available, to December 2017 as a training set for the learning model. The last 12 months of the training set were used solely for grid search. The test set was used to validate the accuracy of the predicted values of the outcomes and consisted of data from January 2018 to December 2019, the latest period before the COVID-19 pandemic. In Japanese cases, this study used data from January 2008, the earliest period for which monthly suicides are available, to December 2017 as a training set for the learning model. As in the US case, the last 12 months of the training set were utilized solely for the grid search. The test set comprised data collected between January 2018 and December 2019.

Table 1: Search Queries in Google Utilized in Model (2) - Model(4) in This Study.

Variables	Search queries (Japanese)
SRSV	suicide (jisatsu), depression (utsu)
IRSV	suicide (jisatsu), depression (utsu), abuse (gyakutai), suicide methods (jisatsu houhou), commit suicide (jisatsu) divorce (rikon), anxiety (fuan), unemployment (shitsugyou), bipolar disorder (soukyokusei shougai), major depression (utsubyou) loneliness (kodoku), insomnia (fumin), marriage (kekkon), relationship breakup (shitsuren), cancer (gan) headache (zutsuu), how to commit suicide (jisatsu douyatte), asthma (zensoku), anxiety disorder (fuan shougai), antidepressant (koututsuzai) social welfare (shakai fukushi), allergy (arerugi), pain (itami), schizophrenia (tougoushichoushou), stress (sutoresu) alcohol (sake), teen suicide (wakamono jisatsu), suicide help (jisatsu tasuke), job (shigoto), suicide ideation (jisatsu ganbou) suicide hotline (jisatsu hottorain), drunkenness (nomisungi), hanging (kubitsuri), severe depression (utsubyou), illicit drugs (ihou yakubutsu) suicide attempt (jisatsu misui), suicidal thoughts (jisatsu kangae), social benefits (shakai fukushi), how to kill yourself (jisatsu houhou), hydrogen sulfide (ryuuka suiso) stock market (kabushiki shijou), domestic violence (DV: diibui), hypnotics (suiminyaku), charcoal burning (rentan), chronic illness (mansei shikkan) kill yourself (jisatsu), suicide prevention (jisatsu yobou), lawsuit (soshou), alcohol abstinence (kinshu), religious belief (shinkou) manic depression (souutsubyou)

2.2 Model Comparison

This study compares model (1) to (3).

$$
\begin{aligned}
y_t &= \alpha_t + \tau_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \\
\alpha_t &= \alpha_{t-1} + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2) \\
\tau_t &= -\sum_{s=t-11}^{t-1} \delta_s + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2)
\end{aligned}
\tag{1}
$$

First, model (1) is a basic structural time series model that does not utilize the RSV of search queries but is composed only of trend and seasonality, serving as a naive model in this study. y_t is the scaled monthly suicide rates, where the minimum value was set to 0 and the maximum value was set to 100 in the training set. We then excluded the last 12 months of the training set (from January 2017 to December 2017) from all scaling because these months were utilized for model parameter choice in the grid search. α_t is a local level trend assumed to follow a random walk, and τ_t is seasonality with a 12-month cycle.

$$
\begin{aligned}
y_t &= \alpha_t + \tau_t + \beta_{it}\mathrm{SRSV}_{it} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \\
\alpha_t &= \alpha_{t-1} + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2) \\
\tau_t &= -\sum_{s=t-11}^{t-1} \delta_s + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2) \\
\beta_{it} &= \beta_{it-1} + \omega_{it}, \quad \omega_{it} \sim N\left(0, \sigma_{\omega i}^2 / \sigma_{\mathrm{SRSV}i}^2\right) \\
&\quad 1/\sigma_{\omega i}^2 \sim Ga(a, b) \\
&\quad \sqrt{b/a} \sim Ga(1, 0.01\sigma_y) \\
&\quad a \sim Ga(1, 10)
\end{aligned}
\tag{2-1}
$$

Model (2-1) describes the selected pattern approach. This model adds the SRSV_{it} to the basic structural time-series model (1). As described in Table 1, this model utilizes the RSV of two search queries, ‘suicide’ and ‘depression,’ which have been most commonly used in previous studies to directly represent phenomena related to public mental health and suicides (Barros et al., 2019; Burnett et al., 2020; Chang et al., 2011; Ekinci et al., 2023; Halford et al., 2020; Jimenez et al., 2020; Knipe et al., 2021; Kristoufek et al., 2016; McCarthy, 2010; Sueki, 2011; Tran et al., 2017; Yang et al., 2011). Similar to y_t, SRSV_{it} was also scaled, with the minimum value set to 0 and the maximum value set to 100 in the training set. Regarding β_{it}, which are the coefficients of SRSV_{it}, and referring to Scott (2024), this study introduces the assumption that the coefficients change over time follow a random walk and that $\sigma_{\omega i}^2$, which is the variance of β_{it} scaled by the variance of SRSV_i, have a prior distribution determined by the gamma distribution. This assumption could be interpreted as assuming that the coefficients of SRSV_{it} change over time while preventing the change from being too large.

$$
\begin{aligned}
y_t &= \alpha_t + \tau_t + \beta_{it}\mathrm{IRSV}_{it} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \\
\alpha_t &= \alpha_{t-1} + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2) \\
\tau_t &= -\sum_{s=t-11}^{t-1} \delta_s + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2) \\
\beta_{it} &= \beta_{it-1} + \omega_{it}, \quad \omega_{it} \sim N\left(0, \sigma_{\omega i}^2 / \sigma_{\mathrm{IRSV}i}^2\right) \\
&\quad 1/\sigma_{\omega i}^2 \sim Ga(a, b) \\
&\quad \sqrt{b/a} \sim Ga(1, 0.01\sigma_y) \\
&\quad a \sim Ga(1, 10)
\end{aligned}
\tag{2-2}
$$

Model (2-2) describes the integrated pattern approach: This model sets the structure of the trend and seasonality to be the same as in model (2-1). However, unlike model (2-1), this model utilizes IRSV_{it}, which represents the RSV of the 51 search queries employed in two or more previous studies, as described in Table 1. The assumptions for the coefficients of IRSV_{it} are the same as those for the coefficients of SRSV_{it} in model (2-1), where the coefficients change over time according to a random walk. In other words, this model increases the dimensions of model (2-1) by incorporating a larger number of search queries.

$$
\begin{aligned}
y_t &= \alpha_t + \tau_t + \beta \mathrm{IRSV}_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \\
\alpha_t &= \alpha_{t-1} + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2) \\
\tau_t &= -\sum_{s=t-11}^{t-1} \delta_s + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2) \\
p(\beta, \gamma, \sigma_\varepsilon^2) &= p(\beta_\gamma | \gamma, \sigma_\varepsilon^2) p(\sigma_\varepsilon^2 | \gamma) p(\gamma) \\
\gamma &\sim \prod_{i=1}^{K} \pi_i^{\gamma_i} (1 - \pi_i)^{1-\gamma_i} \\
\beta_\gamma &| \sigma_\varepsilon^2, \gamma \sim N(0, \sigma_\varepsilon^2 (\Omega_\gamma^{-1})^{-1}) \\
1/\sigma_\varepsilon^2 &\sim Ga(\nu/2, ss/2)
\end{aligned} \tag{3}
$$

Finally, model (3) expresses the integrated and selected pattern approach, which this study introduces as a novel method. This model assumes that the structure of the trend and seasonality are the same as those in the other models. In addition, this model utilizes IRSV_{it}, similar to model (2-2).

However, the coefficients of the search queries in this model differ from those in other models. In this model, this study introduces one of the sparse variable selection techniques called ‘ spike ’ and ‘ slab ’ regression, which was proposed by Ishwaran et al. (2010), Ishwaran and Rao (2005) or Scott and Varian (2014). $p(\gamma)$ is the ‘ spike ’ part in the model. This is determined by the Bernoulli prior, and by setting the parameter $\pi = p/K$ (where K is the total number of coefficients of variables and p is the number of nonzero coefficients), it controls the complexity of the models by setting almost all coefficients of variables to zero (Scott and Varian, 2014). This study conducts a grid search over $\pi = 0.05, 0.1, 0.15$, and 0.20 to determine the optimal proportion of nonzero coefficients for IRSV_{it}.

If the coefficients are not zero, the second to last row and the last row of the equation for model (3) describes the prior distribution of the coefficient. This is the ‘ slab ’ part of the model. In this part, the coefficients are determined by very weakly informative distributions $N(0, \sigma_\varepsilon^2 (\Omega_\gamma^{-1})^{-1})$, where Ω_γ^{-1} denotes that the rows and columns of $\Omega^{-1} = \kappa \left(\omega X^T X + (1-\omega) \mathrm{diag}(X^T X) \right) / n$ correspond to $\gamma_i = 1$ (Scott and Varian, 2014). Here, X (IRSV in this model) denotes the design matrix. In addition, the ss in the prior distribution of inverse σ_ε^2 is determined by the prior set weight ν, the prior set expected R^2 of the regression, and the marginal standard deviation of y (s_y^2), where the relationship is given by $ss/\nu = (1 - R^2) s_y^2$ (Scott and Varian, 2014). Given the limited prior information available regarding the underlying features of the data used in this study, we adopt relatively neutral priors following Scott (2024) and Scott and Varian (2014): $\kappa = 1$, $\omega = 1/2$, $\nu = 0.01$, $R^2 = 0.5$ as prior set values.

In summary, model (3) combines the aspects of both the integrated and selected pattern approaches. On the one hand, it utilizes a large number of search queries for prediction, similar to the integrated pattern approach. However, it simplifies the model, akin to the selected pattern approach.

This model offers four advantages. First, it prevents the risk of overlooking search queries that do not explicitly represent the phenomenon but are valuable for generating predictions. Second, it mitigates the risk of decreased model accuracy due to the use of high-dimensional models. Third, it is more efficient and practical than manual variable selection approaches, such as stepwise approach (Chatterjee and Hadi, 2006; Yanti and

Table 2: Mean Absolute Percentage Error (MAPE) for Suicide Rates and Mean Absolute Error (MAE) for Suicide Numbers in Test Set (Jan 2018 - Dec 2019) for the United States and Japan

	MAPE (Suicide Rates)	**MAE (Suicide Numbers)**
US		
Model (1)	5.86	238.09
Model (2-1)	6.59	267.04
Model (2-2)	3.43	139.61
Model (3)	3.25	133.30
Japan		
Model (1)	5.15	86.94
Model (2-1)	4.97	83.80
Model (2-2)	8.84	147.71
Model (3)	5.63	94.74

Rahardiantoro, 2019; Zhou et al., 2012), as the 'spike' component automatically eliminates unnecessary variables from the model. Moreover, in contrast to alternative variable selection techniques such as Lasso (Li and Chen, 2014; Nguyen and Braun, 2018; Tibshirani, 1996), which assume stationarity and frequently necessitate preprocessing (e.g., differencing) for non-stationary data, this model can directly incorporate trend and seasonal components, thus preventing information loss and the additional burden of transforming non-stationary data into stationary[3].

The parameters and predicted values in all models were calculated using Markov chain Monte Carlo methods (MCMC). This study utilize R package ‘ bsts (version:0.9.10) ’ (Scott, 2024) run MCMC with seed 42.

In this study, a grid search was conducted to determine optimal model parameters. The number of iterations varied between 2,000, 4,000, 8,000, and 16,000, whereas the burn-in period was set to 10%, 20%, 30%, or 40% of the total iterations. For all grid-search evaluations, the last 12 months (January 2017 to December 2017) of the training set were used. The optimal model is selected based on the results of a comprehensive grid search [4].

3 Results

Table 2 presents the accuracy of models (1)-(3) for the US and Japanese test sets, showing the mean absolute percentage error (MAPE) for suicide rates and the mean absolute error (MAE) for suicide numbers.

For the US case, model (3), which uses the integrated and selected pattern approach, is the most optimal among the four models. The model outperformed both the selected pattern approach, which utilized a smaller number of search queries directly related to public mental health status and suicides, and the integrated pattern approach, which utilized the same large number of search queries. It had a MAPE of 3.25% for suicide rates and an MAE of 133.30 for suicide numbers in the test set.

[3] In particular, for the suicide rates used in this study, the Augmented Dickey-Fuller test results shown in STable1 suggest that the US data exhibits non-stationarity.

[4] All of the code and the results of the grid search can be replicated using the materials available on the author's GitHub [URL here].

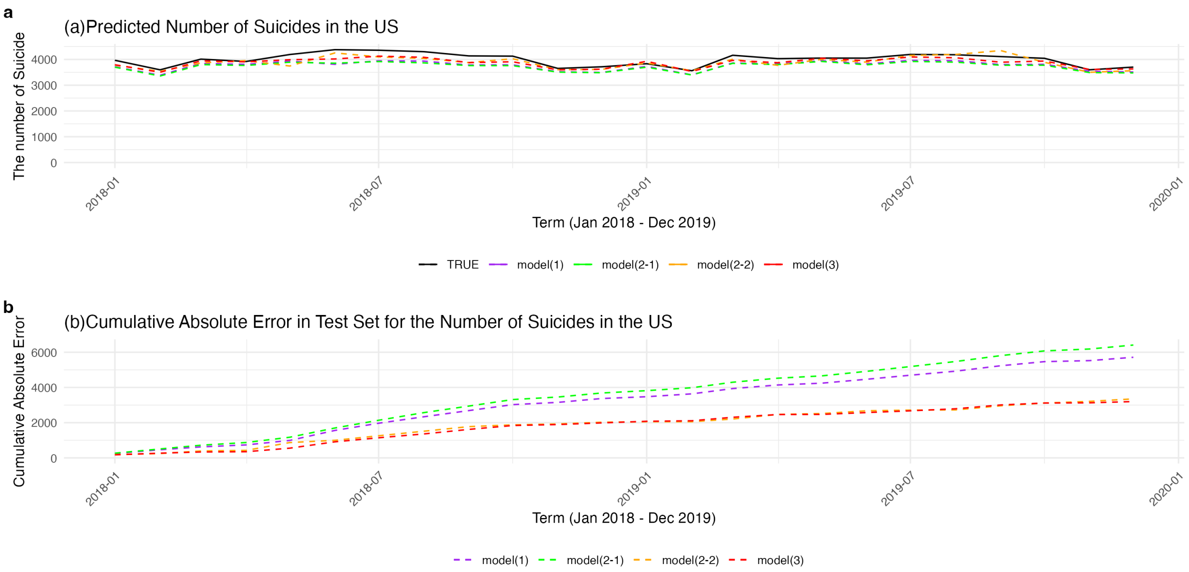

Figure 1: Accuracy of Models for United States Test Set (Jan 2018 - Dec 2019): (a) Trend of Actual Suicide Numbers and Predicted Values for Model (1) - Model (3), (b) Cumulative Absolute Error (AE) for Model (1) - Model (3)

Regarding the stability of the predicted values from the models, Figure 1(a) plots the trend of the actual suicide numbers in the US test set and the predicted values from models (1)-(3), whereas Figure 1(b) plots the cumulative absolute error for the test set. The accuracy of the predicted values from model (3) is remarkably more stable than that of model (1), which uses basic time-series analysis, and model (2-1), which employs the selected pattern approach. It is also slightly more stable than model (2-2), which uses an integrated pattern approach. This stability can be interpreted as consistent accuracy across the entire range of data points in the test set rather than exhibiting high accuracy only at particular points. Therefore, model (3), which uses the integrated and selected pattern approach, is not only the most accurate, but also the most stable model among the four.

Figure 2 plots the inclusion probabilities of search queries in model (3). While the inclusion probabilities of most search queries are approximately zero, the top-ranking search query with an exceptionally high probability of 0.97 is ‘ domestic violence,’ and ‘ pain’ also shows a relatively high probability, ranking in the top 3. This indicates that incorporating search queries that do not directly represent the status of public mental health or suicides, while simultaneously simplifying the model, can contribute significantly to predicting suicides.

In practice, to mitigate the risk of point estimates underestimating (as shown in Figure 1(a)) or overestimating true values, it is necessary to interpret predicted values with their associated uncertainty intervals, as illustrated in SFigure 3.

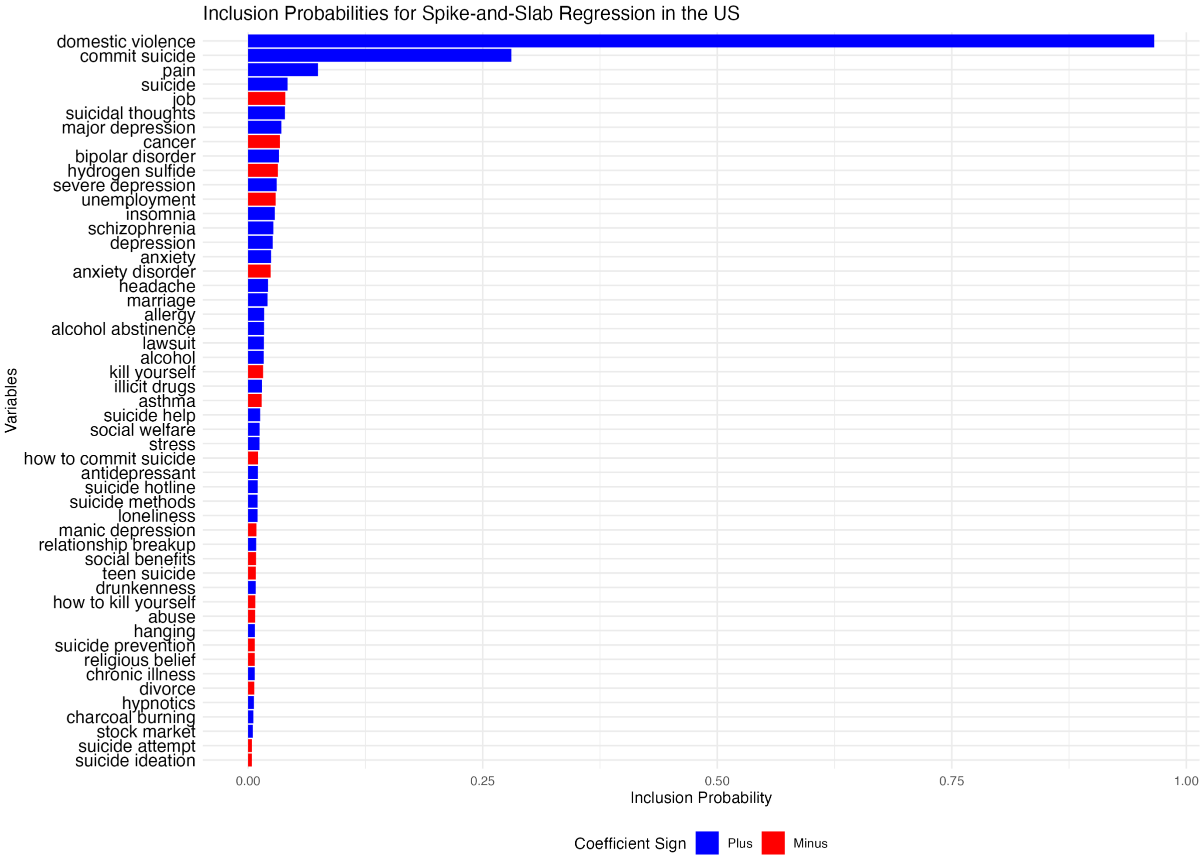

Figure 2: Inclusion Probabilities of Search Queries in Model (3) Using Spike-and-Slab Regression for United States

In contrast, unlike the prediction results for the US case, model (3) does not emerge as the optimal model among the four in the Japanese case. The accuracy of the predicted values for model (3) does not outperform that of model (2-1), which employs the selected pattern approach, or that of model (1), which utilizes the basic structural time-series model as a naive model. In addition, as shown in Figures 3(a) and 3(b), model (3) did not outperform model(1) and model (2-1) in terms of the accuracy stability for the predicted values. The accuracy of model (3) did not deteriorate at any particular point in the test set; rather, it exhibited consistently lower performance across the entire test set.

Finally, Table 3 presents the results of verifying the accuracy of models (1)-(3) by further disaggregating the suicide rates for two countries used in the main verification into male and female suicide rates. These results indicate that the integrated and selected pattern approach demonstrates varying effectiveness across different data subsets, even within the same countries. Even in a country like the US, where model (3), which uses the integrated and selected pattern approach, outperforms other models at the country-level, it does not consistently outperform other models across all data subsets, such as female suicide rates in the US. Conversely, there are cases like female suicide rates in Japan where model (3) outperforms other models, even when it does not perform better than other models at the country-level.

In summary, the objective of the integrated and selected pattern approaches, which combines the advantages of both the selected and integrated pattern approaches while addressing their respective limitations, was achieved in several cases. However, the approach exhibits fragility in accuracy depending on the case, indicating that its robustness remains a significant limitation.

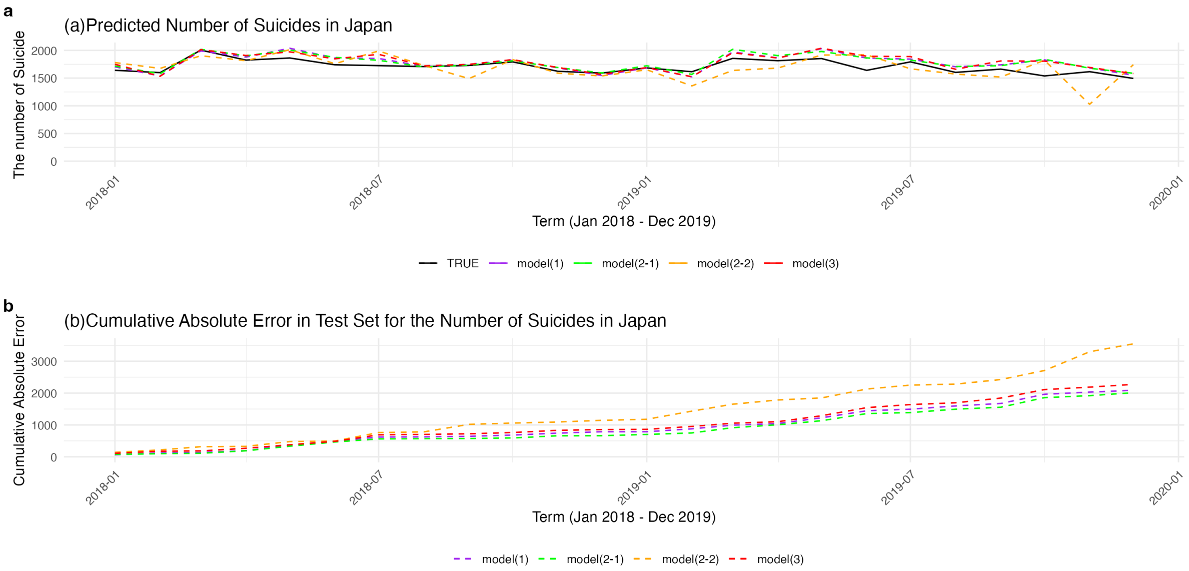

Figure 3: Accuracy of Models for Japanese Test Set (Jan 2018 - Dec 2019): (a) Trend of Actual Suicide Numbers and Predicted Values for Model (1) - Model (3), (b) Cumulative Absolute Error (AE) for Model (1) - Model (3)

Table 3: Mean Absolute Percentage Error (MAPE) for Suicide Rates in Test Set (Jan 2018 - Dec 2019) for the United States and Japan

	All	**Male**	**Female**
US			
Model (1)	5.86	6.89	3.89
Model (2-1)	6.59	7.84	**3.85**
Model (2-2)	3.43	4.40	4.55
Model (3)	**3.25**	**3.77**	3.89
Japan			
Model (1)	5.15	5.29	6.40
Model (2-1)	**4.97**	**5.08**	6.44
Model (2-2)	8.84	10.26	10.06
Model (3)	5.63	5.55	**6.16**

4 Conclusions

This study attempted to predict country-level suicides (suicide rates and suicide numbers) using suicide data from the US and Japan and the RSV of 51 search queries in Google Trends.

Although many studies have predicted the status of public mental health and suicides using search queries (Adam-Troian and Arciszewski, 2020; Barros et al., 2019; Burnett et al., 2020; Chang et al., 2011; Ekinci et al., 2023; Gunn and Lester, 2013; Halford et al., 2020; Jimenez et al., 2020; Knipe et al., 2021; Kristoufek et al., 2016; McCarthy, 2010;

Page et al., 2011; Sueki, 2011; Taira et al., 2021; Tran et al., 2017; Yang et al., 2011), this study introduces a new approach to this body of work. Previous studies mainly employed either the ‘ selected pattern approach, ’ which utilized few search queries directly representing the status of public mental health and suicides, and the ‘ integrated pattern approach, ’ which utilized a large number of search queries. However, this study introduced the ‘integrated and selected pattern approach, ’ combining the advantages of both while addressing their respective limitations. This approach utilizes a large number of potentially useful search queries for prediction while mitigating two risks: overlooking search queries that do not explicitly represent the phenomenon but are valuable for generating predictions, and the decreased model accuracy typically associated with high-dimensional models. This was achieved by employing spike-and-slab regression, a sparse variable selection technique (Ishwaran et al., 2010; Ishwaran and Rao, 2005; Scott and Varian, 2014), to select search queries that were efficient for prediction.

This study demonstrates that the model using the integrated and selected pattern approach has the highest accuracy and stability of predicted values in predicting suicide rates and the number of suicides in the US when compared to models using the basic structural time series model, selected pattern approach, and integrated pattern approach. In the modeling process, the inclusion probabilities of most search queries were close to zero, whereas search queries such as ‘domestic violence’ and ‘pain,’ which do not necessarily directly represent the status of public mental health and suicides, were included in the model with a high probability among the 51 search queries. The model achieved its aim of efficiently predicts suicides by mechanically selecting a few effective search queries from a large pool of potential search queries.

Therefore, this study concludes that the integrated and selected pattern approach is one of the potentially effective methods for predicting suicides by utilizing search queries from search engines. This approach may complements the existing selected pattern approaches (Chang et al., 2011; Gunn and Lester, 2013; Knipe et al., 2021; Kristoufek et al., 2016; McCarthy, 2010; Page et al., 2011; Sueki, 2011; Taira et al., 2021) and integrated pattern approaches (Adam-Troian and Arciszewski, 2020; Barros et al., 2019; Ekinci et al., 2023; Halford et al., 2020; Jimenez et al., 2020; Tran et al., 2017; Yang et al., 2011). Given that monitoring the current situation is essential for suicide prevention (World Health Organization, 2014), this study contributes to this effort by introducing one of the potentially effective methods.

At the same time, however, this study also suggests that the effectiveness of the selected and integrated pattern approach is not universal, as its efficacy varies depending on linguistic, cultural and demographic characteristics. In the case of Japan, which represents different linguistic and cultural contexts from the US, the accuracy and stability of the predicted values for this approach are lower than those of the selected pattern approach and the basic structural time-series model used as a naive model that does not utilize any search queries. Furthermore, even in the US, where the integrated and selected pattern approach outperforms other approaches at the country-level, it does not consistently outperform other models across all data subsets. For instance, while showing superior performance at the country-level, the approach shows varying effectiveness for different subsets such as female suicide rates; conversely, in Japan, the approach performs better for female suicide rates despite its lower performance at the country-level.

Previous studies have demonstrated the effectiveness of a selected pattern approach for suicides prediction using Japanese data (Sueki, 2011; Taira et al., 2021). Additionally, some studies have shown that suicides can be predicted using only simple time-series models with trends and seasonality, without any other variables (Rostami et al., 2019; Swain et al., 2021). These studies suggest that it may be more appropriate to employ a simpler model, tailored to the specific circumstances at hand, such as cases where the available data is limited or where the predictive power of search queries for suicides is inherently low.

In addition to these modeling problems, it also implicates that potential differences in the meaning of suicide across various cultural and community contexts (Eskin et al., 2020; Eskin, 2013; Mueller et al., 2021; Mueller and Abrutyn, 2016) may influences the relationship between search queries and suicides. For instance, Eskin et al. (2020) conducted a multilevel analysis using data from 12 countries, including the US and Japan, with a focus on the cultural dimension of individualism-collectivism. Their findings indicated that cultural factors significantly influence individual-level suicidal ideation, suicide attempts, and attitudes toward suicide. While empirical evidence regarding the impact of cultural factors on the relationship between search queries and suicides is currently limited, these studies indicate that differences in the results between US and Japan observed in this study may be attributable to the distinct cultural orientations of the US (characterized by relatively individualistic values) and Japan (characterized by relatively collectivistic values).

Moreover, several studies also have indicated that demographic factors play a significant role in the development of suicidal ideation and behaviors across diverse populations (Freeman et al., 2017; Miranda-Mendizabal et al., 2019; Richardson et al., 2023). For instance, Freeman et al. (2017) suggested that while gender differences are evident in suicidal ideation and attempts, the extent and characteristics of these discrepancies vary significantly across countries. From these perspectives, the varying efficacy of the selected and integrated pattern approach across distinct data subsets in this study indicates that demographic characteristics (gender) may influence the relationship between search queries and suicides.

Therefore, this study concludes that while the integrated and selected pattern approach is one of the potentially effective methods, its practical implementation requires careful validation through comparing its efficacy with several other models.

References

Adam-Troian, J., Arciszewski, T., 2020. ‘ Absolutist Words From Search Volume Data Predict State-Level Suicide Rates in the United States ’ , *Clinical Psychological Science*, 8, 788–793. doi: https://doi.org/10.1177/2167702620916925

Arensman, E., Scott, V., De Leo, D., Pirkis, J., 2020. ‘ Suicide and Suicide Prevention From a Global Perspective ’ , *Crisis*, 41, S3–S7. doi: https://doi.org/10.1027/0227-5910/a000664

Barros, J.M., Melia, R., Francis, K., Bogue, J., O ’ Sullivan, M., Young, K., Bernert, R.A., Rebholz-Schuhmann, D., Duggan, J., 2019. ‘ The Validity of Google Trends Search Volumes for Behavioral Forecasting of National Suicide Rates in Ireland ’ , *International Journal of Environmental Research and Public Health*, 16, 3201. doi: https://doi.org/10.3390/ijerph16173201

Burnett, D., Eapen, V., Lin, P.-I., 2020. ‘Time Trends of the Public’s Attention Toward

Suicide During the COVID-19 Pandemic: Retrospective, Longitudinal Time-Series Study ’ , *JMIR Public Health and Surveillance*, 6, e24694. doi: https://doi.org/10.2196/24694

Centers for Disease Control and Prevention. ‘Vital Statistics Online Data Portal’ . URL https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm#Downloadable (accessed 5.13.24).

Chang, S.-S., Page, A., Gunnell, D., 2011. ‘ Internet Searches for a Specific Suicide Method Follow Its High-Profile Media Coverage ’ , *American Journal of Psychiatry*, 168, 855–857. doi: https://doi.org/10.1176/appi.ajp.2011.11020284

Chatterjee, S., Hadi, A.S., 2006. ‘Regression Analysis by Example: Chatterjee/Regression’, *Wiley Series in Probability and Statistics.* doi: https://doi.org/10.1002/0470055464

Ekinci, O., Gencay, F.B., Koyuncu, A.N., Soy, F.N., Cetin, O., Yuce, F., 2023. ‘ Assessment of the relationship between Google Trends search data and national suicide rates in Türkiye ’ , *Annals of Medical Research*, 30, 767–773. doi: https://doi.org/10.5455/annalsmedres.2023.04.083

Eskin, M., Tran, U.S., Carta, M.G., Poyrazli, S., Flood, C., Mechri, A., Shaheen, A., Janghorbani, M., Khader, Y., Yoshimasu, K., Sun, J.-M., Kujan, O., Abuidhail, J., Aidoudi, K., Bakhshi, S., Harlak, H., Moro, M.F., Phillips, L., Hamdan, M., Abuderman, A., Tsuno, K., Voracek, M., 2020. ‘ Is Individualism Suicidogenic? Findings From a Multinational Study of Young Adults From 12 Countries’, *Frontiers in Psychiatry*, 11. doi: https://doi.org/10.3389/fpsyt.2020.00259

Eskin, M., 2013. ‘ The effects of individualistic - collectivistic value orientations on non - fatal suicidal behavior and attitudes in Turkish adolescents and young adults ’ , *Scand. J. Psychol. 54*, 493-501. doi: https://doi.org/10.1111/sjop.12072

Freeman, A., Mergl, R., Kohls, E., Székely, A., Gusmao, R., Arensman, E., Koburger, N., Hegerl, U., Rummel-Kluge, C., 2017. ‘ A cross-national study on gender differences in suicide intent ’ , BMC Psychiatry 17, 234. doi: https://doi.org/10.1186/s12888-017-1398-8

Garnett, M.F., Curtin, S.C., Stone, D.M., 2022. ‘Suicide Mortality in the United States, 2000-2020 ’ , *NCHS Data Brief*, 1–8.

Google. ‘ Google Trends ’ . URL https://trends.google.com/trends/ (accessed 5.13.24).

Gunn, J.F., Lester, D., 2013. ‘ Using google searches on the internet to monitor suicidal behavior ’ , *Journal of Affective Disorders*, 148, 411–412. doi: https://doi.org/10.1016/j.jad.2012.11.004

Halford, E.A., Lake, A.M., Gould, M.S., 2020. ‘ Google searches for suicide and suicide risk factors in the early stages of the COVID-19 pandemic’, *PloS One*, 15, e0236777. doi: https://doi.org/10.1371/journal.pone.0236777

Ishwaran, H., Kogalur, U.B., Rao, J.S., 2010. ‘ spikeslab: Prediction and Variable Selection Using Spike and Slab Regression ’ , *R Journal*, 2, 68. doi: https://doi.org/10.32614/RJ-2010-018

Ishwaran, H., Rao, J.S., 2005. ‘ Spike and slab variable selection: Frequentist and Bayesian strategies ’ , *Annals of Statistics*, 33, 730–773. doi: https://doi.org/10.1214/009053604000001147

Jimenez, A., Santed-Germán, M.-A., Ramos, V., 2020. ‘ Google Searches and Suicide Rates in Spain, 2004-2013: Correlation Study ’ , *JMIR Public Health and Surveillance*, 6, e10919. doi: https://doi.org/10.2196/10919

Kandula, S., Olfson, M., Gould, M.S., Keyes, K.M., Shaman, J., 2023. ‘ Hindcasts and forecasts of suicide mortality in US: A modeling study ’ , *PLOS Computational Biology*, 19, e1010945. doi: https://doi.org/10.1371/journal.pcbi.1010945

Knipe, D., Gunnell, D., Evans, H., John, A., Fancourt, D., 2021. ‘ Is Google Trends a useful tool for tracking mental and social distress during a public health emergency? A time-series analysis’, *Journal of Affective Disorders*, 294, 737–744. doi: https://doi.org/10.1016/j.jad.2021.06.086

Kristoufek, L., Moat, H.S., Preis, T., 2016. ‘ Estimating suicide occurrence statistics using Google Trends ’ , *EPJ Data Science*, 5, 32. doi: https://doi.org/10.1140/epjds/s13688-016-0094-0

Li, J., Chen, W., 2014. ‘ Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models’, *International Journal of Forecasting*, 30, 996–1015. doi: https://doi.org/10.1016/j.ijforecast.2014.03.016

McCarthy, M.J., 2010. ‘Internet monitoring of suicide risk in the population’, *Journal of Affective Disorders*, 122, 277–279. doi: https://doi.org/10.1016/j.jad.2009.08.015

Ministry of Health, Labour and Welfare, a. ‘Annual Suicide Statistics’ . URL https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hukushi_kaigo/seikatsuhogo/jisatsu/jisatsu_year.html (accessed 5.13.24).

Ministry of Health, Labour and Welfare, b. ‘The difference between suicide statistics and Vital Statistics’. URL https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hukushi_kaigo/seikatsuhogo/jisatsu/toukeinosyurui.html (accessed 5.13.24).

Ministry of Health, Labour and Welfare, c. ‘Number and rate of suicide deaths based on Vital Statistics ’ . URL https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hukushi_kaigo/seikatsuhogo/jisatsu/jinkoudoutai-jisatsusyasu.html (accessed 6.4.24).

Ministry of Health, Labour and Welfare, d. ‘The report of Vital Statistics in 2022’. URL https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/houkoku22/index.html (accessed 6.8.24d).

Ministry of Internal Affairs and Communications. ‘ Summary of the Results of Population Estimates ’ . URL https://www.stat.go.jp/data/jinsui/2.html (accessed 5.13.24).

Miranda-Mendizabal, A., Castellví, P., Parés-Badell, O., Alayo, I., Almenara, J., Alonso, I., Blasco, M.J., Cebriá, A., Gabilondo, A., Gili, M., Lagares, C., Piqueras, J.A., Rodríguez-Jiménez, T., Rodríguez-Marín, J., Roca, M., Soto-Sanz, V., Vilagut, G., Alonso, J., 2019. ‘ Gender differences in suicidal behavior in adolescents and young adults: systematic review and meta-analysis of longitudinal studies’, Int. J. Public Health 64, 265-283. doi: https://doi.org/10.1007/s00038-018-1196-1

Mueller, A.S., Abrutyn, S., 2016. ‘ Adolescents under Pressure: A New Durkheimian Framework for Understanding Adolescent Suicide in a Cohesive Community’, *American Sociological Review*, 81, 877–899. doi: https://doi.org/10.1177/0003122416663464

Mueller, A.S., Abrutyn, S., Pescosolido, B., Diefendorf, S., 2021. ‘ The Social Roots of Suicide: Theorizing How the External Social World Matters to Suicide and Suicide Prevention ’, *Frontiers in Psychology*, 12. doi: https://doi.org/10.3389/fpsyg.2021.621569

Nguyen, P., Braun, R., 2018. ‘ Time-lagged Ordered Lasso for network inference’, *BMC Bioinformatics*, 19, 545. doi: https://doi.org/10.1186/s12859-018-2558-7

Page, A., Chang, S.-S., Gunnell, D., 2011. ‘Surveillance of Australian suicidal behaviour using the internet? ’, *Australian and New Zealand Journal of Psychiatry*, 45, 1020–1022. doi: https://doi.org/10.3109/00048674.2011.623660

Richardson, C., Robb, K.A., McManus, S., O’Connor, R.C., 2023. ‘Psychosocial factors that distinguish between men and women who have suicidal thoughts and attempt suicide: findings from a national probability sample of adults ’, Psychol. Med. 53, 3133-3141. doi: https://doi.org/10.1017/S0033291721005195

Rostami, M., Jalilian, A., Poorolajal, J., Mahaki, B., 2019. ‘ Time Series Analysis of Monthly Suicide Rates in West of Iran, 2006–2013 ’, *International Journal of Preventive Medicine*, 10, 78. doi: https://doi.org/10.4103/ijpvm.IJPVM_197_17

Scott, S.L., 2024. ‘ bsts: Bayesian Structural Time Series ’ .URL https://cran.r-project.org/web/packages/bsts/bsts.pdf (accessed 8.16.24).

Scott, S.L., Varian, H.R., 2014. ‘ Predicting the present with Bayesian structural time series ’, *International Journal of Mathematics in Operational Research*, 5, 4. doi: https://doi.org/10.1504/IJMMNO.2014.059942

Statcounter. ‘Search Engine Market Share Worldwide’. URL https://gs.statcounter.com/search-engine-market-share/all (accessed 6.9.24).

Sueki, H., 2011. ‘ Does the volume of Internet searches using suicide-related search terms influence the suicide death rate: data from 2004 to 2009 in Japan ’, *Psychiatry and Clinical Neurosciences*, 65, 392–394. doi: https://doi.org/10.1111/j.1440-1819.2011.02216.x

Swain, P.K., Tripathy, M.R., Priyadarshini, S., Acharya, S.K., 2021. ‘Forecasting suicide rates in India: An empirical exposition ’, *PLOS ONE*, 16, e0255342. doi: https://doi.org/10.1371/journal.pone.0255342

Taira, K., Hosokawa, R., Itatani, T., Fujita, S., 2021. ‘Predicting the Number of Suicides in Japan Using Internet Search Queries: Vector Autoregression Time Series Model’, *JMIR Public Health Surveillance*, 7, e34016. doi: https://doi.org/10.2196/34016

Tibshirani, R., 1996. ‘Regression Shrinkage and Selection Via the Lasso’, *Journal of the Royal Statistical Society: Series B*, 58, 267–288. doi: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tran, U.S., Andel, R., Niederkrotenthaler, T., Till, B., Ajdacic-Gross, V., Voracek, M., 2017. ‘Low validity of Google Trends for behavioral forecasting of national suicide rates’, *PloS One*, 12, e0183149. doi: https://doi.org/10.1371/journal.pone.0183149

United States Census Bureau. ‘Population Estimates APIs’ .URL https://www.census.gov/data/developers/data-sets/popest-popproj/popest.html (accessed 5.13.24).

World Health Organization, 2023. ‘Suicide’ .URL https://www.who.int/news-room/fact-sheets/detail/suicide (accessed 5.12.24).

World Health Organization, 2014. ‘Preventing suicide: a global imperative’. World Health Organization, Geneva. URL https://www.who.int/publications/i/item/9789241564779 (accessed 8.16.24).

World Health Organization, 2012. ‘Public health action for the prevention of suicide: a framework’. World Health Organization, Geneva. URL https://www.who.int/publications/i/item/9789241503570 (accessed 8.16.24).

Yang, A.C., Tsai, S.-J., Huang, N.E., Peng, C.-K., 2011. ‘Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004–2009’, *Journal of Affective Disorders*, 132, 179–184. doi: https://doi.org/10.1016/j.jad.2011.01.019

Yanti, Y., Rahardiantoro, S., 2019. ‘Stepwise Approach in Lagged Variables Time Series Modeling: A Simple Illustration’, *IOP Conference Series: Materials Science and Engineering*, 621, 012009. doi: https://doi.org/10.1088/1757-899X/621/1/012009

Zhou, Y., Qureshi, R., Sacan, A., 2012. ‘Data simulation and regulatory network reconstruction from time-series microarray data using stepwise multiple linear regression’, *Network Modeling Analysis in Health Informatics and Bioinformatics*, 1, 3–17. doi: https://doi.org/10.1007/s13721-012-0008-4