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

Sei Takane*

Waseda Institute of Political Economy, Tokyo, Japan

September 3, 2024

Abstract

Background: Suicide prevention has become a global public health and political issue. There have been active attempts to predict the status of public mental health and suicides using search queries from search engines. These studies can be categorized into two main approaches: 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, the selected pattern approach risks overlooking search queries that do not explicitly represent the phenomenon but are still valuable for generating predictions. Conversely, the integrated pattern approach faces the risk of correlations between search queries potentially leading to decreased prediction accuracy and the inefficiency and limitations of manual variable selection.

Methods: To address the limitations of the selected and integrated pattern approaches while complementing their respective strengths, 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.' To test the effectiveness of the 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. We then compared the accuracy of models using the integrated and selected pattern approach against those using the selected and integrated pattern 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. During the modeling process, it successfully selected a few effective search queries from a large pool of potential queries through automated means. However, in the case of Japan, the accuracy and stability of the predicted values for the integrated and selected pattern approach were lower than those of the selected pattern approach and the basic structural time-series model, which was used as a naive model that does not utilize any search queries.

---

*Contact: takamiwa441[at]icloud.com

Limitations: The integrated and selected pattern approach is not universally the most efficient method for predicting suicide across all linguistic and cultural contexts. It remains crucial to emphasize the importance of comparing its efficacy with several other models in practical implementations. Furthermore, this approach should be applied in more diverse cultural contexts and societies, particularly in low- and middle-income countries (LMICs), where public mental health currently presents a more acute challenge.

Conclusion: The integrated and selected pattern approach is one of the effective methods for predicting the status of public mental health and 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 the state of public mental health and suicides 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 suicide deaths 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 suicide. 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 suicide 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 the status of public mental health and 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 the current status of public mental health and suicide. 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.

The results demonstrate the effectiveness of the integrated and selected pattern approach for studies that utilize search queries from search engines to predict the status of public mental health and suicide. The model using the integrated and selected pattern approaches had 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, the selected pattern approach, and the integrated pattern

approach. However, this study also demonstrates the limitations of the integrated and selected pattern approaches. In the Japanese case, the accuracy and stability of the predicted values for this approach were lower than those of the selected pattern approach and the basic structural time-series model used as a naive model.

Future studies should extend this approach to other cultural and linguistic contexts, particularly in low- and middle-income countries (LMICs) where there is a growing need to address public mental health issues (Arensman et al., 2020; World Health Organization, 2014).

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 deaths 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 finalized 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 national 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 rate (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 rate (the number of suicides per 100,000 people).

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

---

[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). 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 are 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.

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). 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 (自殺:jisatsu), depression (鬱:utsu)
IRSV	suicide (自殺:jisatsu), depression (鬱:utsu), abuse (虐待:gyakutai), suicide methods (自殺方法:jisatsu houhou), commit suicide (自殺:jisatsu), divorce (離婚:rikon),
	anxiety (不安:fuan), unemployment (失業:shitsugyou), bipolar disorder (双極性障害:soukyokusei shougai),
	major depression (鬱病:utsubyou), loneliness (孤独:kodoku), insomnia (不眠:fumin),
	marriage (結婚:kekkon), relationship breakup (失恋:shitsuren), cancer (癌:gan),
	headache (頭痛:zutsuu), how to commit suicide (自殺 どうやって:jisatsu douyatte), asthma (喘息:zensoku),
	anxiety disorder (不安障害:fuan shougai), antidepressant (抗鬱剤:koututsuzai), social welfare (社会福祉:shakai fukushi),
	allergy (アレルギー:arerugi), pain (痛み:itami), schizophrenia (統合失調症:tougoushichoushou),
	stress (ストレス:sutoresu), alcohol (酒:sake), teen suicide (若者 自殺:wakamono jisatsu),
	suicide help (自殺 助け:jisatsu tasuke), job (仕事:shigoto), suicide ideation (自殺願望:jisatsu ganbou),
	suicide hotline (自殺 ホットライン:jisatsu hottorain), drunkenness (飲み過ぎ:nomisungi), hanging (首吊り:kubitsuri),
	severe depression (鬱病:utsubyou), illicit drugs (違法薬物:ihou yakubutsu), suicide attempt (自殺未遂:jisatsu misui),
	suicidal thoughts (自殺 考え:jisatsu kangae), social benefits (社会福祉:shakai fukushi), how to kill yourself (自殺方法:jisatsu houhou), hydrogen sulfide (硫化水素:ryuuka suiso), stock market (株式市場:kabushiki shijou), domestic violence (DV:diibui),
	hypnotics (睡眠薬:suiminyaku), charcoal burning (練炭:rentan), chronic illness (慢性疾患:mansei shikkan),
	kill yourself (自殺:jisatsu), suicide prevention (自殺予防:jisatsu yobou), lawsuit (訴訟:soshou),
	alcohol abstinence (禁酒:kinshu), religious belief (信仰:shinkou), manic depression (躁鬱病:souutsubyou)

2.2 Model Comparison

This study compares model (1) to (4).

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

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 rate, 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.

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

Model (2) 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.

$$
\begin{aligned}
y_t &= \alpha_t + \tau_t + \beta_{it}\text{IRSV}_{it} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \\
\alpha_t &= \alpha_{t-1} + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2) \\
\tau_t &= -\sum_{s=t-11}^{t-1} \delta_s + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2) \\
\beta_{it} &= \beta_{it-1} + \omega_{it}, \quad \omega_{it} \sim N\left(0, \sigma_{\omega i}^2/\sigma_{\text{SRSV}i}^2\right) \\
&1/\sigma_{\omega i}^2 \sim Ga(a, b) \\
&\sqrt{b/a} \sim Ga(1, 0.01\sigma_y) \\
&a \sim Ga(1, 10)
\end{aligned}
\tag{3}
$$

Model (3) describes the integrated pattern approach: This model sets the structure of the trend and seasonality to be the same as in model (2). However, unlike model (2), 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), where the coefficients change over time according to a random walk. In other words, this model increases the dimensions of model (2) by incorporating a larger number of search queries.

$$y_t = \alpha_t + \tau_t + \beta\text{IRSV}_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$
$$\alpha_t = \alpha_{t-1} + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2)$$
$$\tau_t = -\sum_{s=t-11}^{t-1} \delta_s + \eta_t, \quad \eta_t \sim N(0, \sigma_\eta^2)$$
$$p(\beta, \gamma, \sigma_\varepsilon^2) = p(\beta_\gamma|\gamma, \sigma_\varepsilon^2)p(\sigma_\varepsilon^2|\gamma)p(\gamma) \qquad (4)$$
$$\gamma \sim \prod_{i=1}^{K} \pi_i^{\gamma_i}(1 - \pi_i)^{1-\gamma_i}$$
$$\beta_\gamma|\sigma_\varepsilon^2, \gamma \sim N(0, \sigma_\varepsilon^2(\Omega_\gamma^{-1})^{-1})$$
$$1/\sigma_\varepsilon^2 \sim Ga(\nu/2, ss/2)$$

Finally, model (4) 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 (3).

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 (4) 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 $\Omega^{-1} = \kappa\left(\omega X^T X + (1-\omega)\text{diag}(X^T X)\right)/n$ 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). Referring to Scott (2024) and Scott and Varian (2014), this study sets $\kappa = 1$, $\omega = 1/2$, $\nu = 0.01$, $R^2 = 0.5$ as prior set values.

In summary, model (4) 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 patterning approach offers three 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

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 Rate)	MAE (Suicide Number)
US		
Model (1)	5.86	238.09
Model (2)	6.59	267.04
Model (3)	3.43	139.61
Model (4)	3.25	133.30
Japan		
Model (1)	5.15	86.94
Model (2)	4.97	83.80
Model (3)	8.84	147.71
Model (4)	5.63	94.74

more efficient and practical than manually including or excluding search queries in the model and selecting only those with high predictive power for outcomes, as unnecessary variables are automatically eliminated by the 'spike' component. In other words, this model bridges the selected and integrated pattern approaches used in previous studies by combining their advantages while addressing their respective limitations.

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 [3].

3   Results

Table 2 presents the accuracy of models (1)-(4) 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 (4), 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 mental health status and suicide, 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.

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)-(4), whereas Figure 1(b) plots the cumulative absolute error (AE) for the test

---

[3]All of the code and the results of the grid search can be referred to on the author's GitHub [URL here].
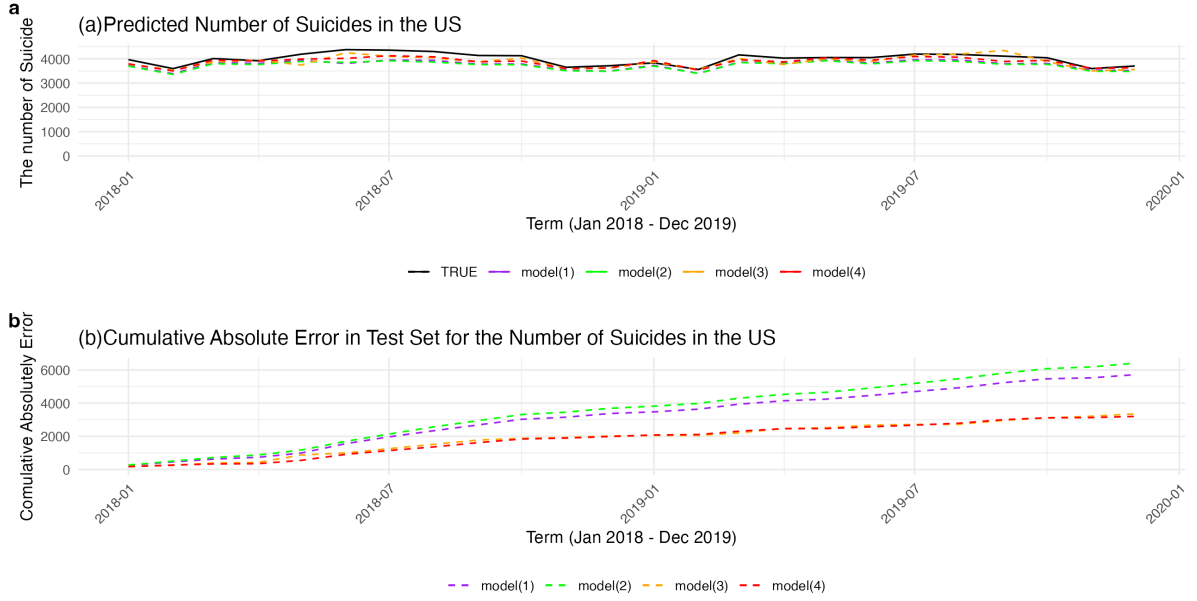
Figure 1: Accuracy of Models for US Test Set (Jan 2018 - Dec 2019): (a) Trend of Actual Suicide Numbers and Predicted Values for Model (1) - Model (4), (b) Cumulative Absolute Error (AE) for Model (1) - Model (4)

set. The accuracy of the predicted values from model (4) is remarkably more stable than that of model (1), which uses basic time-series analysis, and model (2), which employs the selected pattern approach. It is also slightly more stable than model (3), 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 (4), 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 (4). 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 suicide.

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 successfully achieved.
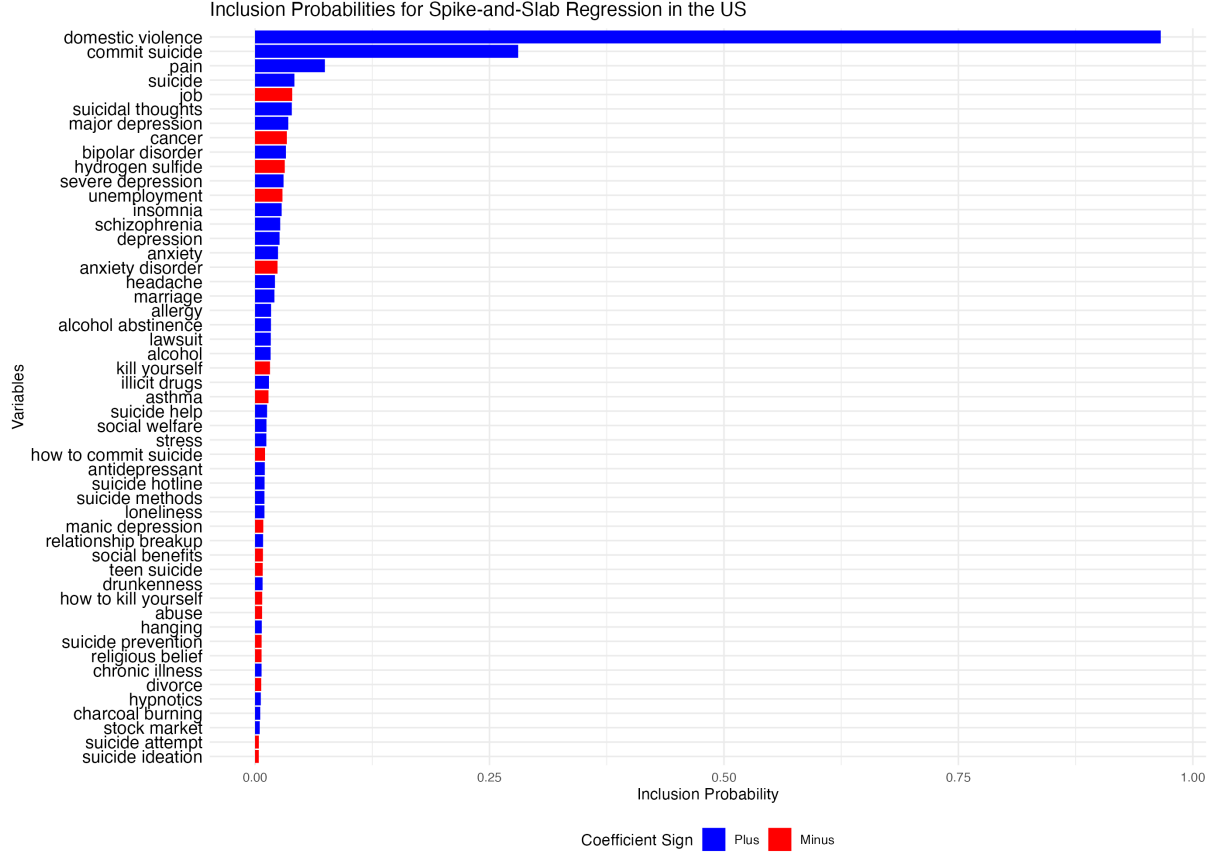
Figure 2: Inclusion Probabilities of Search Queries in Model (4) Using Spike-and-Slab Regression for US

In contrast, unlike the prediction results for the US case, model (4) does not emerge as the optimal model among the four in the Japanese case. The accuracy of the predicted values for model (4) does not outperform that of model (2), 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 (4) did not outperform model (2) in terms of the accuracy stability for the predicted values. The accuracy of model (4) did not deteriorate at any particular point in the test set; rather, it exhibited consistently lower performance across the entire test set.

Finally, Figure 4 shows the inclusion probabilities of search queries in model (4). This shows that while the inclusion probabilities of most search queries are approximately zero, search queries such as 'divorce' or 'stock market,' which do not necessarily directly represent the status of public mental health and suicides, exhibit high inclusion probabilities. However, as presented in Table 2, the accuracy of the model does not outperform that of model (2), which employs the selected pattern approach and utilizes only two search queries that directly represent the status of public mental health and suicide. This indicates that employing sparse variable selection from a large number of search queries does not guarantee efficient prediction of the status of public mental health and suicide, even when it successfully identifies specific variables from a large pool and simplifies the model through mechanical selection processes.
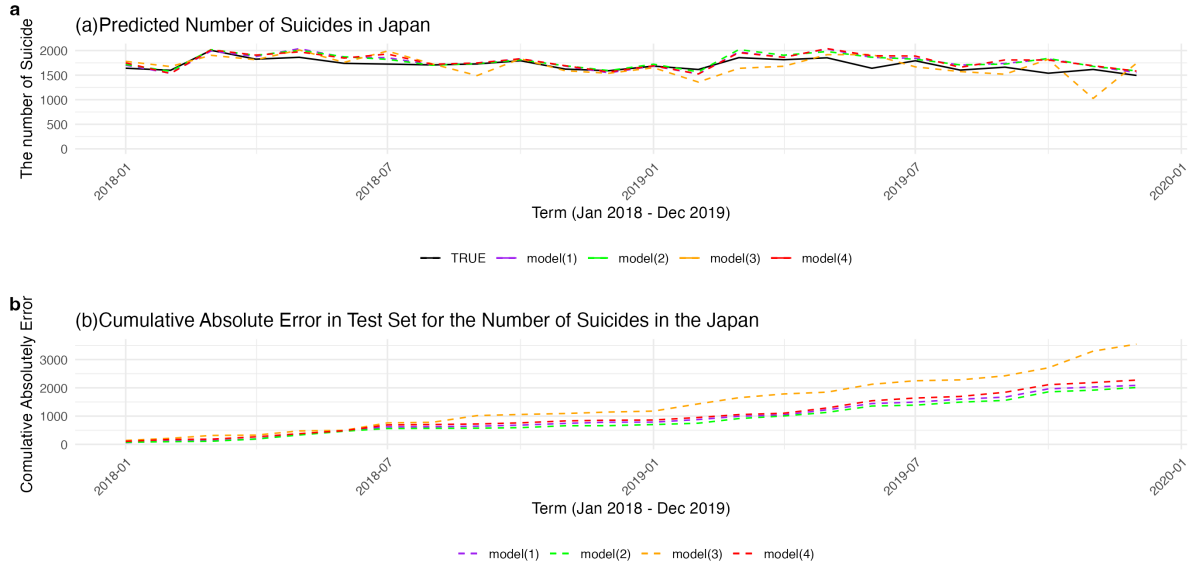
Figure 3: Accuracy of Models for Japanese Test Set (Jan 2018 - Dec 2019): (a) Trend of Actual Suicide Numbers and Predicted Values for Model (1) - Model (4), (b) Cumulative Absolute Error (AE) for Model (1) - Model (4)
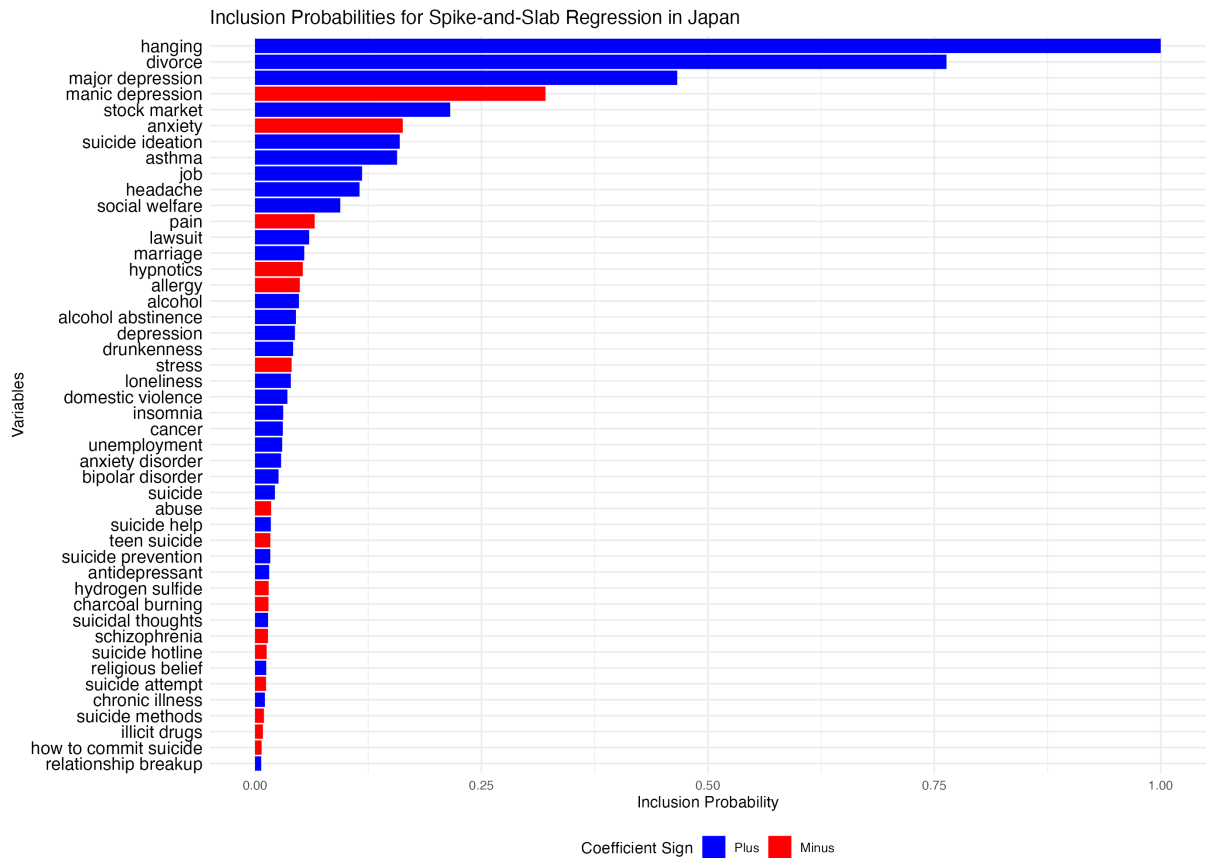

Figure 4: Inclusion Probabilities of Search Queries in Model (4) Using Spike-and-Slab Regression for Japan

4 Conclusions

This study attempted to predict the status of public mental health and suicides using search queries from search engines. This study attempted to predict country-level 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 suicide 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: the decreased model accuracy typically associated with high-dimensional models, and the inefficiency and limitations of manual variable selection. 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 the status of public mental health and suicide 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 an efficient method for predicting the status of public mental health and suicide by utilizing search queries from search engines. This approach 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 a novel and efficient method.

However, at the same time, this study also demonstrated the limitations of the integrated and selected pattern approach. 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.

Previous studies have demonstrated the effectiveness of a selected pattern approach for suicide prediction using Japanese data (Sueki, 2011; Taira et al., 2021). Additionally, some studies have shown that suicide 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). 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; Mueller et al., 2021; Mueller and Abrutyn, 2016) may influences the relationship between search queries and suicides. Therefore, this study, on the one hand, proposes the efficiency of the integrated and selected pattern approach, while on the other hand, emphasizes the importance of comparing its efficacy with multiple other models in practical implementations.

Future studies should focus on applying the integrated and selected pattern approach in more diverse cultural contexts and societies, particularly in Low and Middle Income Countries (LMICs) where public mental health is currently a more serious issue (Arensman et al., 2020; World Health Organization, 2014).

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`

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. ʻIs Individualism Suicidogenic? Findings From a Multinational Study of Young Adults From 12 Countriesʼ, *Frontiers in Psychiatry*, 11. doi: `https://doi.org/10.3389/fpsyt.2020.00259`

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

Google. ʻGoogle Trends ʼ. URL `https://trends.google.com/trends/` (accessed 5.13.24).

Gunn, J.F., Lester, D., 2013. ʻUsing google searches on the internet to monitor suicidal behavior ʼ, *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. ʻGoogle searches for suicide and suicide risk factors in the early stages of the COVID-19 pandemicʼ, *PloS One*, 15, e0236777. doi: `https://doi.org/10.1371/journal.pone.0236777`

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

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

Jimenez, A., Santed-Germán, M.-A., Ramos, V., 2020. ʻGoogle Searches and Suicide Rates in Spain, 2004-2013: Correlation Study ʼ, *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. ʻHindcasts and forecasts of suicide mortality in US: A modeling study ʼ, *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. ʻIs Google Trends a useful tool for tracking mental and social distress during a public health emergency? A time-series analysisʼ, *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. ʻEstimating suicide occurrence statistics using Google Trends ʼ, *EPJ Data Science*, 5, 32. doi: `https://doi.org/10.1140/epjds/s13688-016-0094-0`

McCarthy, M.J., 2010. ʻInternet monitoring of suicide risk in the populationʼ, *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. ʻAnnual Suicide Statisticsʼ. 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).

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`

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`

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`

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`