Advanced Daily Prediction Model for National Suicide Numbers with Social Media Data

Kyung Sang Lee[1]*, Hyewon Lee[2]*, Woojae Myung[3], Gil-Young Song[4],
Kihwang Lee[4], Ho Kim[2,5] ✉, Bernard J. Carroll[6], and Doh Kwan Kim[1] ✉

[1]Department of Psychiatry, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[2]Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, Seoul, Republic of Korea
[3]Department of Neuropsychiatry, Seoul National University Bundang Hospital, Seongnam, Republic of Korea
[4]The Mining Company, Daumsoft, Seoul, Republic of Korea
[5]Institute of Health and Environment, Seoul National University, Seoul, Republic of Korea
[6]Department of Psychiatry, Emeritus, Duke University Medical Center, Durham, NC, USA

Objective Suicide is a significant public health concern worldwide. Social media data have a potential role in identifying high suicide risk individuals and also in predicting suicide rate at the population level. In this study, we report an advanced daily suicide prediction model using social media data combined with economic/meteorological variables along with observed suicide data lagged by 1 week.

Methods The social media data were drawn from weblog posts. We examined a total of 10,035 social media keywords for suicide prediction. We made predictions of national suicide numbers 7 days in advance daily for 2 years, based on a daily moving 5-year prediction modeling period.

Results Our model predicted the likely range of daily national suicide numbers with 82.9% accuracy. Among the social media variables, words denoting economic issues and mood status showed high predictive strength. Observed number of suicides one week previously, recent celebrity suicide, and day of week followed by stock index, consumer price index, and sunlight duration 7 days before the target date were notable predictors along with the social media variables.

Conclusion These results strengthen the case for social media data to supplement classical social/economic/climatic data in forecasting national suicide events. **Psychiatry Investig 2018;15(4):344-354**

Key Words SNS, Sentiment analysis, Social, Warning signs of suicide.

INTRODUCTION

Suicide is a significant public health concern worldwide and is an important cause of death in many member countries of the Organization for Economic Co-operation and Development (OECD). Suicide accounted for over 150,000 deaths in those countries in 2013.[1] South Korea currently has the highest

OECD annual suicide rate, with nearly 30 deaths per 100,000 citizens. In addition, the suicide rate has risen steadily in South Korea over the past two decades, peaking around 2010.[2] World Health Organization (WHO) member states have committed to addressing this global problem of suicide through comprehensive, integrated and responsive mental health and social services in community-based settings.[3]

In recent decades, several clinical instruments have been developed for screening individuals at high risk of suicide.[4-6] As the social science technology advanced, interest has turned to the use of social media data for identification of suicide risk.[7,8] For instance, machine learning and natural language processing have been shown to replicate human classifications of suicide-related posts on social media.[9,10] These studies point to the potential of social media data to identify individuals at high risk of suicide. Additionally, social media data also have a potential role in suicide prediction at the population level. Jashinsky et al.[11] demonstrated a significant correlation be-

tween potential suicide-related tweets in Twitter and regional suicide rates in the United States. In our previous research, we developed a prediction model for the Korean national suicide number, using economic, meteorological, and social media data variables and it showed encouraging prediction accuracy.[12] However, our previous model needs to be refined because of its broad, 3-day prediction epochs and its limited (2-year) reference data base for the training set.

The purpose of this study is to improve the utility and accuracy of our original national suicide prediction model by modifying the design of model. We applied a serial prediction procedure which could reflect changing trends near the prediction date rather than the earlier fixed training set model.[12] We also expanded the period on which data entered in the prediction modeling are based from 2 years to 5 years, and we selected a broader range of terms from the social media content.

METHODS

General design of new prediction models

We used a serial prediction procedure to capture secular trends in the data for the 7-year time period between 1 January 2008 to 31 December 2014. Data for predictor variables were analyzed using a moving period of 5 years minus 1 week (1,818 days) that ended 7 days before the target date (t) (Figure 1). Our prediction model was applied over the final 2 years of the 7-year observation period. Thus, the first prediction model used 5-years of data from 1 January 2008 to 25 December 2012 (for simplicity we will designate this period of 1,818 days as 5 years) to predict the national suicide number for 1 Janu-

ary 2013 with a lag of 7 days. This prediction model advanced by 1 day, every day for 730 days to the end of the study period. The 730th prediction model, the last prediction model, used data from 31 December 2009 to 24 December 2014 to predict the national suicide number for 31 December 2014 with a lag of 7 days. Thus, each prediction was based on data for a unique 5-year period. In that unique period, the best prediction model was selected using a stepwise Akaike Information Criterion (AIC) method where the variables included in the model having the smallest AIC were selected as the good predictors.[13]

Suicide data

The daily national suicide numbers in South Korea from January 1 2008 to December 31 2014 were obtained from the Korea National Statistical Office (KNSO, http://kostat.go.kr/portal/english). The data were thoroughly examined and verified by KNSO. The specific 7-year time period used in our study was chosen for the contemporaneous availability of demographic and social media data. Completed suicides were classified according to the International Classification of Diseases-10 (ICD-10) codes X60-X84, and we included suicides from all causes, including intentional self-poisoning and self-harm.[14] The observed number of suicides on day (t-7) was designated the suicide variable in the prediction models. Thus, the suicide variable in the prediction model did not consider time trends for daily suicide over the 1,818 days of the prediction period.

Social media data

We obtained social media data from Daumsoft, one of the
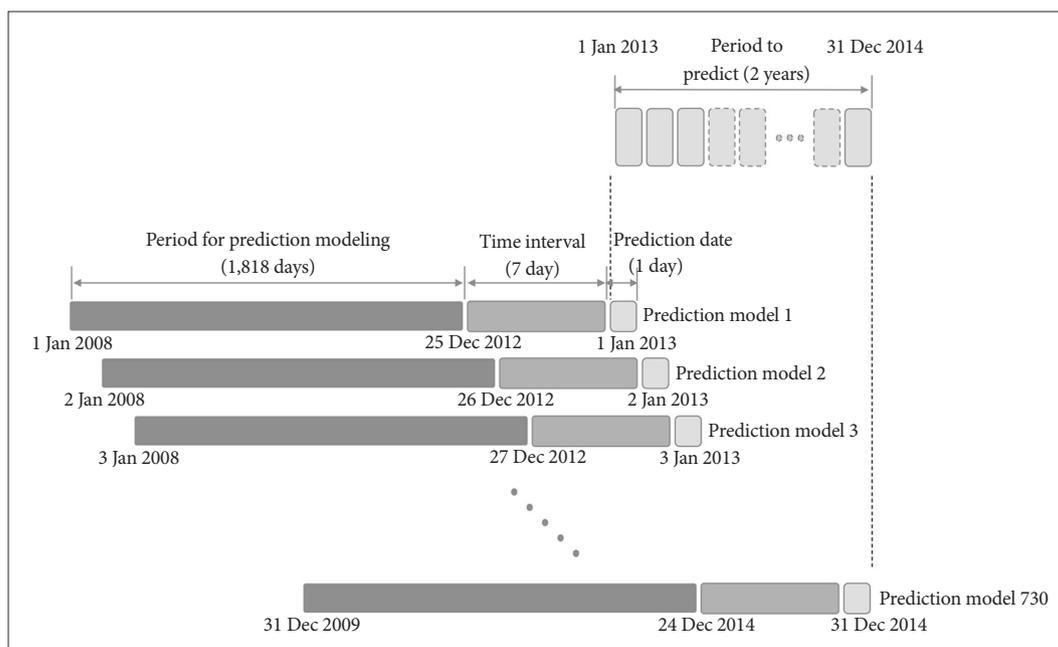
Figure 1. Serial prediction procedure. In this study, a total of 730 individual predictions were executed over 2 years.

leading social media analysis and consulting firms in South Korea.[12] The social media data were drawn from weblog posts on the Naver platform (http://section.blog.naver.com). Naver is the largest portal site for weblog services in South Korea. A set of filtering operations was applied to exclude advertisements and other noisy texts. We analyzed the weblog traffic in the period between January 1, 2008 and December 24, 2014. The weblog service processed 1,106,890,866 posts during that 7-year period. To effectively simplify and quantify the enormous amount of social media data, we recorded the measures according to their frequency of occurrence in the weblog traffic. SOCIALmetrics™, which is a social media analysis system offered by Daumsoft (http://www.daumsoft.com),[12] provides deep level keyword analysis and opinion mining for social media texts and other web documents, and the social media data were obtained using this system.

We began the social media data mining by selecting 10,035 candidate keywords. These keywords comprised 9,709 words from a sentiment word data base and 326 words that are frequently used in suicide websites. The sentiment word data base was initially constructed by repurposing the sentiment lexicon compiled by Daumsoft for its commercial text analytics service. The sentiment word data base was supplemented by sentiment words collected from previous studies on Korean sentiment words.[15,16] We also applied text mining technologies to gather complex words closely related to the seed keywords from Naver blog posts and Tweets, adapting the method used in a previous study on extending the sentiment word list.[17] The weblog count for each Korean candidate keyword was defined as the daily document frequency mentioning that word at least once. For each prediction exercise, 20 words out of the 10,035 candidate keywords were selected. The weblog counts for these 20 words had the highest statistically significant correlation coefficients over 1,818 days with the daily suicide rate on day (t) during each individual prediction period. In addition, we classified the 10,035 candidate keywords as

Table 1. An example of the prediction model for national suicide number on 1 January 2013 by using data from 1 January 2008 to 25 December 2012

Variable	Log (estimates)	Log (standard error)	t	p
Suicide variable				
Suicide (t-7)	4.96×10^{-3}	6.05×10^{-4}	8.20	4.61×10^{-16}
Economic and meteorological variables				
Stock	-1.15×10^{-4}	2.54×10^{-5}	-4.54	$5.95 \times 10{-6}$
Consumer price index	0.014	2.05×10^{-3}	6.73	2.25×10^{-11}
Sunlight	-3.32×10^{-3}	1.13×10^{-3}	-2.94	3.31×10^{-3}
Temperature	-8.33×10^{-4}	4.43×10^{-4}	-1.88	0.060
Celebrity	0.13	0.016	8.07	1.28×10^{-15}
Day of week (reference=monday)				
Tuesday	0.030	0.021	1.38	0.17
Wednesday	-6.69×10^{-3}	0.022	-0.31	0.76
Thursday	-0.036	0.022	-1.66	0.098
Friday	-0.034	0.021	-1.62	0.11
Saturday	-0.041	0.021	-2.00	0.046
Sunday	-0.027	0.017	-1.61	0.11
Social media variables				
Positive words	-1.12×10^{-6}	2.59×10^{-7}	-4.34	1.51×10^{-5}
Weblog count 2: chokjinhada	-2.51×10^{-4}	1.48×10^{-4}	-1.69	0.091
Weblog count 3: gyeongjejeok	1.58×10^{-4}	6.15×10^{-5}	2.57	0.010
Weblog count 5: mogongmakda	1.35×10^{-3}	5.21×10^{-4}	2.60	9.45×10^{-3}
Weblog count 11: jeotanso	1.23×10^{-3}	3.14×10^{-4}	3.91	9.75×10^{-5}
Weblog count 14: uuljeung	1.31×10^{-4}	9.32×10^{-5}	1.41	0.16
Weblog count 15: deodida	4.28×10^{-4}	1.98×10^{-4}	2.16	0.031
Weblog count 20: meoriapeuda	4.82×10^{-4}	1.25×10^{-4}	3.85	1.21×10^{-4}
Weblog count 25: uijonhada	2.45×10^{-4}	1.56×10^{-4}	1.57	0.12
Weblog count 29: buran	1.53×10^{-4}	8.18×10^{-5}	1.87	0.062

Estimates are change of natural logarithm of observed suicide number at prediction date (t) per one increase of predicted variables at t-7

negative, neutral and positive according to the meanings of the words, and we included these 3 classifications independently in the prediction model as social media variables. Thus, a total of 23 social media variables were considered for constructing multivariate regression models along with the other predictive variables at each prediction. An example of a multivariate regression model is shown in Table 1 and 2 lists the social media variables that were used in any of the 730 predictions. These comprised just 30 weblog counts out of the 1,035 candidate keywords, and 3 non-mutually exclusive weblog meaning categories.

Economic, meteorological, and air pollution data

Our study includes economic and meteorological variables which were identified in previous studies of suicide as important variables. The economic data,[18,19] which were extracted from the KNSO, consist of 3 variables: consumer price index; unemployment rate; and stock index valuations (Korea Composite Stock Price Index, KOSPI). The meteorological data, which were obtained from the Korea Meteorological Administration (KMA, http://web.kma.go.kr/eng), consist of two variables: sunlight hours and mean daily temperature.[14,20] Two major air pollution variables [ozone and particulate matter with size of 10 μm in diameter or smaller (PM-10)] were considered based on our previous study for the association between air pollution and suicide.[21] During the study period, the observation station in Seoul was chosen for representative measurement data (http://www.airkorea.or.kr/eng/index). The 730 prediction models each used data from all 1,818 days in each unique prediction period for these variables.

Celebrity suicides

To control the influence of celebrity suicides, we considered their occurrence as a confounding variable. We regarded celebrity suicides as any suicide that received more than two weeks of exposure in news programs of the three major national television networks (KBS, MBC, and SBS).[13] celebrity suicides met this definition during the 7 years of this study. The affected period was defined as a month (30 days) after the first report of the celebrity suicide, according to the study of Phillips.[22] Prediction dates were coded 1 when they were within this 30-day window, while all others were coded 0 on the celebrity variable. The 730 prediction models each used celebrity suicide data from all 1,818 days in each unique prediction period.

Day of week

The day of week was included as a confounding variable to control the variation in the number of suicides by date.[23,24] During the 7-year observation period, the number of suicides varied according to the day of the week. The largest number of suicides occurred on Monday (Average, 43.38; SD, 8.96), followed by Wednesday (Average, 40.87; SD, 9.00), Tuesday (Average, 40.27; SD, 7.39), Thursday (Average, 38.85; SD, 7.45), Friday (Average, 38.57; SD, 7.94), Sunday (Average, 35.65; SD, 7.48), and Saturday (Average, 34.22; SD, 7.84). The 730 prediction models each used data from all 1,818 days in each unique prediction period for this variable.

Ethics statement

Our research analyzes existing data and documents that are publicly available in a manner that does not allow individual subjects to be identified, therefore ethics approval was deemed unnecessary.

Statistical analysis

To avoid multi-collinearity problems and redundancy among the predictor variables, we selected 20 words through the Spearman's correlation test between the daily weblog count for each of the 10,035 candidate keywords on day (t-7) and the daily suicide number 7 days later (t) across the 1,818 days of each unique modeling period. We selected the 20 words that showed the highest correlation coefficient values with suicide number on the prediction date (t), with the added requirement that each word showed p<0.05 in Spearman's correlation test (without correction for multiple comparisons). The multivariate regression model was constructed using weblog counts of these selected words over 1,818 days, along with other predictors [observed number of suicides on day (t-7), economic and meteorological variables, and the sums of the counts of weblog posts that contain positive/neutral/negative words] over 1,818 days. The variables of the prediction models were selected stepwise based on AIC values. Following the initial period of 1,818 days in which we developed the first prediction model, we prospectively predicted national suicide number daily for 2 years from 1 January 2013 to 31 December 2014. For this operation, we used the 'predict' function with 'prediction interval level' set at 0.85 in the R software, which indicates that the observed number of national suicides is expected to fall within the upper and lower boundaries of the prediction interval with 85% probability. If the observed number fell within the prediction interval, we regarded the prediction as correct.[12] Prediction accuracy was defined as the ratio of correct predictions to total predictions. In all statistical computations, dependent variables (national suicide number on prediction date) were seasonally adjusted with a decomposition method using the 'decomposition' function in the Technical Trading Rules (TTR) of the R package. By this procedure, national suicide number was decomposed into seasonal and non-seasonal components, and only the non-seasonal component was included in the statistical model.

Table 2. Prediction variables evaluated in 730 prediction models

Variable	Description	Rates of inclusion	Average estimates	Ranges of estimates (lower, upper)
Suicide variable				
Suicide (t-7)	Observed number of suicides	730/730 (100%)	4.11×10^{-3}	2.55×10^{-3}, 6.05×10^{-3}
Economic and meteorological variables				
Stock	Korean stock index, KOSPI	698/730 (95.6%)	4.34×10^{-5}	-1.43×10^{-4}, 1.37×10^{-4}
Consumer price index	Monthly consumer price index	639/730 (87.5%)	-3.79×10^{-3}	-1.48×10^{-2}, 1.72×10^{-2}
Unemployment	Monthly unemployment rate	332/730 (45.5%)	-0.021	-3.33×10^{-2}, -1.35×10^{-2}
Sunlight	Sunlight duration	616/730 (84.4%)	-2.17×10^{-3}	-3.34×10^{-3}, -1.41×10^{-3}
Temperature	Daily average temperature	401/730 (54.9%)	-1.23×10^{-3}	-1.81×10^{-3}, -5.47×10^{-4}
Ozone	Daily average ozone level	18/730 (2.5%)	-0.89	-1.09, -0.71
PM-10	Daily average particulate matter	0/730 (0%)	-	-
Celebrity	If 7 days before prediction date is within one month after a celebrity suicidal event, 1; else, 0	730/730 (100%)	0.11	8.86×10^{-2}, 0.14
Day of week	Six dummy variables for day of week	730/730 (100%)	-	-
Social media variables*				
Meaning classifications				
Positive words	Sum of count of weblog posts that contain positive words at least once	175/730 (24%)	-8.92×10^{-7}	-1.40×10^{-6}, 3.29×10^{-7}
Neutral words	Sum of count of weblog posts that contain neutral words at least once	502/730 (68.8%)	6.24×10^{-7}	2.14×10^{-7}, 1.54×10^{-6}
Negative words	Sum of count of weblog posts that contain negative words at least once	313/730 (42.9%)	-2.68×10^{-7}	-4.60×10^{-7}, -1.12×10^{-7}
Top 30 weblog counts†				
Weblog count 1: siljejeok	Meaning 'practical'; 'matter-of-fact'; 'businesslike'	503/730 (68.9%)	4.65×10^{-4}	2.94×10^{-4}, 7.32×10^{-4}
Weblog count 2: chokjinhada	Meaning 'facilitate'; 'promote'; 'accelerate'	476/730 (65.2%)	-2.67×10^{-4}	-4.44×10^{-4}, -1.78×10^{-4}
Weblog count 3: gyeongjejeok	Meaning 'economical'; 'financial'	393/730 (53.8%)	1.15×10^{-4}	6.26×10^{-5}, 2.30×10^{-4}
Weblog count 4: jjieogitda	Meaning 'There have been comments about'	383/730 (52.5%)	1.05×10^{-3}	6.62×10^{-4}, 1.21×10^{-3}
Weblog count 5: mogongmakda	Meaning 'blocks the skin's pores'	349/730 (47.8%)	1.11×10^{-3}	6.51×10^{-4}, 1.67×10^{-3}
Weblog count 6: piryoitda	Meaning 'it requires that'; 'need to'	342/730 (46.8%)	-1.03×10^{-4}	-1.37×10^{-4}, -7.84×10^{-5}
Weblog count 7: hyogwageoduda	Meaning 'obtain the desired results'; 'get effect'	342/730 (46.8%)	6.52×10^{-4}	5.45×10^{-4}, 8.39×10^{-4}

Table 2. Prediction variables evaluated in 730 prediction models (continued)

Variable	Description	Rates of inclusion	Average estimates	Ranges of estimates (lower, upper)
Weblog count 8: mijigeunhada	Meaning 'lukewarm'	315/730 (43.2%)	2.11×10^{-4}	1.53×10^{-4}, 3.30×10^{-4}
Weblog count 9: gakgwangbatba	Meaning 'taking spotlight'	176/730 (24.1%)	-3.64×10^{-4}	-4.65×10^{-4}, -2.50×10^{-4}
Weblog count 10: sonsil	Meaning '(economic) loss'	164/730 (22.5%)	2.02×10^{-4}	1.10×10^{-4}, 2.75×10^{-4}
Weblog count 11: jeotanso	Meaning 'low carbon'	152/730 (20.8%)	1.57×10^{-3}	7.65×10^{-4}, 2.07×10^{-3}
Weblog count 12: ganeungseongkeuda	Meaning 'be in with a shout'; 'possible'	144/730 (19.7%)	2.54×10^{-4}	1.79×10^{-4}, 3.23×10^{-4}
Weblog count 13: chimche	Meaning 'recession'; '(economic) depression'	140/730 (19.2%)	1.72×10^{-4}	1.15×10^{-4}, 2.91×10^{-4}
Weblog count 14: uuljeung	Meaning 'depressive disorder'	123/730 (16.8%)	1.49×10^{-4}	1.11×10^{-4}, 1.80×10^{-4}
Weblog count 15: deodida	Meaning 'slow'	119/730 (16.3%)	3.92×10^{-4}	3.24×10^{-4}, 5.01×10^{-4}
Weblog count 16: chansa	Meaning 'praise'; 'compliment'	116/730 (15.9%)	2.16×10^{-4}	1.74×10^{-4}, 2.72×10^{-4}
Weblog count 17: gwallyeonitda	Meaning 'have relevant to'	110/730 (15.1%)	3.20×10^{-4}	2.66×10^{-4}, 3.82×10^{-4}
Weblog count 18: yecheukhada	Meaning 'predict'; 'forecast'	105/730 (14.4%)	2.90×10^{-4}	2.55×10^{-4}, 3.26×10^{-4}
Weblog count 19: jureodeulda	Meaning 'decrease'; 'diminish'	97/730 (13.3%)	-1.04×10^{-4}	-1.30×10^{-4}, -8.51×10^{-5}
Weblog count 20: meoriapeuda	Meaning 'have a headache'	93/730 (12.7%)	4.60×10^{-4}	4.27×10^{-4}, 5.00×10^{-4}
Weblog count 21: miljeopangwangye	Meaning 'intimate relation'	93/730 (12.7%)	1.49×10^{-3}	1.21×10^{-3}, 1.71×10^{-3}
Weblog count 22: uijonjeok	Meaning 'dependent'	90/730 (12.3%)	1.13×10^{-3}	7.84×10^{-4}, 1.47×10^{-3}
Weblog count 23: jeungga	Meaning 'increase'; 'growth'	88/730 (12.1%)	7.31×10^{-5}	5.80×10^{-5}, 8.72×10^{-5}
Weblog count 24: gyujewanhwa	Meaning 'relaxation of regulation'; 'deregulation'	85/730 (11.6%)	7.28×10^{-4}	3.34×10^{-4}, 9.95×10^{-4}
Weblog count 25: uijonhada	Meaning 'depend on'; 'reliance on'	85/730 (11.6%)	2.47×10^{-4}	2.16×10^{-4}, 2.86×10^{-4}
Weblog count 26: ironjeok	Meaning 'theoretical'	79/730 (10.8%)	4.84×10^{-4}	4.08×10^{-4}, 5.22×10^{-4}
Weblog count 27: keodarata	Meaning 'big'; 'huge'; 'large'	76/730 (10.4%)	1.31×10^{-4}	1.22×10^{-4}, 1.68×10^{-4}
Weblog count 28: gogeupwha	Meaning 'gentrification'	74/730 (10.1%)	9.20×10^{-4}	7.69×10^{-4}, 1.11×10^{-3}
Weblog count 29: buran	Meaning 'anxiety'	67/730 (9.2%)	2.05×10^{-4}	1.48×10^{-4}, 2.85×10^{-4}
Weblog count 30: gyeonggihoebok	Meaning 'a business recovery'; 'return to prosperity'	56/730 (7.7%)	9.60×10^{-4}	8.72×10^{-4}, 1.05×10^{-3}

All prediction variables were derived from 7 days before prediction date (t-7) within each unique 5-year data. Estimates are change of natural logarithm of observed national suicide number at prediction date (t) per one increase of predicted variables at t-7, then averaged for 730 prediction models. *weblog posts that contain the word/words at least once, †top 30 weblog count variables that could be included in the prediction with the highest frequency across all 730 prediction models

Because suicide number is not normally distributed, we used natural logarithm transformation of the suicide number in the regression analysis. After the regression analysis, the predicted value of the non-seasonal component of the national suicide number on the prediction date was summed with the seasonal component on the prediction date to calculate the predicted undecomposed national suicide number. All statistical analyses were performed using the R 3.2.3 public statistics software (http://www.r-project.org).

RESULTS

Trend of annual national suicide numbers

Over the 7 years of our study, annual national suicide numbers trended upwards (suicides per 100,000 persons: 26 in 2008; 31 in 2009; 31.2 in 2010; 31.7 in 2011; 28.1 in 2012; 28.5 in 2013; and 27.3 in 2014), reflecting a general trend over 20 years preceding 2010 (Figure 2). The mean of absolute daily suicide number during this study was 38.83 (standard deviation=8.51) with a range from 16 to 72.

An example of prediction model for national suicide numbers

Our model predicted daily national suicide numbers for 2 years, using accumulated information of the preceding 5-year period for prediction modeling. For an example (Table 1), the predicted suicide number on 1 January 2013 (t) was derived from data for the unique 5-year period extending from 1 January 2008 to 25 December 2012. As the result of the stepwise AIC method selecting the best model based on the smallest AIC, 1 suicide variable, 5 of 8 economic and meteorological variables, 9 of 20 weblog counts, and 1 of 3 meaning classifications of social media variables were selected in that unique period. Through this process, the number of suicides on January 1, 2013 was predicted to be 28.75 (prediction interval: 19.60 to 40.75), and the actual number of suicides observed was 32. Among the selected predictors, the strongest single predictor of suicide number on the prediction date (t) was the suicide number on day (t-7) [log (estimates), 4.96×10^{-3}; P, 4.61×10^{-16}]. When expressed as percent change in relative risk, the suicide number on prediction date (t) in-
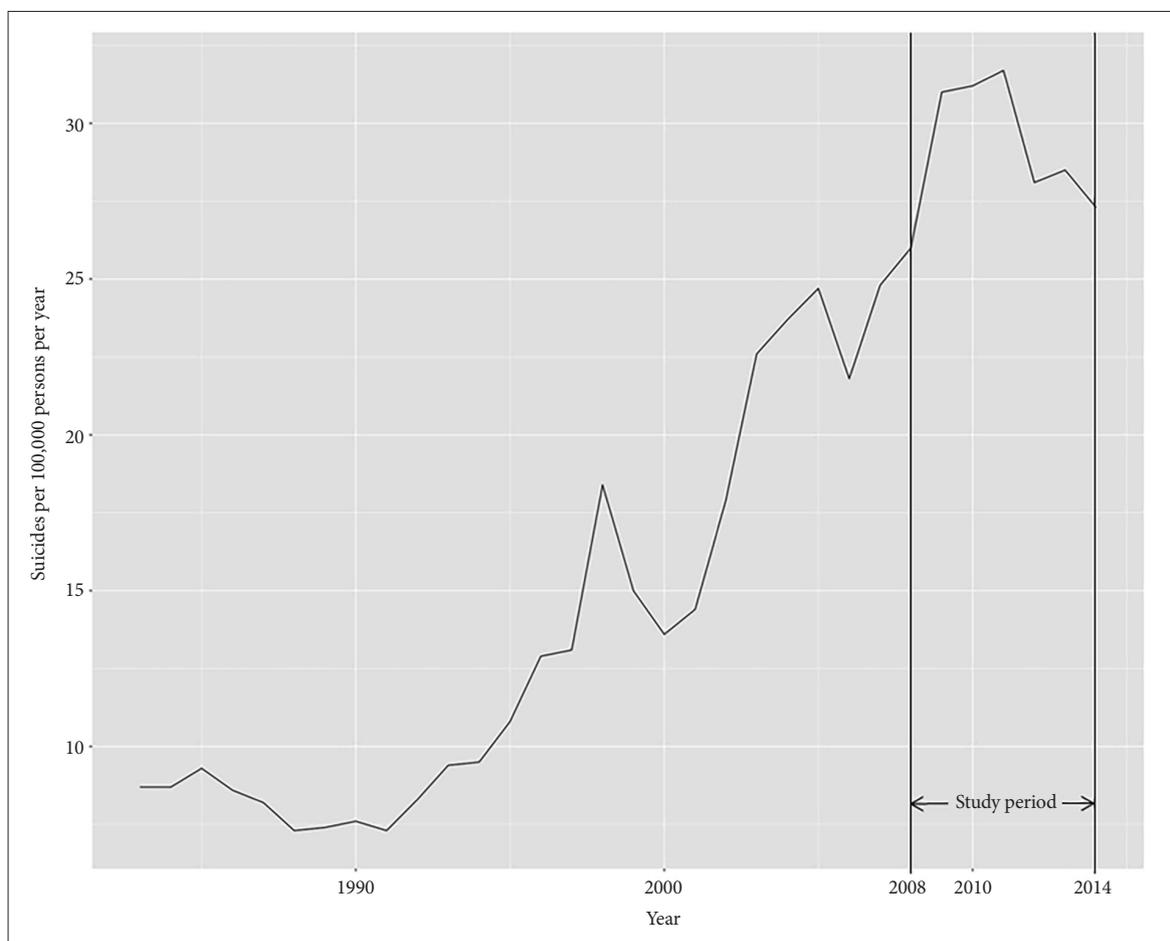
Figure 2. Trend of annual Korean national suicide numbers per 100,000 persons.

creases by 5% (95% CI: 4.9–5.2%) per 10 additional suicide commission on day (t-7). Among the economic and meteorological variables, KOSPI and sunlight hours showed negative associations, whereas the consumer price index and celebrity suicide events showed positive associations with the observed suicide number on the prediction date (t). The log estimate of Friday effect [as the prediction date (t) was Friday] was -0.034 when the reference is Monday, but not statistically significant; People committed suicide less on Friday compared to Monday. Among the social media variables, the sum of positive words showed a negative association, while weblog counts 3, 5, 11, 15, and 20 showed positive associations with the suicide number on prediction date (t). Among 9 words of weblog counts, 3 words were related to economic aspects (Weblog count 3, 11, and 15).

Prediction variables included in 730 individual prediction models

Our model produced 730 individual daily predictions over the 2 years from 1 January 2013 to 31 December 2014. Each prediction was based on a unique set of historical data. As the prediction variables could differ at each prediction date (t), the inclusion rates of the variables showed variation over the 2 years (Table 2). There were three variables [number of suicides on day (t-7), celebrity suicide, and day of week] included in all 730 individual daily predictions, followed by KOSPI, 95.6% inclusion rate; consumer price index, 87.5% inclusion rate; and sunlight duration, 84.4% inclusion rate. Among the

economic and meteorological variables, the inclusion rate for unemployment, ozone, and PM-10 was much lower (unemployment, 45.5%; ozone, 2.5%; PM-10, 0%). In Table 2 the weblog counts are listed 1−30 by rank of their inclusion rates, which ranged from 68.9% to 7.7%. Table 2 also displays the average estimates (AE) and the range of estimates for each variable across the 730 individual daily predictions. During the 2-year prediction period, we compared the predicted numbers with the observed numbers for each 730 days. As a result, of the total 730 days, 605 days of observed numbers fell within the prediction intervals with 85% probability. The prediction accuracy was 82.9%, which is defined as the percentage of correct predictions for total predictions (Figure 3). Correlation between predicted and observed suicide number was highly significant (n=730, rho=0.56, p=5.56×10^{-62}, Spearman correlation test).

The suicide number recorded on day (t-7) showed a positive association [average estimates (AE), 4.11×10^{-3}; range of estimates (RE), 2.55×10^{-3} to 6.05×10^{-3}] with the national suicide number on prediction date (t). This variable had a 100% inclusion rate. Among the economic and meteorological variables, unemployment (AE, -0.021; RE, -3.33×10^{-2} to -1.35×10^{-2}), sunlight (AE, -2.17×10^{-3}; RE, -3.34×10^{-3} to -1.41×10^{-3}), temperature (AE, -1.23×10^{-3}; RE, -1.81×10^{-3} to -5.47×10^{-4}), and ozone (AE, -0.89; RE, -1.09 to -0.71) showed negative associations with national suicide numbers on prediction date (t), while the celebrity suicide variable (AE, 0.11; RE, 8.86×10^{-2} to 0.14) showed a positive association with national suicide
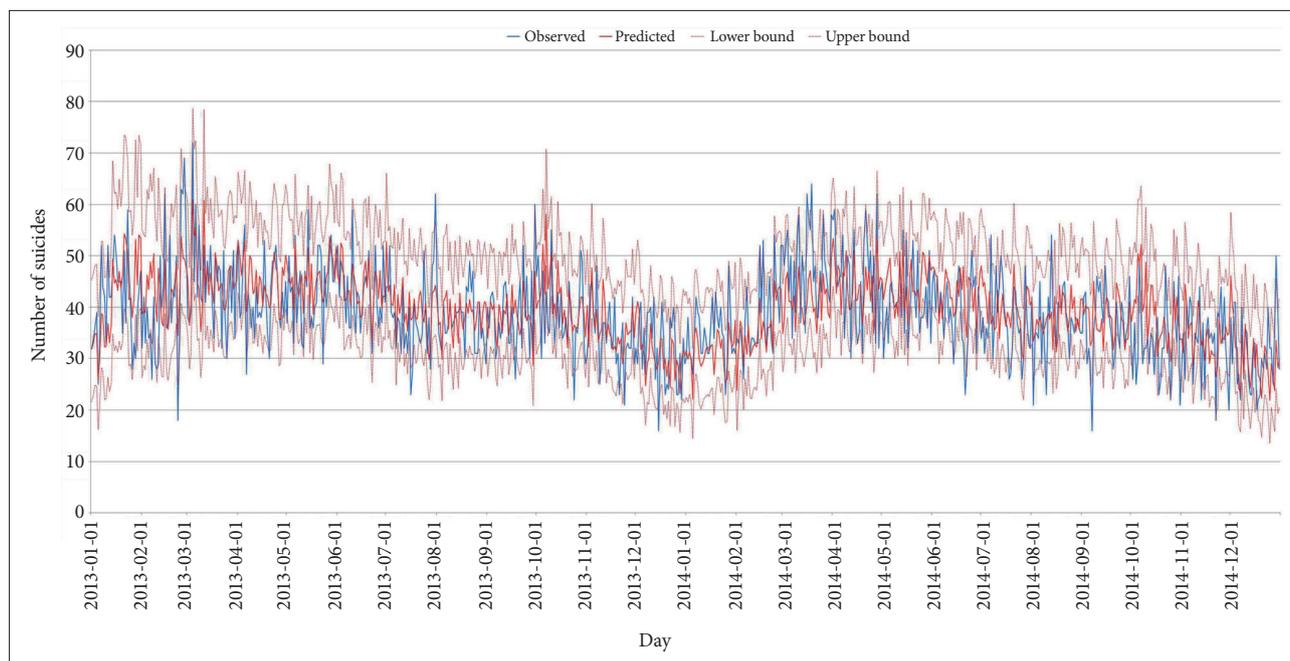
Figure 3. Prediction of daily Korean national suicide number in 2-year prediction period. Observed suicides (blue solid line), predicted suicides (red solid line), and prediction intervals (red dashed lines). The prediction range was computed for 85% probability. Prediction range accuracy was 82.9% for the 2-year prediction period.

numbers on the prediction date (t). For the economic variables, KOSPI (AE, 4.34×10^{-5}; RE, -1.43×10^{-4} to 1.37×10^{-4}) and consumer price index (AE, -3.79×10^{-3}; RE, -1.48×10^{-2} to 1.72×10^{-2}) showed an inconsistent direction of associations with national suicide numbers on the prediction date (t), evidenced by negative to positive values of RE (Table 2).

For social media variables classified by meaning, both positive and negative words showed negative associations with national suicide numbers on the prediction data (t), while neutral words showed a positive association (Table 2). Each of the top 30 weblog counts showed a consistently positive or consistently negative association with the observed national suicide numbers. For example, "Uuljeung" (weblog count 14, meaning 'depressive disorder') and "Buran" (weblog count 29, meaning 'anxiety') consistently showed positive associations with national suicide numbers, evidenced by positive RE values, when these words were included in the prediction models. Among the top 30 weblog counts, 18 words denoted an economic aspect (Weblog count 1, 2, 3, 6, 7, 9, 10, 11, 12, 13, 15, 18, 19, 23, 24, 27, 28, and 30); 2 words denoted mood status (Weblog count 14 and 29). Among the top 30 weblog counts, 4 words (Weblog count 2, 6, 9, and 19) showed negative associations with observed suicide numbers, while the other 26 words showed positive associations (Table 2).

DISCUSSION

The chief aim of this study is to improve the accuracy of national daily suicide rate predictions with an advanced multiple variable prediction model. Our new model predicted national daily suicide number with 82.9% accuracy, which is higher than that of the previous model we developed (79%).[12] Additionally, a prediction time lag of 7 days was feasible and the prediction was more precisely targeted to a single day, in contrast to the 3 day-epoch of the previous model.

Several systematic reviews for suicide prevention strategies have focused principally on gatekeeper training, screening programs, public education, media education, and restricting access to lethal means.[25-29] If suicide high-risk days or periods could be predicted, various suicide prevention strategies could be optimized to reduce suicide number effectively after such vulnerable periods were declared for the population. For an example, after a famous American rock star, Kurt Cobain's death by suicide, many experts worried that "copycat" suicide might occur in the aftermath.[30-32] The City of Seattle and several local radio stations collaborated to organize the vigil and invited a crisis clinic director to speak to the public concern and to educate thousands of mourners with the tape of Kurt Cobain's widow presenting his suicide in a negative fashion. The press and visual media reported Cobain's sui-cide and life with responsible concern, distinguishing Cobain the musician from Cobain the depressed drug abuser and suicide. These various efforts might have counteracted any potential glamorization of his death. Indeed, the record revealed no marked spike in suicides associated with the celebrity suicide.[30]

Traditional approaches to suicide prevention include the use of mailings, brochures, billboards, radio, television, and telephones.[33] Over recent years, the amount of data transmitted via internet, smart devices, and social network services has exponentially increased. The internet has been widely used for suicide prevention efforts, as a means of raising public awareness and increasing access to reliable information.[34] Mobile applications on smart phones can deliver mental health interventions effectively for a range of mental health disorders, including depression, stress, anxiety, smoking cessation, and psychosis.[35-37] As the technology in electronic and ambulatory assessment devices becomes more sophisticated, mental health care would be more accessible, efficient and interactive for patients.[38-41] We expect that our improved daily national suicide prediction model could be utilized in these technologies in the near future and would assist in promoting the positive outcome of suicide prevention efforts.

It is noteworthy that among the social media variables that showed high predictive strength, nearly two thirds (18 of 30 words) were associated with economic welfare. This result might suggest that economic themes give a good reflection of public mood in social media data. This finding is consistent with multiple previous studies about the negative influence of financial stress on the suicide rate, which is observed in several countries, European Union,[42] Russia,[43] East/Southeast Asia,[44-46] Greece,[47] South Korea,[48-51] and even in different continents simultaneously.[52]

We also observed meaningful associations between traditional economic variables and suicide rate. KOSPI and consumer price index showed relatively high inclusion rates (95.6% for KOSPI; 87.5% for consumer price index) in the 730 prediction models. This result suggests that these variables had high predictive power for suicide and it is consistent with previous studies that linked suicide rates with the consumer price index, although the direction of the association was not always consistent.[18,53,54] However, traditional economic variables are somewhat limited in respect of immediacy of predictions, as they cannot be measured daily. Stock market variables such as KOSPI are closed 2 days a week. Consumer price index and unemployment are usually assessed periodically. Even though traditional economic variables were included in our prediction models with relatively high frequency, the majority of selected social media variables were related to economic welfare. This suggests that social media variables

sampled daily might compensate for the limitations of traditional economic variables in predicting suicide numbers.

Unemployment showed a negative association with numbers of suicides, when included in our prediction models. This result is different from previous review studies which demonstrated that unemployment was positively associated with a greater incidence of suicide related behaviors[55-58] and suicide rate.[12] Sunlight and temperature showed negative associations with suicide rate, when each variable was included in our prediction models. This result does not agree with previous studies[59,60] or with our previous study.[12] Though the reason for these different results is unclear, we should note the relatively low inclusion rate of unemployment (45.5%) and temperature (54.9%) in the 730 individual prediction models. Low inclusion rates might imply that these variables had relatively low predictive strength for suicide in comparison to other variables. Therefore, the association of any single variable with suicide rate should be carefully interpreted because our prediction models included multiple variables. Negative words estimates were found to contribute to the reduction of suicide rates when included in 730 individual prediction models. However, since the negative words used in this study are classified within 10,035 candidate keywords, it is unlikely that these results represent the general relationship between the negative words used in the blog and suicide rates.

The top 30 Weblog counts of social media variables included 2 words meaning 'depressive disorder' and 'anxiety' in the prediction models. This result might imply that public mood was sensitively represented by social media data. There have been several studies that suggest social media data could be used to detect public mood sensitively.[61-64] In our previous study, two weblog words denoting 'suicide' and 'dysphoria' were significantly associated with national suicide number.[12] Jashinsky et al.[11] used suicide-related keywords and phrases on tweets which included depression and other psychological disorders for study. They found a strong correlation between regional Twitter-derived data and observed regional suicide data.

In conclusion, our prediction models could predict national suicide number for a single date 7 days in advance with 82.9% accuracy, which is an improvement on our previous study.[12] Overall the correlation between predicted and observed daily suicide rates was highly significant, accounting for 31% of the variance (r=0.56) over a 2-year span of daily predictions. Considering that the data used for these predictions cannot account for subject-specific variables, the correlation between predicted and observed daily suicide numbers is encouraging. When used in combination with traditional economic variables, the social media data, especially words related to economic welfare or mood status, showed high predictive strength. Further studies for utilization of this model to advanced technology are needed.

REFERENCES

1. OECD. Health at a Glance 2015. Paris: OECD Publishing; 2015.
2. Lim D, Ha M, Song I. Trends in the leading causes of death in Korea, 1983-2012. J Korean Med Sci 2014;29:1597-1603.
3. Saxena S, Funk M, Chisholm D. World health assembly adopts comprehensive mental health action plan 2013-2020. Lancet 2013;381:1970-1971.
4. Caldwell CB, Gottesman II. Schizophrenics kill themselves too: a review of risk factors for suicide. Schizophr Bull 1990;16:571-589.
5. Fawcett J, Scheftner W, Clark D, Hedeker D, Gibbons R, Coryell W. Clinical predictors of suicide in patients with major affective disorders: a controlled prospective study. Am J Psychiatry 1987;144:35-40.
6. Large M, Smith G, Sharma S, Nielssen O, Singh SP. Systematic review and meta-analysis of the clinical factors associated with the suicide of psychiatric in-patients. Acta Psychiatr Scand 2011;124:18-29.
7. Cash SJ, Thelwall M, Peck SN, Ferrell JZ, Bridge JA. Adolescent suicide statements on MySpace. Cyberpsychol Behav Soc Netw 2013;16:166-174.
8. Huang YP, Goh T, Liew CL. Hunting suicide notes in web 2.0-preliminary findings. Ninth IEEE International Symposium on Multimedia Workshops, 2007, p.517-521.
9. O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting suicidality on Twitter. Internet Interv 2015;2:183-188.
10. Burnap P, Colombo W, Scourfield J. Machine classification and analysis of suicide-related communication on twitter. Proceedings of the 26th ACM Conference on Hypertext & Social Media, 2015, p.75-84.
11. Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, et al. Tracking suicide risk factors through Twitter in the US. Crisis 2014;35:51-59.
12. Won HH, Myung W, Song GY, Lee WH, Kim JW, Carroll BJ, et al. Predicting national suicide numbers with social media data. PLoS One 2013;8:e61809.
13. Yamashita T, Yamashita K, Kamimura R. A stepwise AIC method for variable selection in linear regression. Commun Stat-Theor Meth 2007;36:2395-2403.
14. Kim YH, Kim H, Kim DS. Association between daily environmental temperature and suicide mortality in Korea (2001-2005). Psychiatry Res 2011;186:390-396.
15. Park IJ, Min KH. Making a list of Korean emotion terms and exploring dimensions underlying them. Korean J Soc Personal Psychol 2005;19:109-129.
16. Rhee JW, Song HJ, Na EK, Kim HS. Classification of emotion terms in Korean. Korean J Journalism Commun Stud 2008;52:85-116.
17. Bollen J, Mao HN, Zeng XJ. Twitter mood predicts the stock market. J Comput Sci 2011;2:1-8.
18. Ceccherini-Nelli A, Priebe S. Economic factors and suicide rates: associations over time in four countries. Soc Psychiatry Psychiatr Epidemiol 2011;46:975-982.
19. Inoue K, Fukunaga T, Okazaki Y. Study of an economic issue as a possible indicator of suicide risk: a discussion of stock prices and suicide. J Forensic Sci 2012;57:783-785.
20. Tsai JF, Cho W. Temperature change dominates the suicidal seasonality in Taiwan: a time-series analysis. J Affect Disord 2012;136:412-418.
21. Kim Y, Myung W, Won HH, Shim S, Jeon HJ, Choi J, et al. Association between air pollution and suicide in South Korea: a nationwide study.

PLoS One 2015;10:e0117929.

22. Phillips DP. The influence of suggestion on suicide: substantive and theoretical implications of the Werther effect. Am Sociol Rev 1974;39: 340-354.

23. Maldonado G, Kraus JF. Variation in suicide occurrence by time of day, day of the week, month, and lunar phase. Suicide Life Threat Behav 1991;21:174-187.

24. Lester D. Temporal variation in suicide and homicide. Am J Epidemiol 1979;109:517-520.

25. Clifford AC, Doran CM, Tsey K. A systematic review of suicide prevention interventions targeting indigenous peoples in Australia, United States, Canada and New Zealand. BMC Public Health 2013;13:463.

26. Robinson J, Cox G, Malone A, Williamson M, Baldwin G, Fletcher K, et al. A systematic review of school-based interventions aimed at preventing, treating, and responding to suicide- related behavior in young people. Crisis 2013;34:164-182.

27. Szumilas M, Kutcher S. Post-suicide intervention programs: a systematic review. Can J Public Health 2011;102:18-29.

28. Lapierre S, Erlangsen A, Waern M, De Leo D, Oyama H, Scocco P, et al. A systematic review of elderly suicide prevention programs. Crisis 2011; 32:88-98.

29. Mann JJ, Apter A, Bertolote J, Beautrais A, Currier D, Haas A, et al. Suicide prevention strategies: a systematic review. JAMA 2005;294:2064-2074.

30. Jobes DA, Berman AL, O'Carroll PW, Eastgard S, Knickmeyer S. The Kurt Cobain suicide crisis: perspectives from research, public health, and the news media. Suicide Life Threat Behav 1996;26:260-269; discussion 269-271.

31. Martin G, Koo L. Celebrity suicide: did the death of Kurt Cobain influence young suicides in Australia? Arch Suicide Res 1997;3:187-198.

32. Gould MS. Suicide and the media. Ann N Y Acad Sci 2001;932:200-221; discussion 221-204.

33. Luxton DD, June JD, Kinn JT. Technology-based suicide prevention: current applications and future directions. Telemed J E Health 2011; 17:50-54.

34. Krysinska KE, De Leo D. Telecommunication and suicide prevention: hopes and challenges for the new century. Omega (Westport) 2007;55: 237-253.

35. Whittaker R, McRobbie H, Bullen C, Borland R, Rodgers A, Gu Y. Mobile phone-based interventions for smoking cessation. Cochrane Database Syst Rev 2012;11:Cd006611.

36. Harrison V, Proudfoot J, Wee PP, Parker G, Pavlovic DH, Manicavasagar V. Mobile mental health: review of the emerging field and proof of concept study. J Ment Health 2011;20:509-524.

37. Palmier-Claus JE, Rogers A, Ainsworth J, Machin M, Barrowclough C, Laverty L, et al. Integrating mobile-phone based assessment for psychosis into people's everyday lives and clinical care: a qualitative study. BMC Psychiatry 2013;13:34.

38. Intille SS. Technological Innovations Enabling Automatic, Context-Sensitive Ecological Momentary Assessment. In: Stone A, Editors. The Science of Real-Time Data Capture: Self-Reports in Health Research. Oxford: Oxford University Press, 2007, p.308-337.

39. Luxton DD, McCann RA, Bush NE, Mishkind MC, Reger GM. mHealth for Mental Health: Integrating Smartphone Technology in Behavioral Healthcare. Prof Psychol Res Pr 2011;42:505-512.

40. Kelly J, Gooding P, Pratt D, Ainsworth J, Welford M, Tarrier N. Intelligent real-time therapy: harnessing the power of machine learning to optimise the delivery of momentary cognitive-behavioural interventions. J Ment Health 2012;21:404-414.

41. Kramer I, Simons CJ, Hartmann JA, Menne-Lothmann C, Viechtbauer W, Peeters F, et al. A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial. World Psychiatry 2014;13:68-77.

42. Stuckler D, Basu S, Suhrcke M, Coutts A, McKee M. The public health effect of economic crises and alternative policy responses in Europe: an empirical analysis. Lancet 2009;374:315-323.

43. Gavrilova NS, Semyonova VG, Evdokushkina GN, Gavrilov LA. The response of violent mortality to economic crisis in Russia. Popul Res Policy Rev 2000;19:397-419.

44. Chang SS, Gunnell D, Sterne JAC, Lu TH, Cheng AT. Was the economic crisis 1997-1998 responsible for rising suicide rates in East/Southeast Asia? A time-trend analysis for Japan, Hong Kong, South Korea, Taiwan, Singapore and Thailand. Soc Sci Med 2009;68:1322-1331.

45. Jeon SY, Reither EN, Masters RK. A population-based analysis of increasing rates of suicide mortality in Japan and South Korea, 1985-2010. BMC Public Health 2016;16:356.

46. Wada K, Kondo N, Gilmour S, Ichida Y, Fujino Y, Satoh T, et al. Trends in cause specific mortality across occupations in Japanese men of working age during period of economic stagnation, 1980-2005: retrospective cohort study. BMJ 2012;344:e1191.

47. Economou M, Madianos M, Theleritis C, Peppou LE, Stefanis CN. Increased suicidality amid economic crisis in Greece. Lancet 2011;378: 1459.

48. Watts J. Seoul - Suicide rate rises as South Korea's economy falters. Lancet 1998;352:1365-1365.

49. Park JS, Lee JY, Kim SD. A study for effects of economic growth rate and unemployment rate to suicide rate in Korea. Korean J Prev Med 2003;36:85-91.

50. Kwon JW, Chun H, Cho SI. A closer look at the increase in suicide rates in South Korea from 1986-2005. BMC Public Health 2009;9:72.

51. Kim H, Song YJ, Yi JJ, Chung WJ, Nam CM. Changes in mortality after the recent economic crisis in South Korea. Ann Epidemiol 2004;14: 442-446.

52. Chang SS, Stuckler D, Yip P, Gunnell D. Impact of 2008 global economic crisis on suicide: time trend study in 54 countries. BMJ 2013;347:f5239.

53. Berk M, Dodd S, Henry M. The effect of macroeconomic variables on suicide. Psychol Med 2006;36:181-189.

54. Zhang X, Fuehres H, Gloor PA. Predicting stock market indicators through twitter "I hope it is not as bad as I fear". Procedia Soc Behav Sci 2011;26:55-62.

55. Platt S. Unemployment and suicidal behaviour: a review of the literature. Soc Sci Med 1984;19:93-115.

56. Wilson SH, Walker GM. Unemployment and health: a review. Public Health 1993;107:153-162.

57. Milner A, Page A, LaMontagne AD. Long-term unemployment and suicide: a systematic review and meta-analysis. PLoS One 2013;8:e51333.

58. Jin RL, Shah CP, Svoboda TJ. The impact of unemployment on health: a review of the evidence. CMAJ 1995;153:529-540.

59. Lambert G, Reid C, Kaye D, Jennings G, Esler M. Increased suicide rate in the middle-aged and its association with hours of sunlight. Am J Psychiatry 2003;160:793-795.

60. Page LA, Hajat S, Kovats RS. Relationship between daily suicide counts and temperature in England and Wales. Br J Psychiatry 2007;191:106-112.

61. Bollen J, Mao HN, Pepe A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM 2011;11:450-453.

62. Woo H, Cho Y, Shim E, Lee K, Song G. Public Trauma after the Sewol Ferry Disaster: The Role of Social Media in Understanding the Public Mood. Int J Environ Res Public Health 2015;12:10974-10983.

63. Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE, et al. Feeling bad on Facebook: depression disclosures by college students on a social networking site. Depress Anxiety 2011;28:447-455.

64. Kumar M, Dredze M, Coppersmith G, De Choudhury M. Detecting changes in suicide content manifested in social media following celebrity suicides. Proceedings of the 26th ACM Conference on Hypertext & Social Media, 2015, p.85-94.