Using Wikipedia to study global suicide patterns

Julie Barberio  
**Health: Science, Society, and Policy Program, Brandeis University, Waltham, MA**

Mike P. Petkov  
**Department of Anesthesia, Perioperative and Pain Medicine, Boston Children’s Hospital, Harvard Medical School, MA**

Clas Linnman  
**Department of Anesthesia, Perioperative and Pain Medicine, Boston Children’s Hospital, Harvard Medical School, MA**

We conduct much of our lives online, and these activities leave traces that may be useful in epidemiological research. This study evaluated if global, seasonal and weekly variability in suicide incidence is reflected in the number of visits to Wikipedia’s article on suicide. Data on daily visits to the Wikipedia article on “suicide” were obtained for 36 languages, and compared to the literature on regional suicide incidence, and to seasonal and weekday patterns when available. There was no correlation between suicide incidence and the number of visits to Wikipedia’s article on suicide in respective language regions. Further, suicide incidence and article views according to season were found to be different across all language regions. Free and publically available information sources on online behavior holds promise to complement traditional epidemiological methods. However, at present, the publicly available Wikipedia user data does not appear to provide a good proxy for rapid assessments of global trends in the epidemiology of suicide.

Keywords: Wikipedia, Internet, global, seasonal

**Corresponding author:** Clas Linnman, Boston Children’s Hospital, 9 Hope Avenue, Waltham, MA 02453, USA  
clas.linnman@childrens.harvard.edu

**Acknowledgements:**  
We would like to thank Erik Zachte and Leila Zia for a helpful discussion, and Dr. Vasiliy S. Usenko for kindly providing data on suicides in Dnipropetrovsk, Ukraine.
The International Telecommunication Union estimates that three billion people used the Internet in 2014, and Internet user data provides new sources for scientific investigations on behaviors and interests across the globe (Golder & Macy, 2011). Prior studies indicate that searches for “suicide” or “how to commit suicide” may be an indicator for suicidal behavior (Gunn & Lester, 2013; Song et al., 2014; Sueki, 2011), although the precision of existing data needs to be improved (Fond et al., 2015; Page, Chang, & Gunnell, 2011). Social support and coping methods may be found online, but pro-suicide webpages and information on suicide methods are also present (Daine et al., 2013; Dunlop, More, & Romer, 2011; Gunnell et al., 2012; Luxton, June, & Fairall, 2012; Wong et al., 2013). Studies have found suicide-related Twitter messages (Jashinsky et al., 2014) and dysphoria-related weblog (Won et al., 2013) to correlate with regional suicide frequencies. However, such studies are often language specific, making cross-cultural comparisons difficult. Wikipedia, the free-access, free-content Internet encyclopedia, supported and hosted by the non-profit Wikimedia Foundation, may provide the possibility to study behavior cross-culturally, as the encyclopedia has millions of users in multiple languages.

Suicide incidence varies considerably by country (Varnik, 2012), season (Ajdacic-Gross et al., 2010a) and week-day (Maldonado & Kraus, 1991). Here, we explore if statistics related to the traffic to Wikipedia’s article on “suicide” in different languages can be used as an indicator of such regional and temporal variations in suicide incidence.

We hypothesized that the number of Wikipedia suicide article visits would display

i) variations by language region that correlate to suicide rates in the respective language regions, and

ii) seasonal- and

iii) weekly- variations that correlate to published reports of seasonal and weekly variability in the respective language regions.

Methods

Suicide incidence

National suicide prevalence rates were obtained from (WHO, 2012) and (Varnik, 2012).

Suicide seasonality and weekday variations

Data on seasonal variability in suicide incidence was obtained for 25 countries from the following sources: Brazil (Bando & Volpe, 2014), Croatia (Marcikic et al., 2003), Denmark (Postolache et al., 2010), Finland (Partonen et al., 2004), France (Souetre et al., 1987), Germany (Lukaschek et al., 2014; Massing & Angermeyer, 1985), Greece (Christodoulou et al., 2009; Vougiouklakis et al., 2005), Hungary (Toro et al., 2009; Zonda et al., 2008), Italy (Altamura et al., 1999; Rocchi et al., 2007), Iran (Shojaei et al., 2013), Japan (Liu et al., 2013; Ohtsu et al., 2009), Kazakhstan (Grijbovskii et al., 2013), Lithuania (Kalediene & Petrauskiene, 2004), Netherlands (Hoogenboezem & Netherlands, 2003), Norway (Morken, Lillegård, & Linaker, 2002; Norway, 2012), Poland (Młodozeniec et al., 2010), Portugal (Campos et al., 2002), Romania (Jung, Matei, & Hecser, 2009), Serbia (Penev, 2014), Slovakia (Kancirova & Kudela, 2014), South Korea (Korea, 2014), Sweden (Makris et al., 2013), Turkey (Asirdizer et al., 2010; Enginyurt et al., 2014), Ukraine (Usenko et al., 2014), and the United States of America (USA) (Miller et al., 2012).

Data on weekday variability in suicide incidence was available for seven countries:
Croatia (Marcikic et al., 2003), Hungary (Zonda et al., 2008), Italy (Altamura et al., 1999), Japan (Ohtsu et al., 2009), Lithuania (Kalediene & Petrauskiene, 2004), the Netherlands (J. & Netherlands, 2002) and USA (Miller et al., 2012).

Wikipedia data

Publicly available data on the daily number of visits to the Wikipedia article on “suicide” was extracted across 36 languages between 2008 and 2013. Three metrics were used: The Wikipedia page rank, the number of daily visits to the suicide page, and the number of daily visits to the Wikipedia main page (as a means for controlling for general traffic variations).

First, the “suicide” page view statistics for Wikimedia projects was extracted from http://dumps.wikimedia.org/other/pagecounts-raw/ by means of the online tool http://stats.grok.se/. We extracted the page ranking of the suicide article, and daily number of visits from the earliest available source (tracking of daily page visits began in early 2008) until December 31, 2013. We also extracted the daily number of visits to the Wikipedia “main” page as a means of controlling for general temporal fluctuations in Wikipedia traffic.

We extracted data for the largest Wikipedia languages, with a cutoff at >0.05% of the global share of Wikipedia traffic. This included 36 languages (Arabic, Bulgarian, Catalan, Chinese, Croatian, Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hebrew, Hungarian, Indonesian, Italian, Japanese, Kazakh, Korean, Lithuanian, Malay, Norwegian, Persian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Spanish, Swedish, Thai, Turkish, Ukrainian and Vietnamese) accounting for 95.5% of all the Wikipedia traffic. User data from alternative methods of accessing Wikipedia content were excluded: “portal” (3.3% of the global share), “mobile” (0.09%) and “Simple English” (0.06%).

Data considerations

Both Internet penetrance and Wikipedia use varies across nations. Wikipedia does not provide a breakdown by country for traffic to specific articles, hampering our ability to link user statistics to epidemiological data on suicide incidence from individual countries. For some languages, such as Kazakh, Thai and Hebrew, over 95% of the Wikipedia traffic comes from Kazakhstan, Thailand and Israel, respectively, providing relatively fine-grained spatial information. For global languages, however, the Wikipedia sites are shared across countries, for example, the public data does not distinguish between a visit to the Spanish article on “Suicidio” originating from Spain, Chile or Mexico. Wikipedia does however report the total traffic by country, which can be used as an approximation of regional web traffic. Thus, a suicide-incidence-by-language-region index was calculated, where we weighted each nation’s suicide rate (from epidemiological literature reports, (WHO, 2012) and (Varnik, 2012)) by the relative national use of Wikipedia. For example, the Spanish Wikipedia was accessed from Mexico 25%, Spain 20%, Colombia 12.5%, Argentina 9.5%, Venezuela 6.6%, Chile 4.4, and Peru 4.4% of the time. The Spanish suicide-incidence-by-language-region index is thus a composite of those respective countries’ suicide incidence weighted by relative Wikipedia use. Countries accounting for less than 3% of visits for a particular language were excluded.

Over the time period investigated (2008 to 2013), most regions have experienced an increase in Internet access. In 2008, 23% of the world population had Internet access,
whereas 38% had access in 2013 (www.internetlivestats.com). The rate of increase differs between countries — in South Korea 81% of the population had Internet access in 2008 and 85% in 2013. In the same time period the Ukraine went from 11% to 42%. The number of visits to Wikipedia parallels this increase in access. To control for such generally increasing trends, and to control for temporal fluctuations in Wikipedia traffic, we used the number of visits to the Wikipedia main page in each language region to calculate the ratio between suicide and main page visits. For the seasonal and weekday analysis, we averaged user data over the six years, and over each week of the six years respectively. We limited Wikipedia data collection to 2008-2013 as changes in Wikipedia usage (mobile platforms) are ongoing, and suicide frequency statistics was most available for this time (i.e. official registry data is typically delayed).

**Results**

Data on daily Wikipedia suicide page visits was available for 93% of time between February 1, 2008 and December 31, 2013 in the 36 languages investigated. The articles on suicide ranked between the 512th (Finnish) and 7162nd (Slovak) most commonly visited articles on Wikipedia. There was no significant correlation between the suicide-/main-page visit ratio and suicide incidence, \(n = 36, r^2 = 0.05, p = 0.21\), or between the page ranking and the suicide incidence in the respective language regions \(n = 36, r^2 = 0.07, p = 0.12\).

**Seasonal variations**

Pooling the literature reports on seasonality, we found a general pattern of peak suicide incidence in May, and a December trough (Figure 1). Wikipedia suicide page visits displayed varying degrees of seasonality, with a general trend of peaks in November and troughs in July-August (Figure 1). Further, in the 25 countries investigated, there was no significant correlation between seasonality of Wikipedia suicide page visits and seasonality of suicide in any of the investigated countries, with the exception of Hungary/Hungarian, where there was a significant negative correlation between suicide seasonality and Wikipedia suicide page visits \(r = -0.59, p = 0.043\).

**Weekday variability**

The weekly variation in suicide incidence paralleled the weekly number of visits to the Wikipedia suicide page in Croatian (“samoubojstvo”, \(r = 0.96, p = 0.0003, \text{one tailed}\), Dutch (“zelfmoord”, \(r = 0.91, p = 0.002, \text{one tailed}\), English (“suicide” \(r = 0.89, p = 0.004, \text{one tailed}\), Italian (“suicidio”, \(r = 0.84, p = 0.009, \text{one tailed}\), Japanese (“自殺” \(r = 0.76, p = 0.02, \text{one tailed}\), and Lithuanian (“savįžudybė”, \(r = 0.67, p = 0.049, \text{one tailed}\) but not for Hungarian (“öngyilkosság”, \(r = 0.51, p = 0.12, \text{one tailed}\), see Figure 2.

Among Jewish men in Israel, suicide incidence is highest on Sundays (start of the working week) and lowest on Fridays (Sabbath) (suicide incidence data for all week days was not reported) (Modan, Nissenkorn, & Lewkowski, 1970). This weekly pattern was reflected in the number of visits to the Wikipedia suicide article in Hebrew (“תודבאתה”).

However, when controlling for general Wikipedia usage with the suicide-/main-page visits ratio, none of the above correlations remained significant.
Contrary to our hypothesis, we found that the ranking of Wikipedia’s article on suicide in various languages did not correlate to suicide incidence across the included countries. While the weekly patterns of Wikipedia visits corresponded well to the weekly distribution of suicide in six out of seven language regions, this effect disappeared when controlling for general Wikipedia usage across the week. The seasonal distribution of Wikipedia visits had no relation to the seasonality of suicide observed in many countries. Rather, the number of visits to the article on suicide appears to display an opposite seasonal pattern compared to actual suicides (Figure 1).
Seasonal variations in suicide incidence is a longstanding observation, and a multitude of explanations have been put forward, such as amount of sunshine, temperature related alterations in neuroendocrine function, sociological aspects, and the availability of means to commit suicide (Ajdacic-Gross et al., 2010; Durkheim, 1897). The observation that Wikipedia suicide page visits peak in November and trough in the summer months deserves some speculation: There is a widespread belief among non-experts that suicide is more common in the fall and winter and less common in late spring and summer (Voracek, Tran, & Sonneck, 2007). This myth corresponds quite well to the observed Wikipedia page visits, perhaps suggesting that, at least for North America and Europe, dark rainy and cold November days may have people thinking and reading about suicide, while actual suicide appears triggered by other factors. As such, Wikipedia user data may be more sensitive to cultural trends and beliefs than actual suicide incidence.

**Limitations**

Our findings highlight the potential — and the challenges — in Internet user generated data sources for epidemiological studies. Online behavior data are attractive for their unprecedented quantities of large amounts of data sampled in close-to real time, often from diverse geographical areas. The “organic” nature of Wikipedia, i.e. rapid growth, rapid technical development (such as mobile traffic) and continuous user editing makes it a moving target, where both user statistics and information structure on Wikipedia itself is an important object of study (Bizer et al., 2009). There are several limitations, both to the current study, and the use of online data, outlined below.

i) Data sampling and noise: The user data provided by Wikipedia are not unique visitors, it is “pages loaded” when accessed at the normal Uniform Resource Locator (URL), i.e. http://en.wikipedia.org/wiki/Suicide. It is not sampled data, has not checked for outliers, contains much impure data (such as bots loading a page continuously), and it does not include requests to the mobile site, which is expected to serve about half of the pageviews at some point in 2015 (Killiondude, 2015).

ii) Data granularity: The publicly available Wikipedia user statistics used here provide only very coarse grained data on geographical location, especially for Portuguese, Spanish, English and Arabic, given the wide spread of these languages. Much better estimates of regional associations could potentially be obtained by tracking IP addresses (or GPS coordinates for mobile users). However, such data may infringe on user privacy and is therefore not provided by the Wikimedia foundation (Wikimedia, 2015).

iii) Sample characteristics: In 2010, 79% of the American population used the Internet, whereof 56% male and 50% female users reported using Wikipedia as a source of information (Zickuhr & Rainie, 2011). Wikipedia usage was most popular among Internet users with a college degree, with broadband connections, annual incomes over $50,000 and amongst young adults. This stands in contrast to a four times higher suicide rate among American men than women, and low socioeconomic status as a risk factor (Li et al., 2011). Moreover, in countries with lower Internet penetrance, it is conceivable that Internet access is most limited amongst individuals with lower socioeconomic status. Thus, while usage and Internet access are increasing, Wikipedia user statistics may not (yet) be representative for the global population.

iv) Meaning of metric: Visiting Wikipedia’s article on suicide is, in the majority of cases, not a sign of suicidal ideation. A more operational definition may be that the Wikipedia
user statistics provide data representing a select population’s deliberate information seeking.

vi) Article content: The length and quality of Wikipedia’s article on suicide varies considerably between languages. For example, the English article contains 11,500 words and 185 references (July 7, 2015), while the Slovak article contains 88 words. It is likely that more interest in the topic will lead to more extensive authoring and editing of content, and more extensive articles may similarly generate higher usage.

Defining benchmarks: Official suicide data are influenced with regard to the judicial use of the word suicide in determining the cause of death (Kelleher et al., 1996), and there may be uncertainty as to whether a death is a pure unintentional fatality or a suicide. Variations in methods registering unnatural deaths may reduce the validity of suicide mortality data and comparing suicide rates between countries is no simple task (Atkinson, Kessel, & Dalgaard, 1975; Gjertsen et al., 2013). Further, official suicide data rarely indicates the ethnicity of the victims, which further complicates comparisons. For example, in Saudi Arabia, 16 to 19% of suicide victims are Saudi nationals, while foreign workers (primarily Indians) comprise more cases (Al Madni et al., 2010; Helaly, Ali, & Zidan, 2015). Similarly, in Dubai suicide rates among expatriates is seven times higher than among nationals (Dervic et al., 2012).

In conclusion, Wikipedia user statistics on visits to the article on suicide do not appear to reflect regional suicide incidence variability, and we found no evidence that such data can be used to study seasonal suicidality trends. Internet user data (webpage statistics, search engine data, social media, etc.) are an attractive complement — but not replacement — to traditional epidemiological sources, and care needs to be taken when evaluating these types of data.

References


Using Wikipedia to study global suicide patterns


