Beyond demographics: How search engine data can enhance the prediction of suicide rates in India and inform prevention

Authors:
Natalia Adler, UNICEF, New York, USA
Ciro Cattuto, ISI Foundation, Turin, Italy
Kyriaki Kalimeri, ISI Foundation, Turin, Italy
Daniela Paolotti, ISI Foundation, Turin, Italy
Michele Tizzoni, ISI Foundation, Turin, Italy
Elad Yom-Tov, Microsoft Research
Stefaan Verhulst, The Governance Lab, New York University
Andrew Young, The Governance Lab, New York University

Abstract

Background
India is home to 20% of the world’s suicide deaths. In India, and around the world, young people are especially at risk of suicide. While statistics regarding suicide in India are distressingly high, data and cultural issues likely contribute to a widespread underreporting of the problem. Social stigma and only recent de-criminalization of suicide are but two factors hampering official agencies’ collection and reporting of suicide rates.

Objective
As the product of a data collaborative – the cross-sector exchange of data to create new public value – this paper leverages private-sector search engine data toward gaining a fuller, more accurate picture of the suicide issue among young people in India. By combining official statistics on suicide with data generated through search queries, this paper seeks to: 1) add an additional layer of information to more accurately represent the magnitude of the problem; 2) determine whether search query data can serve as an effective proxy for factors contributing to suicide that are not represented in traditional datasets; and 3) consider how data collaboratives built on search query data could inform future suicide prevention efforts in India and beyond.

Methods
We combined official statistics on demographic information with data generated through search queries from Bing to predict suicide rates per state in India as reported by the National Crimes Record Bureau of India.

We have extracted English language queries on five topics ("suicide", "depression", "hanging", "pesticide", "poison"). For each query, we recorded the time and date of the query, the state in India from which the user made the query, and the text of the query. We have then collected data on demographic information at state level in India, including: Urbanization, Growth Rate, Sex Ratio, Internet Penetration, Population. We have modeled the suicide rate per state as a function of the queries on each of the 5 topics considered as linear independent variables. We also built a second model by integrating the demographic information on Urbanization, Growth Rate, Sex Ratio, Internet Penetration and Population, all considered as additional linear independent variables in the model.

**Results**

Results of the first model fit ($R^2$) when predicting the suicide rates from the fraction of queries in each of the 5 topics, as well as the fraction of all suicide methods, show a correlation of about 0.5. The correlation increases significantly with the removal of even 3 outliers, and improves slightly when 5 outliers are removed. In all cases, statistically significant correlation is reached, but the best correlation is obtained for suicide methods (hanging, pesticide, and poison), and only to a lesser extent for depression. Results for the second model fit using both query data and demographic data show that for all categories, if no outliers are removed, demographic data predict suicide rates better than query data. However, when 3 outliers are removed, query data about pesticides or poisons improves the model over using demographic data.

**Conclusions**

Internet search data has been shown in previous work to serve as a proxy for many health-related behaviors, enabling the measurement of rates of different conditions ranging from influenza to suicide. In this work, we used both search data and demographics to predict suicide rates. In this way, search data serves as a proxy for unmeasured (hidden) factors corresponding to suicide rates. Moreover, our procedure for outlier rejection serves to single out states where the suicide rates have substantially different correlations with both demographic factors and query rates.
Introduction

According to the World Health Organization (WHO) [1], close to 800,000 people kill themselves by suicide every year, with 78% of global suicides occurring in low- and middle-income countries [1]. Teenagers and young adolescents are particularly at risk, as suicide represents the second leading cause of death among 15-29 year-olds worldwide [1]. These concerning figures do not even fully capture the magnitude of the problem. The WHO estimates that good quality data on suicide exists for only 60 countries worldwide.

According to official statistics, India is home to 20% of the world’s suicide deaths [2], yet the issue attracts limited national public health attention [3]. In addition, statistics on suicides released by the Indian National Crime Records Bureau (NCRB) are insufficient to understand the magnitude of the problem. In 2013, the NCRB reported that 134,799 people died of suicide, making the suicide rate 11% of total death [4]. However, evidence from other studies shows that the NCRB’s suicide rate data is grossly underreported. For instance, WHO reported 170,000 cases of suicide deaths in India, which is about 35,000 higher than the NCRB’s data [3]. Similarly, the Registrar General of India implemented a nationally representative mortality survey indicating that about 3% of the surveyed deaths (2,684 of 95,335) in individuals aged 15 years or older were due to suicide, corresponding to about 187,000 suicide deaths in India in 2010 (as reported in Patel [3]).

The factors contributing to these inconsistencies are likely manifold – including both data collection barriers and cultural challenges. The deep-rooted stigma associated with mental disorders, coupled with limited suicide prevention and mental health services, makes it difficult to address suicide as a major public health problem in India [5]. And until recently, suicide was a criminal offense in the country – likely compelling families to report suicides as death by an illness or accidents to avoid punishment [6]. Moreover, analysis (if any) of suicide records is limited to demographic correlations. Patel [3], for instance, only focused on age- and sex-variables to analyze the survey findings.

This paper seeks to contribute to this discussion by describing how data gaps in Indian suicide reporting can be filled through the creation of data collaboratives. Data
collaboratives are “an emerging form of public-private partnership in which actors from different sectors exchange information to create new public value.” [7]. Data collaboratives are increasingly being tested as means for improving evidence-based policymaking and targeted service delivery around the world, including notably data sharing arrangements between corporations and national statistical offices [8].

By combining official statistics on suicide with data generated through search queries, this paper seeks to: 1) add an additional layer of information to more accurately represent the magnitude of the problem; 2) determine whether search query data can serve as an effective proxy for factors contributing to suicide that are not represented in traditional datasets; and 3) consider how data collaboratives built on search query data could inform future suicide prevention efforts in India and beyond.

**What is known on who commits suicide in India?**

The literature varies and sometimes offers a convoluted picture of who is at risk of committing suicide – in India and beyond. There seems to be consensus that young people aged 15-29 are particularly at risk. Patel [3] found that 40% of suicides among men, and 56% of suicides among women occurred between the ages of 15 and 29. However, while the prevalence of suicide among the young is generally accepted, the male-to-female suicide ratio in India varies greatly, ranging from 1.04 to 1.63 according to some studies [6], while official statistics from the NCRB shows a male-to-female suicide ratio at 2:1. Patel [3] on the other hand, found a rate of 26.3 suicides among boys and men over 15 per 100,000 people compared to 17.5 for girls and women – demonstrating the need for more clarity. The paper also found that Indian boys and men had a 1.7% cumulative risk of dying by suicide between the ages of 15 and 80, compared to a 1.0% risk among girls and women [4].

There exists a strong correlation between educational backgrounds and suicide risks, with the less educated accounting for 70.4% of suicide cases as recorded in the NCRB data [6]. Counterintuitively, suicide among students in India is also increasing – moving from 5.5% of total suicides in 2010 to 6.2% in 2013 [9].
Suicides rate vary by occupation [4]. Housewives accounted for about 18% of the total victims, while farmers comprise 11.9% of the total victims followed by those working in the private sector (7.8%), unemployed (7.5%) and public sector (7.8% and 2.2%, respectively). A study reported that approximately 16,000 farmers in India commit suicide every year [10]. Patel [3] found that about half of suicide deaths in India arose from poisoning, especially resulting from the ingestion of pesticides.

*Risk Factors*

Some population groups are more at risk than others. For example, a study shows that 27.2% of primary care patients suffer from *depressive disorder*, and 21.3% of them have attempted suicide, demonstrating how depression is one of the underlying factors that drives suicide [11].

Risk factors cut across geographical lines, with official statistics showing a significantly higher rate of suicide taking place in *southern states*, such as Tamil Nadu, Andhra Pradesh, Karnataka, Kerala, West Bengal, and Maharashtra, where 63.6% of suicide cases occurred [4].

These data paint a picture of the breadth and diversity of the suicide issue in India. But given the stigma associated with suicide, poor quality data, and the still-recent decriminalization of suicide attempts, these statistics confuse as much as they elucidate. The country’s underreporting challenge – and the likely neglect of certain population groups altogether – creates major challenges for meaningfully determining who is at risk.

*Internet data as a source for health information*

In response to the issue of suicide underreporting in India, this paper looks at a specific cohort of people (*English-speaking Internet users*) to add another layer of understanding about those at risk of committing suicide. The majority - or one third - of Internet users are young (18-35 years) [12], which coincides with the age group most at risk for suicides (15-29 years).

With its 360 million users (26% of the population), India is home to more Internet users than any country save China [13]. Men dominate internet usage in India [14] with 71%
to women’s 29%. Internet usage is more prevalent in northern (27%) and western states (25%) compared to the South (19%) and East (16%) [12]. Yet not all Indians accessing the Internet do so in English. The Indian Constitution recognizes 22 official languages and the number of Indian language Internet users have grown dramatically over the years, surpassing English users: 234 million compared to 175 million, respectively.

Internet data has been used to monitor health behaviors, for a variety of conditions ranging from infectious diseases [15]–[17] to mental health conditions [18]. Anonymous venues online, especially search engines [19] allow people to seek information on sensitive topics. Monitoring such venues can thus offer a window into behaviors which are otherwise difficult to study. Specifically, in the case of suicides, social media has been used to detect suicidality [20], and search engine logs were utilized to analyze suicides in general [21], [22] and the Werther Effect (“copycat suicides”) in particular [21].

Here we examine Internet search engine logs for information about suicides. Search engine logs, as analyzed here, focus on a population of English-speaking people in India. The market share of Bing in India has been reported to be around 7% at the time of collection [23]. Moreover, as shown in Fisher and Yom-Tov [21], people seeking information on suicides via search engines are (at least in the United States) people who are contemplating suicide, not people who may necessarily successfully commit suicide.

*Internet search queries as a source for suicide-related information*

This paper’s methodology, described below, builds on previous work leveraging search query data analysis. Numerous studies have found that search query data is reflective of behaviors in the physical world [24]. In the United States, for example, people searching for actionable information about suicides (how to commit suicides) correspond to the population that attempts suicide – but not the population that successfully commits suicide [25].

Several studies have analyzed Google Trends, an aggregate measure of search query volume, and found correlations between search queries for suicide and rate of suicide. Gunn and Lester [19] found a correlation between volume of queries about suicide and actual number of suicides by analyzing search words and phrases like “how to suicide.”
Hagihara et al. [22] conducted a study in Japan that shows how suicide queries spike in the period before there is an increase in suicide rate [24]. This method was replicated in Taiwan and Australia, but those studies yielded contradicting results [21]. Other studies are most skeptical about the correlation between Google queries and suicide rates, concluding that a tool to identify relevant search queries must be further developed in order to create a more precise prediction mechanism [27].

Finally, a study by Kristoufek, et al. [28] studied how data on the number of Google searches for the terms ‘depression’ and ‘suicide’ in England relate to the number of suicides between 2004 and 2013. The researchers found that estimates drawing on Google data were significantly more accurate than estimates relying on previous suicide data alone. Interestingly, their findings show that a greater number of searches for the term ‘depression’ is related to fewer suicides, whereas a greater number of searches for the term ‘suicide’ is related to more suicides, though the correlation is not extremely high ($R^2$ of about 0.4).

**Methods**

**Search engine data**

We extracted all English language queries from the Bing search engine submitted by people from India between November 2016 and February 2017 (inclusive). For each query we recorded the time and date of the query, the state in India from which the user made the query, and the text of the query. The correlation between the number of queries per state and the population of that state, multiplied by internet penetration, provided a positive Spearman correlation coefficient of rho=0.93 (P-value < $10^{-12}$). The queries on five topics were identified by testing whether the text of the queries contained one or more of the inclusion terms in Table 1, and not containing any of the exclusion terms. The exclusion terms were found by identifying the most common words and word pairs appearing in conjunction with the inclusion terms, and identifying those which were unrelated to suicidal intentions.
State data on suicide rates

The suicide rate per state was obtained from the 2014 Accidental Deaths & Suicides in India report from the National Crimes Record Bureau of India [29]. (See supplementary material for a table with the data).
<table>
<thead>
<tr>
<th>Topic</th>
<th>Inclusion terms</th>
<th>Exclusion terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicide</td>
<td>Suicide, kill myself</td>
<td>'suicide squad', 'song', 'download', 'skill', 'killer', 'movie', 'video', 'bill',</td>
</tr>
<tr>
<td></td>
<td></td>
<td>'game', 'lyrics', 'mp3', 'suicide girl', 'militia', 'mockingbird', 'ghandi',</td>
</tr>
<tr>
<td></td>
<td></td>
<td>'akame ga kill', '3 days to kill', 'wifi kill', 'kill dil', 'kill zone', 'killzone',</td>
</tr>
<tr>
<td></td>
<td></td>
<td>'kill em with kindness', 'kill me heal me', 'rkill'</td>
</tr>
<tr>
<td>Depression</td>
<td>Depression, depressed</td>
<td></td>
</tr>
<tr>
<td>Hanging</td>
<td>Hang, hanging</td>
<td>'wall hanging', 'hanging garden', 'macrame'</td>
</tr>
<tr>
<td>Pesticide</td>
<td>Pesticide</td>
<td></td>
</tr>
<tr>
<td>Poison</td>
<td>Poison</td>
<td>'poison ivy', 'poisonous snakes', 'poisoned thoughts', 'poison thoughts', 'food poisoning',</td>
</tr>
<tr>
<td></td>
<td></td>
<td>'hanging boobs', 'hanging lights'</td>
</tr>
</tbody>
</table>

Table 1: Exclusion and inclusion terms for each of the 5 topics related to suicides.

Demographic data

In the data analysis, we have included demographic information at the state level, including: Urbanization, Growth Rate, Sex Ratio, Internet Penetration, Population. Data were obtained as follows:

1. **Sex Ratio, Population, Urbanization, Growth Rate**: From Wikipedia [30].
3. **Internet penetration**: From data published by The Hindu newspaper [32].
4. **Enrolment in higher education**: Gross Enrolment Ratio in Higher Education, available from the Statistical Year Book of India, 2016 (Table 29.6) [33]

Modeling for the search engine data

Data were analyzed for their temporal patterns (diurnal and weekly) as well as their variation by state. We modeled the suicide rate per state as a function of the queries on each of the 5 topics. Thus, the dependent variable in our models was the expected number of
suicides in each state, which are the product of the suicide rate by the size of the population. The independent variables included the fraction of queries (with respect to the number of internet users in the state) from each state in each topic. Outliers were removed using an iterative process: Up to either 3 or 5 states were removed by finding the state which, if removed, increased model fit (R^2) by the greatest amount, and repeating this process three or five times, as desired. The model used throughout is a linear model, unless otherwise stated. We report model fit for different levels of outlier rejection below.

Risks and Data Responsibility

To be clear, the use of data related to suicide, suicidal ideation, and mental health creates some level of risk across the data lifecycle. The analysis described in this paper adhered to strict data responsibility principles, ensuring that sensitive data was not shared or compromised, and that aggregated rather than personally identifiable data informed our findings. For the field at large to effectively and legitimately leverage data collaboratives to improve public understanding of suicide rates and devise evidence-based prevention strategies, data responsibility methods and tools are needed for both sides of data-sharing arrangements. The research that informed this paper is contributing to the development of data responsibility frameworks to aid the field in assessing if, when, and how data can be shared in a responsible manner as part of a data collaborative. This study is considered exempt by the Microsoft Institutional Review Board.

Results

Temporal analysis

Figure 1 shows the percentage of queries about suicide, depression, and suicide methods as a function of the hour of the day and the day of the week. As these figures show, these queries broadly follow the baseline (all queries made in India). However, closer inspection reveals that relevant queries are approximately 20% less likely
compared to the baseline during early morning hours, and up to 15% more likely during evening to late night hours. The largest difference between the baseline and relevant queries when stratified by day of the week is smaller than 5%.

![Figure 1: Diurnal and weekly patterns of relevant queries (suicide, depression, and suicide methods), compared to the baseline of all queries made in India. Days are numbered sequentially from Sunday (1) through Saturday (7).](image)

**Using query data to predict suicide rate**

Table 2 shows the model fit ($R^2$) when predicting the suicide rates from the fraction of queries in each of the 5 topics, as well as the fraction of all suicide methods. As the table shows, the correlation increases significantly with the removal of even 3 outliers, and improves slightly when 5 outliers are removed. In all cases, statistically significant correlation is reached, but the best correlation is obtained for suicide methods (hanging, pesticide, and poison), and only to a lesser extent for depression. This indicates that people who are considering suicide are not asking about the term itself, but rather about possible precursors (depression) and methods of suicide.

<table>
<thead>
<tr>
<th>Query</th>
<th>$R^2$ (no outliers removed)</th>
<th>$R^2$ after removal of 3 outliers</th>
<th>$R^2$ after removal of 5 outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hanging</td>
<td>0.29</td>
<td>0.49</td>
<td>0.65</td>
</tr>
<tr>
<td>Pesticide</td>
<td>0.16</td>
<td>0.71</td>
<td>0.80</td>
</tr>
<tr>
<td>Poison</td>
<td>0.33</td>
<td>0.65</td>
<td>0.72</td>
</tr>
<tr>
<td>All methods of suicide</td>
<td>0.47</td>
<td>0.68</td>
<td>0.80</td>
</tr>
<tr>
<td>(hanging, pesticide, poison)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suicide</td>
<td>0.13</td>
<td>0.65</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Table 2: Model fit ($R^2$) for predicting the expected number of suicides in each state from the fraction of queries in each topic. Excluded are states with less than 0.25% of the Indian population (n=20).

We next analyze the outliers and whether, given a model that was constructed after exclusion of an outlier, more suicides would be predicted by query volume, compared to official statistics, or vice versa. A negative outlier would thus indicate that more suicides are predicted according to the volume of queries in a topic than are reported by official data. A positive outlier would indicate the reverse: the official reported suicide rate is greater than that which would be inferred from the queries.

Analyzing the models after excluding 5 outliers per model we find that there are slightly more negative outliers than positive ones: 16 negative outliers compared to 14 positive outliers. Figure 2 shows the outliers. From this figure, it can be seen that the outliers are not distributed randomly: In all but one case, a state will either have all positive or all negative outliers. The states with most negative outliers are Jammu & Kashmir and Jharkhand (4 and 3 outliers, respectively), and the ones with most positive outliers are Telangana and Gujarat (5 and 4 outliers, respectively).
Our findings singled out two states, **Jammu & Kashmir** and **Jharkhand**, as having more queries indicative of suicide than would be expected, given the published suicide rates in these states. On the opposite, **Gujarat** and **Telangana** have more reported suicides than predicted by search data. We attribute the latter to the fact that agriculture plays a big role in these two states, and thus suicides by farmers, who may not be regular internet users, are a large population of people who commit suicide but are unobserved by search data.

*Addition of demographic data to query data*

Demographic data are correlated with suicide rate. Therefore, in this section we investigate whether query data can add to the prediction of suicide rates, beyond what demographic data can provide. To overcome the dimensionality of additional variables we employ a stepwise linear model, which selects the most predictive variables under the criteria that the p-value for an F-test of the change in the sum of squared error is maximally reduced by adding a variable.

Table 3 shows the model fit for models using solely query data and for models using both query data and demographic data. The variables selected when using just demographic data are urbanization rate and sex ratio, both positively correlated with suicide rates.

Table 3 shows that for all categories, if no outliers are removed, demographic data predict suicide rates better than query data. However, when 3 outliers are removed, query data about pesticides or poisons improves the model over using demographic data.

<table>
<thead>
<tr>
<th>Query</th>
<th>Without demographic data</th>
<th>With demographic data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R^2 (no outliers removed)</td>
<td>R^2 after removal of 3 outliers</td>
</tr>
<tr>
<td>Hanging</td>
<td>0.28</td>
<td>0.49</td>
</tr>
</tbody>
</table>
Using both demographics and query data, the outlying states for pesticides are all negative (more suicides predicted by web data than reported): Andhra Pradesh, Kerala, Punjab. The outliers for poison are: Telangana (positive), Delhi (negative), Jharkhand (negative). Thus, the sole positive outlier is the same as using query data alone.

Comparing the outliers, we obtained with the query data only with those obtained when including the demographics data, results are rather similar (Table 4). We report only outliers for pesticides and poison because these are the only keywords for which query data was selected for inclusion in the model.

### Table 3: Model fit \( (R^2) \) for predicting the expected number of suicides in each state from the fraction of queries in each topic, with and without demographics, using a stepwise model. Stars denote cases where query data was not selected for inclusion in the model. Excluded are states with less than 0.25% of the Indian population (n=20).

<table>
<thead>
<tr>
<th></th>
<th>Demographics + query data model</th>
<th>Query data only</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pesticide</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.16 0.71 0.51* 0.91</td>
<td></td>
</tr>
<tr>
<td><strong>Poison</strong></td>
<td>0.33 0.65 0.51* 0.76</td>
<td></td>
</tr>
<tr>
<td><strong>All methods of suicide (hanging, pesticide, poison)</strong></td>
<td>0.47 0.68 0.47 0.74</td>
<td></td>
</tr>
<tr>
<td><strong>Suicide</strong></td>
<td>0.13 0.34 0.51* 0.75*</td>
<td></td>
</tr>
<tr>
<td><strong>Depression</strong></td>
<td>-0.01 0.65 0.51* 0.75*</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: Outliers (with rejection of 3 states) for pesticides and poison queries, for models which use demographic data and query data and for models which only use query data. In parenthesis are the direction of the outlier.

<table>
<thead>
<tr>
<th></th>
<th>Demographics + query data model</th>
<th>Query data only</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pesticides</strong></td>
<td>● Andhra Pradesh (-) ● Kerala (-) ● Punjab (-)</td>
<td>● Andhra Pradesh (-) ● Jammu and Kashmir (-) ● Punjab (-)</td>
</tr>
<tr>
<td><strong>Poison</strong></td>
<td>● Telangana (+) ● Delhi (+) ● Jharkhand (-)</td>
<td>● Telangana (+) ● Jharkhand (-) ● Madhya Pradesh (+)</td>
</tr>
</tbody>
</table>
Discussion

Internet search data has been shown in previous work to serve as a proxy for many health-related behaviors, enabling the measurement of rates of different conditions ranging from influenza to suicide. Here we use these data to predict suicide rates in India. Internet data can be influenced and biased by the population using it. Similarly, official suicide statistics might be susceptible to under-reporting by statistical agencies. In our work we therefore applied two tools that could mitigate these biases. First, we used both search data and demographics to predict suicide rates. In this way, search data serves as a proxy for unmeasured (hidden) factors corresponding to suicide rates. Second, our procedure for outlier rejection serves to single out states where the suicide rates have substantially different correlations with both demographic factors and query rates. To emphasize the difference between these two influences, consider the following cases: First, suppose that in one state there exists a population who commits suicide but does not use the Internet to search for it beforehand. In that case, such a state would be identified as an outlier in our data. Similar cases would occur if there is underreporting in the state, in which case the state would be highlighted as a positive outlier. A different case can occur if a demographic factor which is unavailable to the model predicts suicide rates. In this case, if search queries are influenced by this factor, they will serve as a proxy to suicide rates, and will improve model correlations. We believe that in the data analyzed, both effects are at work. Specifically, some (mostly agricultural states) were found to be negative outliers, suggesting that in those states a population who does not use the internet, or the Bing search engine, are at higher risk of suicide. Similarly, several states were identified as positive outliers, suggesting that in those states underreporting might be occurring. Moreover, search data was shown to improve the predictive ability of models, implying that other demographic factors are also correlative to suicide rates. Future work will identify such factors. Cases in point are Telangana and Andhra Pradesh. The former was recently separated from the latter and declared an independent state. Both are states where agriculture is a major industry, and so where farmer suicides may be expected. However, the first of these was identified as a positive outlier (where fewer suicides were predicted compared from query data) whereas the latter was identified as the opposite, for both models. Several
explanations may be offered for this difference. First, Telangana has registered more suicides due to poverty and unemployment [29]. Second, Telangana has a higher urbanization rate, compared to Andhra Pradesh (39% and 30%, respectively). Third, Telangana has a higher education rate (35% vs. 29%). Both variables are taken into account by the demographics and could thus predict a higher suicide rate in Telangana than is expected solely from query data. This is coupled with the fact that Telangana has recorded a recent increase in student suicides [34]. Finally, in an attempt to curb farmer suicides, Digital India has improved access to information by providing farmers with Internet access [35], a factor which may have contributed to a higher than expected rate of queries from farmers in our data.

Regarding methods of suicide, queries for two methods were found to improve prediction of suicide rate over demographics. Interestingly, searches for depression or the phrase “suicide” did not. This result has two implications. First, individuals searching for methods of suicide, like hanging or poison, are likely at greater risk of suicide than those searching for more general terms related to taking one's own life. Second, efforts to predict the incidence of suicide toward taking preventative action are more likely to find success if they focus on queries about specific methods of committing suicide as opposed to keywords related to depressive symptoms or the concept of suicide more generally.

Most Internet search engines nowadays provide information on helpline numbers in highlighted information boxes above search results when users search for information on suicides. Such pointers to crisis support, particularly in smartphone apps, have been shown to be effective in some studies [36]. To the best of our knowledge, these boxes are only displayed when people explicitly search for the term “suicide”. Our results suggest that future research should investigate the display of these boxes also in cases where people search for methods of suicide. However, it is unclear how to distinguish between searches of methods which are related to suicides and those which are not (e.g., a farmer searching for information on pesticides).

Public awareness campaigns may be a driver of people searching for information about suicides. However, in the data analyzed we found that searches for methods of suicide are better correlated with suicide rate. We suggest that this shows that our data is not
affected by such exogenous factors. We also suggest that additional research is needed to explore the feasibility of creating new restrictions on the sale of pesticides in India, particularly sales to young people.

Building on this work, and drawing upon cross-sector data, including but not limited to search queries, we are conducting research aimed at increasing our understanding of the drivers of suicide among young people in India, how those drivers differ across regions, and how those findings can inform suicide prevention efforts in the country going forward.

Acknowledgments

All the authors of the work would like to express their gratitude to Brinda Adige, Global Concerns India; Trisha Shetty, SheSays; and Purwa Sawant at Unicef for their insightful input and their valuable comments and suggestions. Thanks also to Michelle Winowatan, The Governance Lab, for important research support.

DP, MT, CC and KK acknowledge partial support from the Lagrange Project of the ISI Foundation funded by the CRT Foundation.
References


