Marquette University
e-Publications@Marquette

Computer Science Faculty Research and Publications | Computer Science, Department of

2022

Privacy Concerns with Using Public Data for Suicide Risk Prediction Algorithms: A Public Opinion Survey of Contextual Appropriateness

Michael Zimmer
Marquette University, michael.zimmer@marquette.edu

Sarah Logan
University of Rochester

Follow this and additional works at: https://epublications.marquette.edu/comp_fac

 Part of the Computer Sciences Commons

Recommended Citation

Zimmer, Michael and Logan, Sarah, "Privacy Concerns with Using Public Data for Suicide Risk Prediction Algorithms: A Public Opinion Survey of Contextual Appropriateness" (2022). *Computer Science Faculty Research and Publications*. 69.
https://epublications.marquette.edu/comp_fac/69

Marquette University

e-Publications@Marquette

Computer Sciences Faculty Research and Publications/College of Arts and Sciences

***This paper is NOT THE PUBLISHED VERSION*.**
Access the published version via the link in the citation below.

Journal of Information, Communication and Ethics in Society, Vol. 20, No. 2 (2022): 257-272. DOI. This article is © Emerald Group Publishing Ltd. and permission has been granted for this version to appear in e-Publications@Marquette. Emerald Group Publishing Ltd. does not grant permission for this article to be further copied/distributed or hosted elsewhere without the express permission from Emerald Group Publishing Ltd.

Privacy Concerns with Using Public Data For Suicide Risk Prediction Algorithms: A Public Opinion Survey Of Contextual Appropriateness

Michael Zimmer
Marquette University, Milwaukee, Wisconsin
Sarah Logan
University of Rochester, Rochester, New York

Abstract

Purpose

Existing algorithms for predicting suicide risk rely solely on data from electronic health records, but such models could be improved through the incorporation of publicly available socioeconomic data –

such as financial, legal, life event and sociodemographic data. The purpose of this study is to understand the complex ethical and privacy implications of incorporating sociodemographic data within the health context. This paper presents results from a survey exploring what the general public's knowledge and concerns are about such publicly available data and the appropriateness of using it in suicide risk prediction algorithms.

Design/methodology/approach

A survey was developed to measure public opinion about privacy concerns with using socioeconomic data across different contexts. This paper presented respondents with multiple vignettes that described scenarios situated in medical, private business and social media contexts, and asked participants to rate their level of concern over the context and what factor contributed most to their level of concern. Specific to suicide prediction, this paper presented respondents with various data attributes that could potentially be used in the context of a suicide risk algorithm and asked participants to rate how concerned they would be if each attribute was used for this purpose.

Findings

The authors found considerable concern across the various contexts represented in their vignettes, with greatest concern in vignettes that focused on the use of personal information within the medical context. Specific to the question of incorporating socioeconomic data within suicide risk prediction models, the results of this study show a clear concern from all participants in data attributes related to income, crime and court records, and assets. Data about one's household were also particularly concerns for the respondents, suggesting that even if one might be comfortable with their own being used for risk modeling, data about other household members is more problematic.

Originality/value

Previous studies on the privacy concerns that arise when integrating data pertaining to various contexts of people's lives into algorithmic and related computational models have approached these questions from individual contexts. This study differs in that it captured the variation in privacy concerns across multiple contexts. Also, this study specifically assessed the ethical concerns related to a suicide prediction model and determining people's awareness of the publicness of select data attributes, as well as which of these data attributes generated the most concern in such a context. To the best of the authors' knowledge, this is the first study to pursue this question.

Introduction

Suicide is currently the 10th leading cause of death in the USA, and suicide rates have continued to increase in recent years (Suicide Statistics, 2019). About half of suicide decedents have contact with the health-care system in the month before their death (Ribeiro *et al.*, 2017), indicating that there is a significant opportunity to identify patients who are at risk for a suicide attempt when they visit their health-care provider. Most algorithms for predicting suicide risk rely solely on data from electronic health records (Barak-Corren *et al.*, 2017), but it has been proposed that algorithmic solutions for predicting suicide risk could be improved through the incorporation of publicly available socioeconomic data – such as financial, legal, life event, derogatory and sociodemographic data – alongside electronic health record data (Barak-Corren *et al.*, 2017; O'Connor and Portzky, 2018; Turecki and Brent, 2016).

If incorporation of publicly available socioeconomic data into the suicide prediction model – and other predictive models – is to be pursued, we must understand the complex ethical and privacy implications so as to not erode the trust between a patient and the health-care system. To start this effort, we present results from a survey exploring what the general public's knowledge and concerns are about such publicly available data.

Background

In an age of rapidly increasing technological advancements, scholars from all fields of expertise are investigating how big data can aid them in their work. From bioinformaticians who analyze genetic code to political scientists who analyze government records to sociologists who analyze social media data, everyone wants in on the benefits of big data and the algorithmic systems it empowers (Boyd and Crawford, 2012). Although the potential uses for big data and algorithm-based systems are nearly endless, it is important and necessary to consider the ethical and privacy concerns related to these endless possibilities (Metcalf *et al.*, 2016).

One critical question focuses on the ethical and privacy complications of using publicly available data in research. Significant gaps among researchers on what constitutes "public" data that does not require explicit consent prior to harvesting (Zimmer, 2010, 2016), whether a platform's terms of service might allow the automated scraping of public data (Fiesler *et al.*, 2020), and even at what stage does computational research become human subjects research requiring particular ethical protection (Metcalf and Crawford, 2016). Further, users are often not aware of the types of access researchers have to public data via social media platforms and their application programming interfaces (Fiesler and Proferes, 2018). Given the above, uncertainty persists among the research community on how to address the ethical and privacy complications of using ostensibly public data in computational research projects (Shilton, 2015; Vitak *et al.*, 2016).

These challenges become even more complicated within the context of medical data, where users' knowledge and attitudes about the public availability of their health data is particularly muddled. While some studies have shown users expressing little concern over privacy issues related to their personal fitness information, noting such data was not inherently sensitive and expressing ambivalence over the possibility of sharing that data with third parties (Zimmer *et al.*, 2020), other studies have shown users of self-tracking technologies are frequently unaware of the details of external data access to which they agree in the context of clicking "accept" to "terms of use" (Bietz *et al.*, 2016). Further, a recent study assessing privacy concerns related to the use of publicly available health-related tweets data in research found that the acceptability of harvesting such data depended greatly on the nature of the health ailment, who was collecting it and the context of use (Reuter *et al.*, 2019).

The potential of incorporating publicly available socioeconomic data – typically obtained from commercial data brokers – into medical research complicates things even further. The role of socioeconomic factors in various aspects of health care we well studied (Fiscella *et al.*, 2000; Marmot *et al.*, 2008) and their potential for assisting in the particular challenge of suicide prevention are promising (Barak-Corren *et al.*, 2017; O'Connor and Portzky, 2018; Turecki and Brent, 2016). But integrating such a wide range of data points – such as financial, legal, life event, derogatory and sociodemographic data – into different contexts is often met with resistance, irrespective of how

public the information might be (Crain, 2018; Hoofnagle, 2004; Martin and Nissenbaum, 2017; Tene and Polonetsky, 2013).

To help assess the ethics of incorporating public socioeconomic data in the development of algorithms to predict suicide risk, we invoke Nissenbaum's (2010) theory of privacy as contextual integrity (CI). CI rejects the traditional dichotomy of public versus private information, as well as the notion that privacy preferences and decisions in one context universally apply to other contexts. Instead, CI rests on the understanding that our interactions with other people, institutions and technologies occur within particular contexts. Norms of appropriateness govern people's expectations of how personal information should flow within any given context. Therefore, responding to a data ethics question – e.g. should third-party socioeconomic data be integrated with health data? needs to start not with privacy as a static set of principles but with an understanding of norms of appropriateness within the context in which the data is being collected and used, and whether it is deemed appropriate to move information from one context – such as socioeconomic data – and apply it within a new context – such as medical research in suicide risk prediction.

Study objective

We rely on CI to investigate people's privacy concerns about integrating publicly available socioeconomic data within various contexts, including a suicide risk prediction algorithm. We aim to measure public opinion about privacy concerns across medical-, business- and social media-related contexts, as well as specifically investigate the privacy concerns related to a suicide prediction model. To do this, we first assess participants' general privacy concerns and their privacy knowledge. Then we evaluate participants' opinions and concerns toward various contexts, which we will then be able to compare to concerns toward the use of socioeconomic data in the suicide prediction algorithm. Overall, we aim to determine whether it is ethical to combine publicly available socioeconomic data with health data to improve a suicide prediction algorithm through an assessment of individuals' awareness of this very possibility as well as their comfort with the appropriateness of this use of said data within this particular context.

We approach this through the following research questions:

RQ1. To what extent do people understand that their socioeconomic data – such as financial, legal, life event, derogatory and sociodemographic data – is publicly available?
RQ2. In which contexts do individuals find the use of publicly available information most concerning?
RQ3. Regarding suicide risk prediction algorithms specifically, which socioeconomic data points are most problematic for inclusion?

Methods

Survey instrument

A survey was developed to measure public opinion about privacy concerns with using various socioeconomic data points across various contexts. The survey consisted of five sections that allowed us to collect information about demographics of respondents, general privacy concerns, awareness of what types of information are publicly available, privacy concerns associated with specific contexts and concerns over the use of 30 publicly available data attributes in the context of a suicide risk prediction

algorithm. To inquire about personal data use across a broad range of contexts, we presented respondents with ten vignettes that described scenarios situated in medical, private business and social media contexts (see Appendix 1). Following each vignette, we asked participants to rate their level of concern over the context, and we asked what factor contributed most to their level of concern. Specific to suicide prediction, we presented respondents with various data attributes [1] that could potentially be used in the context of a suicide risk algorithm and asked participants to rate how concerned they would be if each attribute was used for this purpose (see Appendix 2). The survey was tested for clarity and consistency, and the research protocol received Institutional Review Board approval.

Data collection

The survey was deployed on Qualtrics from 17 July to 24 July 2020. We contracted Qualtrics to recruit respondents approximately between the ages of 26 and 99 who live in the USA. Respondents with a response time of fewer than 2.35 min, which was one-half of the median soft launch time, were excluded from the data. Respondents who left comments that were unrelated to what we were asking and exhibited a lack of thoughtful consideration were also removed from the data. In total, we had 420 respondents in our data set.

Data analysis

Descriptive statistics were first used to analyze all questions from the survey. All descriptive statistical procedures were done in Microsoft Excel (Version 2006) and R (Version 3.6.1). We then computed a privacy knowledge score by giving a participant 1 point for every question in the privacy knowledge section that they answered with “Publicly available” and ½ point for every question they answered “Could be determined based on other publicly available information.” Their points were summed to obtain their privacy knowledge score. We computed a privacy concern score for each participant by averaging their responses to the nine questions in the general privacy section. A Pearson correlation coefficient was computed in IBM SPSS Statistics for Windows, version 24, to assess the relationship between privacy concern score and privacy knowledge score.

To further assess privacy concern, we created two new data sets from the original, one containing participants with a high privacy concern score (greater than 3) and one containing participants with a low privacy concern score (less than or equal to 3) to assess how each group responded to all of the survey questions. Chi-square tests were conducted to determine whether there was a relationship between all of the demographic factors and general privacy concern rating (high or low) and between concern over each vignette and privacy concern rating. All chi-square analyses were done in SPSS. A *t*-test was performed in R to determine if there was a significant difference between the privacy knowledge score means of the high general privacy concern group and low general privacy concern group. To evaluate the relationship between privacy concern and knowledge over specific data attributes, we asked about data attributes from related categories in both the concern and knowledge section of the survey. For each concern question, we split the data into two groups, some concern (participants who answered extremely, moderately or somewhat concerned) and little-to-no concern (participants who answered slightly or not at all concerned), and analyzed how much knowledge each group had regarding that specific data attribute. We isolated all participants an additional time based on their concern toward the suicide prediction model vignette. Participants who selected extremely,

moderately or somewhat concerned in response to this vignette were placed in a “some concern” group, and participants who selected slightly or not at all concerned were placed in a “little-to-no concern” group. We then analyzed the groups’ responses to their concern over the use of various data attributes in a suicide prediction model.

Results

Demographics

Over half (251/420, 59.8%) of the participants were aged 26–45, while the other 169 participants (40.2%) ranged from 46 to 66 or more years old. The mean age was 44.97 years (SD = 14.92). Exactly half of the participants were female (210/420, 50.0%). Participants could select multiple ethnicities: 327 participants were white, 46 were African American/Black, 17 were Asian/Pacific Islander, 25 were Hispanic/Latinx, six were Middle Eastern and seven were Native American/Indigenous. Most participants were married or in a domestic partnership (260/420, 61.9%), and 100 participants (23.8%) were single and never married. Education levels varied; 330 participants (78.6%) had at least some level of college experience, whereas 88 (21.0%) had either less than a high school degree, a high school degree or equivalent or had attended trade school. Most participants (274/420, 65.2%) were employed, and 72 (17.1%) were retired. There was a broad range of household income; 69.8% of participants (293/420) had an income of less than $100,000, whereas 28.1% of participants (118/420) had an income of greater than $100,000. Household sizes tended to be small (286/420, 68.1%), although 129 participants (30.7%) lived with 3–5 members and five participants (1.2%) lived with six or more members.

Knowledge of publicly available information

We assessed the knowledge of participants regarding the publicness of various socioeconomic data attributes by testing their awareness of 15 data elements that are publicly available. Summary results are provided in Table 1. Overall, 4,179 of the total 6,300 answer responses (66.3%) to the privacy knowledge questions were correct (participant selected “Could be determined based on other publicly available information” or “Publicly available”), and 2,121 responses (33.7%) were incorrect (participant selected “Not publicly available” or “I don’t know”). The publicness of several data attributes was fairly common knowledge: whether you own or rent at your current address (77.4% correctness), whether or not you are registered to vote (76.4% correctness), the amount of time you lived at your previous address (75.2% correctness), the last recorded sale price of your current address (74.3% correctness) and the amount of time since you last moved (72.4% correctness). The most missed question asked about the publicness of your total number of relatives and associates that own a boat or airplane (46.0% correctness). Questions asking about knowledge of the publicness of derogatory record data all had between 64% and 70% correctness: time since your most recent arrest (64.5% correctness), the total number of misdemeanor convictions (68.3% correctness), the total count of household members with felony convictions (68.6% correctness), whether you have been housed in a correctional facility (69.0% correctness) and total bankruptcy filings (69.3% correctness). Questions with approximately half of the correct responses were: amount of time since your last car accident (53.3% correctness), the total number of relatives and associates who have attended college (56.9% correctness), number of members in your household with licenses for concealed weapons (59.8% correctness) and your estimated household income range (63.6% correctness).

Vignette privacy concerns

Respondents were presented with ten vignettes to gauge their privacy concerns across different types of information gathered across different contents. Overview results are provided in Table 2. The vignette that generated the most concern was about a public health worker collecting global positioning system (GPS) data to assess adherence to stay-at-home orders during the COVID-19 pandemic (319/420, 76% answered that they were somewhat, moderately or extremely concerned about this vignette). The vignette that generated the least concern was about a restaurant owner conducting surveys to improve the restaurant's quality of service (184/420, 43.8% answered that they were not at all concerned about this vignette).

The vignette discussing the use of socioeconomic data in a suicide prediction algorithm showed considerable concern, with two-thirds of participants answering that they were somewhat, moderately or extremely concerned about this use of data, whereas 32.1% answered that they were not at all concerned or slightly concerned.

Factors contributing to concern

Overall, the factors that contributed most to concern over the vignettes were the *purpose of collecting the data* (784/3780 total responses to factor questions, 20.7%) and the *potential future use of data* (737/3780 total responses to factor questions, 19.5%). The *purpose of data collection* was the greatest contributing factor for the vignettes about fitness data being used to develop a weight loss product, purchase transactions being used to stock products, search history being used for targeted advertising, Twitter "Following" lists being used to identify accounts of people getting information about the Black Lives Matter movement and email tracking being used to understand a company's target audience. *Future use* was the greatest contributing factor for the vignettes about tracking COVID-19 stay-at-home orders and the use of genomic data to identify a cancer-causing mutation.

For the vignette discussing the use of socioeconomic data in a suicide prediction algorithm, a technical error prevented users from seeing the options of *potential future use of the data* and *none*. Based on the choices available to respondents for this vignette, the *type of data being collected* generated the most concern (139/420, 33.1%) for this vignette.

Concern over data attributes in suicide prediction model

Specific to the development of a suicide prediction algorithm, we presented respondents with various data attributes that would be used for that purpose (see Appendix 2). These results are summarized in Table 3. Attributes with the greatest expressed concern (indicating some, moderate or extreme concern) include those about annual income (78.6%), ownership of assets (71.2%) or value of real estate (71.4%), court appearances (69.0%), arrest records (68.3%) and felony records (67.1%) and whether one holds a license for concealed weapons (64.8%). Concerns were also evident in data attributes about one's entire household, with many exceeding the concern expressed for the individual data attribute.

Attributes with the least amount of expressed concern include possessing a hunting or fishing license (50.0%), whether one attended college (51.7%) or the number of times in a car accident (54.5%).

General concerns translate to specific concerns

Based on participants' responses to the general privacy questions, we computed a privacy concern score and assigned participants to a high or low general privacy concern group. In total, 215 participants (51.2%) were assigned to the high general privacy concern group and 205 (48.8%) were assigned to the low general privacy concern group. The mean privacy concern score for all participants was 3.06 out of 5 (SD = 0.60). The mean privacy concern score for the high concern group was 3.53 (SD = 0.33), and the mean privacy concern score for the low concern group was 2.56 (SD = 0.38).

Overall, and as expected, those with low general privacy concerns tended to have lower concerns with the vignettes, while those with greater general privacy concerns found the vignettes more concerning. For the vignette discussing the use of socioeconomic data in a suicide prediction algorithm showed considerable concern, 80.0% of respondents with high overall privacy concerns found this particular vignette concerning. And of the respondents with low overall privacy concerns, 55.1% found this vignette concerning.

We then isolated the responses of participants in the high privacy concern group and the low privacy concern group to evaluate how each group responded to the appropriateness of specific socioeconomic data attributes being used in a suicide prediction model. The overall privacy concern groupings tended to be indicative of participants' responses to the specific data attributes. Of the 6,450 total responses to questions regarding concern over the use of various data attributes from the high privacy concern group, there were 3,865 moderately or extremely concerned responses (59.9%). Of the 6,150 total responses to questions regarding concern over the use of various data attributes from the low privacy concern group, there were only 1,767 moderately or extremely concerned responses (28.7%) and 1,842 not at all concerned responses (30.0%).

Among the high overall privacy concern group, considerable concern was expressed for all data attributes, with the lowest-rated factor being "Whether you have a hunting or fishing license" with only 64.2% expressing some, moderate or extreme concern. For those with low overall privacy concerns, some attributes still presented considerable concern: 70.2% of this group expressed some, moderate or extreme concern about "Your estimated annual income," and 70.7% expressed similar concern over "The total estimated annual income for your entire household." Data attributes referencing household members also tended to rate higher than other attributes for this group.

Discussion

Overall findings

Although the benefits of big data are manifold, it is necessary to consider the ethical questions and privacy concerns that arise when integrating data pertaining to various contexts of people's lives into algorithmic and related computational models. Previous studies have approached these questions from a variety of contexts, including personal fitness data (Bietz *et al.*, 2016; Zimmer *et al.*, 2020) and social media monitoring (Reuter *et al.*, 2019). This study differs in that we aimed to capture the variation in privacy concerns across several contexts, spanning from medicine to business to social media. We were also specifically interested in assessing the ethical concerns related to a suicide prediction model and determining people's awareness of the publicness of select data attributes, as well as which of these data attributes generated the most concern in such a context.

Addressing *RQ1*, we found that overall, two-thirds of our respondents correctly determined that the data elements in the survey were publicly available (either directly available or through some sort of inference). While this can be viewed positively, the fact remains that one-third did not have complete awareness of the extent of the publicness of various socioeconomic data points.

Addressing *RQ2*, we found considerable concern across the various contexts represented in our vignettes. With the exception of vignette 7 (a restaurant using customer satisfaction surveys to improve quality), the majority of respondents expressed some level of concern about the data use proposed within the hypothetical vignettes. The highest levels of concern centered on GPS tracking for social distancing compliance, and marketers monitoring email and search engine activities. General concerns over the collection and use of personal data during the COVID-19 pandemic might be a contributing factor to concerns over vignette 1. Overall, concern was greatest (68.0% expressing some, moderate or extreme concern) in vignettes that focused on the use of personal information within the medical context (vignettes 1, 2 and 10).

Our results also show that the most common factors contributing concerns across the various vignettes were the *purpose of data collection* and the *potential future use of data*, a finding supported by existing research showing consumers are most concerned about how companies are and might be using their personal information in the future (Hoffman *et al.*, 1999; Phelps *et al.*, 2018).

By splitting the data into a high privacy concern group and a low privacy concern group, we were able to identify trends within and across these groups. While we expected people with generally high levels of overall privacy concerns to, therefore, express concerns with our vignettes, we were more curious as to whether individuals who typically have low privacy concerns might suddenly express concern for a particular scenario. As with the general findings, even those with low privacy concerns expressed considerable worry about GPS tracking during the COVID-19 pandemic, as well as having researchers monitor Black Lives Matter activity on Twitter. Here, our low privacy concern respondents expressed similar worries from the high concern group about how such data might be used for other purposes.

Data concerns with suicide risk prediction modeling

RQ3 reflects on our specific interest in measuring individuals' comfort with incorporating socioeconomic data within suicide risk prediction models. While this particular vignette ranked in the middle of overall concern, various data elements stood out as particularly problematic among our respondents. Our results show a clear concern from all participants in data attributes related to income, crime and court records, and assets. This is consistent with other research, indicating that most consumers were unwilling to share information about household income and other financial information. Data about one's household – beyond just the individual – were also particularly concerns for our respondents, suggesting that even if one might be comfortable with their own being used for risk modeling, data about other household members is more problematic. This held true even for respondents with generally low privacy concerns, suggesting these data elements are particularly troublesome when used within this context.

Connected to *RQ1*, a concerning finding is that many attributes that a majority of respondents failed to recognize were publicly available were also flagged as particularly concerning in the detailed assessment of data used within suicide risk prediction algorithms. For example, 54.0% of respondents

did not recognize that the “count of relatives and associates that own a boat or airplane” was publicly available, yet 75.5% found it somewhat, moderately or extremely concerning that the data element “Whether or not anyone in your household owns assets (such as a watercraft, an aircraft or real estate property)” might be used in a suicide risk prediction algorithm. Similarly, over 40% of respondents did not realize “number of members in your household with licenses for concealed weapons” was publicly available, while 66.4% found using such data concerning. This suggests many respondents have concerns over the use of certain data elements while underestimating the general availability of the data.

Study limitations

We recognize that participants recruited through Qualtrics are likely digitally savvy individuals and of a high enough socioeconomic status to own a device on which to take the survey. We acknowledge that these characteristics likely had some impact on our results. To help mitigate the effects of these characteristics, we requested that Qualtrics provide us with a specific distribution of individuals across age, income and gender. Nonetheless, these characteristics undoubtedly had an influence on how participants responded, in particular, to the vignettes.

We also recognize that had we framed the suicide prediction model vignette in slightly different terms, it could have elicited a different response from participants. For example, had we put a greater emphasis on the benefits and societal good of creating such an algorithm and had we clarified that all personal information would be de-identified, perhaps participants would have been less concerned over the use of data in this way. Future work could focus more specifically on participants’ concerns over a suicide risk prediction algorithm and include vignettes all with the main purpose of creating a suicide prediction model but altering more minor factors about the vignettes, such as the type of data used, how it was obtained and whether the algorithm would be used by someone other than clinicians.

Conclusion

In this study, we measured public opinion regarding the use of data in various contexts. In particular, we were interested in assessing opinion over the use of publicly available socioeconomic data in a suicide risk prediction algorithm. To aid in our analysis of these contexts, we also measured public knowledge of select data attributes and concern over the incorporation of these attributes into the suicide prediction model.

Combining socioeconomic data with existing medical records gives researchers the opportunity to improve suicide prediction models. It is clear that the overall goal of this initiative, minimizing suicide attempts, is good and beneficial to society. However, informed by the lens of CI, the incorporation of socioeconomic data within suicide risk prediction models threatens to violate existing norms of what information is appropriate within the medical context. We found that over two-thirds of participants have at least some concern level toward using socioeconomic data in the suicide prediction algorithm. In comparison to the response to the nine other vignettes, this suicide prediction model vignette fell approximately in the middle in terms of the level of concern. This indicates that while this case is less concerning than some popular uses of data today, such as tracking of search history or email tracking, it is undoubtedly more concerning than researchers accessing genomic data from an ancestry website or fitness data from a wearable device.

We also found that the publicness of some data attributes was well known, such as voter registration records and address records, whereas the publicness of other types of information was less well known, such as asset records of relatives and accident records. We highlighted certain data attributes that were particularly sensitive to individuals who exhibited both high and low privacy concern, such as data related to income, assets and criminal records. Taken together, medical patients may have a lack of awareness that their doctors have access to their socioeconomic data and data about their household members, which has been aggregated by a third party, and some of those data elements are particularly problematic.

Ultimately, we were able to determine that the appropriateness of incorporating personal data within various computational applications is contextually dependent, with the appropriateness often determined by the type of use and concern over future uses of data. We found that the use of certain data attributes is more concerning than others, and that individuals often lack full knowledge of the availability of public data, especially certain sensitive socioeconomic data attributes about our lives and our broader households. Specifically, we determined that participants were most concerned about the use of income records, asset data and criminal records in suicide risk prediction models, with asset data also being among the data elements participants were least aware were publicly available. Therefore, researchers hoping to rely on such data need to take steps to fully consider the broader ethical and privacy implications of relying on such data, despite their possible predictive value.

In the broadest sense, we have shown how confronting the ethical and privacy implications of incorporating publicly available socioeconomic data into algorithmic models presents a unique challenge that requires more than simply relying on the public availability of such data. Researchers – and the general public – are better off when we rely on robust conceptual frameworks such as CI and engage in social science-based research to better understand the knowledge and expectations of the general public. Algorithmic models like those to help predict suicide risk can be of great public benefit, but only if pursued in an ethically informed manner.

Note

1. Data attributes were based on a list of over 400 “Socioeconomic Health Attributes” marketed by LexisNexis to improve predictive modeling: https://risk.lexisnexis.com/products/socioeconomic-health-attributes

This material is based upon work supported by the National Science Foundation REU site grant no. IIS-1950826 “Data Science Across the Disciplines.” The authors also thank Dr Jordan Smoller and his colleagues at Harvard Medical School and in the Psychiatric and Neurodevelopmental Genetics Unit (PNGU) at Massachusetts General Hospital for their feedback and support.

References

Suicide Statistics (2019), “American foundation for suicide prevention”, 15 November, available at: https://afsp.org/suicide-statistics/ (accessed 4 October 2020).

Barak-Corren, Y., Castro, V.M., Javitt, S., Hoffnagle, A.G., Dai, Y., Perlis, R.H., Nock, M.K., Smoller, J.W. and Reis, B.Y. (2017), “Predicting suicidal behavior from longitudinal electronic health records”, *American Journal of Psychiatry*, Vol. 174 No. 2, pp. 154-162.

Bietz, M.J., Bloss, C.S., Calvert, S., Godino, J.G., Gregory, J., Claffey, M.P., Sheehan, J. and Patrick, K. (2016), "Opportunities and challenges in the use of personal health data for health research", *Journal of the American Medical Informatics Association*, Vol. 23 No. e1, pp. e42-e48.

Boyd, D. and Crawford, K. (2012), "Critical questions for big data", *Information, Communication and Society*, Vol. 15 No. 5, pp. 662-679.

Crain, M. (2018), "The limits of transparency: data brokers and commodification", *New Media and Society*, SAGE Publications, Vol. 20 No. 1, pp. 88-104.

Fiesler, C. and Proferes, N. (2018), "'Participant' perceptions of Twitter research ethics", *Social Media Society*, Vol. 4 No. 1, p. 2056305118763366.

Fiesler, C., Beard, N. and Keegan, B.C. (2020), "No robots, spiders, or scrapers: legal and ethical regulation of data collection methods in social media terms of service", *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14, pp. 187-196.

Fiscella, K., Franks, P., Gold, M.R. and Clancy, C.M. (2000), "Inequality in quality: addressing socioeconomic, racial, and ethnic disparities in health care", *JAMA*, Vol. 283 No. 19, p. 2579.

Hoffman, D.L., Novak, T.P. and Peralta, M. (1999), "Building consumer trust online", *Communications of the ACM*, Vol. 42 No. 4, pp. 80-85.

Hoofnagle, C. (2004), "Big brother's little helpers: how choice point and other commercial data brokers collect and package your data for law enforcement", *North Carolina Journal of International Law*, Vol. 29 No. 4, p. 595.

Marmot, M., Friel, S., Bell, R., Houweling, T.A. and Taylor, S. (2008), "Closing the gap in a generation: health equity through action on the social determinants of health", *The Lancet*, Vol. 372 No. 9650, pp. 1661-1669.

Martin, K. and Nissenbaum, H. (2017), "Privacy interests in public records: an empirical investigation", *Harvard Journal of Law and Technology*, Vol. 31, p. 111.

Metcalf, J. and Crawford, K. (2016), "Where are human subjects in big data research? The emerging ethics divide", *Big Data and Society*, Vol. 3 No. 1, p. 2053951716650211.

Metcalf, J., Keller, E.F. and Boyd, D. (2016), "Perspectives on big data, ethics, and society", Council for Big Data, Ethics, and Society, available at: https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/ (accessed 19 April 2019).

Nissenbaum, H. (2010), *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, Stanford Law Books, Stanford, California.

O'Connor, R.C. and Portzky, G. (2018), "Looking to the future: a synthesis of new developments and challenges in suicide research and prevention", *Frontiers in Psychology*, Vol. 9, p. 2139, doi: 10.3389/fpsyg.2018.02139.

Phelps, J., Nowak, G. and Ferrell, E. (2018), "Privacy concerns and consumer willingness to provide personal information", *Journal of Public Policy and Marketing*, SAGE Publications, Los Angeles, CA, Vol. 19 No. 1, doi: 10.1509/jppm.19.1.27.16941.

Reuter, K., Zhu, Y., Angyan, P., Le, N., Merchant, A.A. and Zimmer, M. (2019), "Public concern about monitoring Twitter users and their conversations to recruit for clinical trials: survey study", *Journal of Medical Internet Research*, Vol. 21 No. 10, p. e15455.

Ribeiro, J.D., Gutierrez, P.M., Joiner, T.E., Kessler, R.C., Petukhova, M.V., Sampson, N.A., Stein, M.B., Ursano, R.J. and Nock, M.K. (2017), "Health care contact and suicide risk documentation prior

to suicide death: results from the army study to assess risk and resilience in service members", *Journal of Consulting and Clinical Psychology*, Vol. 85 No. 4, pp. 403-408.

Shilton, K. (2015), "Emerging ethics norms in social media research", presented at the Workshop on Beyond IRBs: Ethical Review Processes for Big Data Research, available at: https://bigdata.fpf.org/papers/emerging-ethics-norms-in-social-media-research/ (accessed 28 December 2016).

Tene, O. and Polonetsky, J. (2013), "Big data for all: privacy and user control in the age of analytics", *Northwestern Journal of Technology and Intellectual Property*, Vol. 11 No. 5, pp. 239-273.

Turecki, G. and Brent, D.A. (2016), "Suicide and suicidal behaviour", *The Lancet*, Vol. 387 No. 10024, pp. 1227-1239.

Vitak, J., Shilton, K. and Ashktorab, Z. (2016), "Beyond the Belmont principles: ethical challenges, practices, and beliefs in the online data research community", *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, ACM, *New York, NY*, pp. 941-953.

Zimmer, M. (2010), "'But the data is already public': on the ethics of research in Facebook", *Ethics and Information Technology*, Vol. 12 No. 4, pp. 313-325.

Zimmer, M. (2016), "OkCupid study reveals the perils of big-data science", *Wired*, 14 May, available at: www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/ (accessed 28 May 2016).

Zimmer, M., Kumar, P., Vitak, J., Liao, Y. and Chamberlain Kritikos, K. (2020), "'There's nothing really they can do with this information': unpacking how users manage privacy boundaries for personal fitness information", *Information, Communication and Society*, Vol. 23 No. 7, pp. 1020-1037.

Table 1. Knowledge of publicly available information

Data attribute	I don't know	Not publicly available	Could be determined based on other information	Publicly available
1. The amount of time since your last car accident	8319.8%	11326.9%	13331.7%	9121.7%
2. Whether you own or rent at your current address	4210.0%	5312.6%	15737.4%	16840.0%
3. Total count of your relatives and associates that own a boat or airplane	11026.2%	11727.9%	12529.8%	6816.2%
4. Last recorded sale price of your current address	4310.2%	6515.5%	13031.0%	18243.3%
5. Time since your most recent arrest	6014.3%	8921.2%	10124.0%	17040.5%
6. Your total number of misdemeanor convictions	6014.3%	7317.4%	12630.0%	16138.3%
7. Total count of your household members with felony convictions	7317.4%	5614.0%	13832.9%	15035.7%
8. Whether or not you've been housed in a correctional facility	6916.4%	6114.5%	13532.1%	15536.9%
9. Your total number of bankruptcy filings	5613.3%	7317.4%	14334.0%	14835.2%
10. Total number of relatives and associates who have attended college	8319.8%	9823.3%	13131.2%	10825.7%
11. Your estimated household income range	4911.7%	10424.8%	17742.1%	9021.4%
12. The number of members in your household with licenses for concealed weapons	7818.6%	9121.7%	14334.0%	10825.7%
13. Amount of time since you last moved	409.5%	7618.1%	17140.7%	13331.7%
14. Whether or not you are registered to vote	4410.5%	5513.1%	14133.6%	18042.9%
15. The amount of time you lived at your previous address	389.0%	6615.7%	14935.5%	16739.8%

Table 2. Vignette privacy concerns

Vignette	Not at all concerned	Slightly concerned	Somewhat concerned	Moderately concerned	Extremely concerned
1. GPS data for tracking stay-at-home order adherence	378.8%	6415.2%	7217.1%	8119.3%	16639.5%

2. Genomic data for identification of cancer-causing mutation	7417.6%	9322.1%	8019.0%	7217.1%	10124.0%
3. Fitness data for a weight loss product	7217.1%	7618.1%	9021.4%	7417.6%	10825.7%
4. Transaction records for stocking purposes	5713.6%	9121.7%	8019.0%	8119.3%	11126.4%
5. Search history for targeted advertising	4410.5%	7618.1%	7517.9%	9723.1%	12830.5%
6. Following lists for tracking users involved in Black Lives Matter movement	6716.0%	6014.3%	7718.3%	8420.0%	13231.4%
7. Customer satisfaction surveys to improve restaurant quality	18443.8%	5412.9%	6214.8%	5713.6%	6315.0%
8. Email tracking for understanding target audience	337.9%	8219.5%	9221.9%	7818.6%	13532.1%
9. Comments to improve a social media ad	11026.2%	7818.6%	7718.3%	7116.9%	8420.0%
10. Socioeconomic data for a suicide prediction model	6214.8%	7317.4%	9422.4%	9522.6%	9622.9%

Table 3. Concern over data attributes in suicide prediction model

Data attribute	**Not at all concerned**	**Slightly concerned**	**Somewhat concerned**	**Moderately concerned**	**Extremely concerned**
Number of times you have been in a car accident	9723.1%	9422.4%	7818.6%	7417.6%	7718.3%
Distance (in miles) between you and your nearest relative	7818.6%	6214.8%	10124.0%	8821.0%	9121.7%
Whether or not you are registered to vote	10825.7%	5412.9%	8119.3%	9121.7%	8620.5%
The number of phone numbers associated with you	5513.1%	7016.7%	8420.0%	8921.2%	12229.0%
Whether or not you attended college	12429.5%	7918.8%	7918.8%	7016.7%	6816.2%
Total number of household members who attended college	11727.9%	6515.5%	7517.9%	8921.2%	7417.6%
Your estimated annual income	378.8%	5312.6%	9322.1%	10926.0%	12830.5%
The total estimated annual income for your entire household	358.3%	5513.1%	10224.3%	9422.4%	13431.9%

The original mortgage dollar amount at your current address	7417.6%	7818.6%	8720.7%	8620.5%	9522.6%
The estimated market value of your previous address	9823.3%	4911.7%	9322.1%	10124.0%	7918.8%
The difference in neighborhood median household income between your address and your most recent address	8119.3%	7818.6%	8420.0%	9723.1%	8019.0%
The number of multi-family properties in your neighborhood	12830.5%	5011.9%	8821.0%	8720.7%	6716.0%
Your current address' neighborhood crime index, based on law enforcement data	11026.2%	6315.0%	8821.0%	8520.2%	7417.6%
Your previous address' neighborhood crime index, based on law enforcement data	11226.7%	5212.4%	8219.5%	9923.6%	7517.9%
Whether or not you own assets (such as a watercraft, aircraft or real estate property)	4510.7%	7618.1%	10424.8%	9823.3%	9723.1%
Whether or not anyone in your household owns assets (such as a watercraft, an aircraft or real estate property)	419.8%	6214.8%	11026.2%	10725.5%	10023.8%
The total value for all real estate properties you own	5513.1%	6515.5%	8620.5%	10324.5%	11126.4%
The total value for all real estate properties everyone in your household owns	4911.7%	5914.0%	9221.9%	10324.5%	11727.9%
Total number of real estate properties sold within past five years	8821.0%	4510.7%	8720.7%	9021.4%	11026.2%
The total number of court records (including felony, misdemeanor, lien, judgment, bankruptcy or eviction) listed in your name	8119.3%	4911.7%	9121.7%	8319.8%	11627.6%
The total number of court records (including felony, misdemeanor, lien, judgment, bankruptcy or eviction) for your entire household	7517.9%	5412.9%	9522.6%	8520.2%	11126.4%
Your total number of arrests	9021.4%	4310.2%	7217.1%	10224.3%	11326.9%
Total number of arrests for your entire household	8520.2%	5011.9%	7317.4%	9522.6%	11727.9%
Your total number of felony convictions	8821.0%	5011.9%	7317.4%	9422.4%	11527.4%

Total number of felony convictions for your entire household	9221.9%	5212.4%	7116.9%	9923.6%	10625.2%
Whether you have a hunting or fishing license	14033.3%	7016.7%	6515.5%	7918.8%	6615.7%
Whether anyone in your household has a hunting or fishing license	13732.6%	5112.1%	8319.8%	7918.8%	7016.7%
Whether you have a license for concealed weapons	8821.0%	6014.3%	7818.6%	9121.7%	10324.5%
Whether anyone in your household has a license for concealed weapons	8720.7%	5412.9%	9522.6%	9522.6%	8921.2%
The number of times you have changed addresses in the past five years	8921.2%	6515.5%	8720.7%	8019.0%	9923.6%

Appendix 1. Vignettes

Survey respondents were presented with ten vignettes that described scenarios situated in medical, private business and social media contexts and were asked to provide their level of concern for their data being used in this way:

(1) A public health worker is collecting location data from your phone's GPS system. The data will be used to track your adherence to stay-at-home orders during the COVID-19 pandemic.

(2) A medical researcher is collecting genomic data from an ancestry website that collects saliva samples. The data will be used to confirm the identity of a cancer-causing mutation.

(3) A research team for a fitness company is collecting data about your activity levels from your wearable fitness tracker. The data will be used to develop and market a new weight loss product.

(4) A researcher for a private company is collecting your purchase transaction records from their stores. The data will be used to analyze customer purchase habits and then stock items accordingly.

(5) A marketer is collecting your search history from a popular search engine. The data will be used to advertise their products to you based on your interests.

(6) A university researcher is collecting the "Following" list from Twitter accounts that have used the hashtag #BlackLivesMatter. The data will be used to identify the accounts of people who are getting information about the Black Lives Matter movement.

(7) A manager at a local restaurant is collecting your responses to a customer satisfaction survey. The data will be used to improve the quality of service.

(8) A marketer is collecting information about when and where you opened an email from an email tracking service. The data will be used to gain a better understanding of their target audience.

(9) A marketer is collecting comments from one of their social media advertisements. The data will be used to assess the reaction to the ad and improve the ad accordingly.

(10) A medical researcher wants to collect socioeconomic data from public databases and combine it with medical records of people in your community to improve an algorithm that identifies patients with suicide risk.

Respondents were also asked what factor contributed most to their level of concern:

- Who is obtaining your data.
- The type of data being collected.
- Where the data is being obtained from.
- The purpose of collecting the data.
- The potential future use of the data.
- None of these.

Appendix 2. Data attributes for suicide risk prediction

Survey respondents were presented various data attributes that would be used in the context of a suicide risk algorithm and asked participants to rate how concerned they would be if each attribute was used for that purpose:

- Number of times you have been in a car accident.
- Distance (in miles) between you and your nearest relative.
- Whether or not you are registered to vote.

- The number of phone numbers associated with you.
- Whether or not you attended college.
- Total number of household members who attended college.
- Your estimated annual income.
- The total estimated annual income for your entire household.
- The original mortgage dollar amount at your current address.
- The estimated market value of your previous address.
- The difference in neighborhood median household income between your address and your most recent address.
- The number of multi-family properties in your neighborhood.
- Your current address' neighborhood crime index, based on law enforcement data.
- Your previous address' neighborhood crime index, based on law enforcement data.
- Whether or not you own assets (such as a watercraft, an aircraft or real estate property).
- Whether or not anyone in your household owns assets (such as a watercraft, an aircraft or real estate property)
- The total value for all real estate properties you own.
- The total value for all real estate properties everyone in your household owns.
- Total number of real estate properties sold within past five years.
- The total number of court records (including felony, misdemeanor, lien, judgment, bankruptcy or eviction) listed in your name.
- The total number of court records (including felony, misdemeanor, lien, judgment, bankruptcy or eviction) for your entire household.
- Your total number of arrests.
- Total number of arrests for your entire household.
- Your total number of felony convictions.
- Total number of felony convictions for your entire household.
- Whether you have a hunting or fishing license.
- Whether anyone in your household has a hunting or fishing license.
- Whether you have a license for concealed weapons.
- Whether anyone in your household has a license for concealed weapons.
- The number of times you have changed addresses in the past five years.