Internet search data and mental health: promises and pitfalls

ALEXANDRE LOCH, Miranda Wolpert, Lynsey Bilsland, Elena Netsi, Matthew Brown, Gwydion Williams, Shuranjeet Takhar, Christopher Christofi, Roman Kotov

Submitted to: JMIR Mental Health
on: May 20, 2024

Table of Contents

Internet search data and mental health: promises and pitfalls

ALEXANDRE LOCH[1] PhD, MD; Miranda Wolpert[2]; Lynsey Bilsland[2]; Elena Netsi[2]; Matthew Brown[2]; Gwydion Williams[2]; Shuranjeet Takhar[2]; Christopher Christofi[2]; Roman Kotov[3]

[1]Institute of Psychiatry, University of Sao Paulo São Paulo BR
[2]Wellcome Trust London GB
[3]Renaissance School of Medicine Stony Brook University New York US

Corresponding Author:
ALEXANDRE LOCH PhD, MD
Institute of Psychiatry, University of Sao Paulo
Rua Dr. Ovidio Pires de Campos 785
4 andar sala 4N60
São Paulo
BR

Abstract

Internet is now integral to everyday life, and users' online search data could be of strategic importance in mental healthcare. As shown by previous studies, it may provide valuable insights into individuals mental state, and can be of great value in early identification and in helping pathways to care. Internet search data can potentially provide real-time identification and alert mechanisms to timely interventions, for instance. In this viewpoint paper, we discuss the various problems related to the use of this data in research and in clinical practice such as privacy, integration with clinical information, and other technical challenges. We propose solutions to address these issues, and possible directions to follow.

(JMIR Preprints 20/05/2024:60754)
DOI: https://doi.org/10.2196/preprints.60754

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Internet search data and mental health: promises and pitfalls

Alexandre Andrade Loch[1]; Miranda Wolpert [2]; Lynsey G. Bilsland[2]; Elena Netsi[2]; Matthew Brown[2]; Gwydion Williams[2]; Shuranjeet Takhar[2]; Christopher Christofi[2]; Roman Kotov[3]

[1] Laboratorio de Neurociencias (LIM 27), Instituto de Psiquiatria, Hospital das Clinicas HCFMUSP, Faculdade de Medicina, Universidade de Sao Paulo, Sao Paulo, SP, BR, Brazil

[2] Wellcome Trust, London, NW1 2BE, UK

[3] Renaissance School of Medicine, Stony Brook University, NY, USA

Abstract

Internet is now integral to everyday life, and users' online search data could be of strategic importance in mental healthcare. As shown by previous studies, it may provide valuable insights into individuals mental state, and can be of great value in early identification and in helping pathways to care. Internet search data can potentially provide real-time identification and alert mechanisms to timely interventions, for instance. In this viewpoint paper, we discuss the various problems related to the use of this data in research and in clinical practice such as privacy, integration with clinical information, and other technical challenges. We propose solutions to address these issues, and possible directions to follow.

Keywords:

privacy; stigma; online; prevention;

Introduction

Internet is now integral to everyday life, and the world's population has never been more interconnected. It is an important source of communication and mainly of information, fuelled by powerful search engines linking users do websites. As such, internet search data could be of strategic importance in mental healthcare. The World Health Organization (WHO) adopted in 2020 a global strategy, presenting a roadmap to link the latest developments in innovation and digital health to improve health outcomes. This has been proven to be particularly important for mental illnesses, as evidence suggests an increase in its prevalence rates after the pandemic[1], with digital technologies demonstrating a great potential to aid this area[2]. In this sense, internet search data could potentially be of great value in early identification and in helping pathways to care, reducing the everlasting delays in seeking help and improving outcomes.

Value proposition

Internet search data may provide valuable insights into users' mental state. A systematic review conducted in 2021 observed a significant increase in terms related to mental illnesses during the COVID-19 pandemic[3]. This link was confirmed by associating the number of COVID-19 cases or deaths with internet search volume for specific keywords, as well as by content and sentiment analysis. Search data is not limited to internet search terms used, but may include their timestamp, location history, and YouTube search data and videos viewed—available through the Google Takeout app, for instance[4]. As such, it provides a diverse and rich data source for mental health research, offering potential insights into users' behavior, thoughts, and concerns[5]. Content of these searches can be linked to clinical data obtained from electronic health records or direct patient assessment to identify clinically relevant language. Studies may focus on key words (e.g., searching for "voices" or "psychiatrist"), categories of words (e.g., negative emotions lexicon), composites of words and phrases derived by machine learning, or contextual embeddings identified by deep learning models[6]

Recent studies suggest that signals from search history and YouTube data may be correlated with depression and anxiety levels,[4] predict suicidal behavior,[7] and predict diagnoses of schizophrenia spectrum disorders as well[8]. Exemplifying, one such study observed that individuals with mood disorders conducted internet searches significantly more than subjects with schizophrenia spectrum disorders from 6pm to 12am, and used significantly more words related to negative emotions compared to healthy volunteers[8]. Accordingly, internet search data might offer real-time alert and opportunities for timely and early interventions, potentially improving mental health outcomes. Most important, it might be an useful tool to address the delay in seeking help for mental disorders[9]. Researchers are also investigating whether it could potentially be used as an adjunct to ongoing clinical management, with a small number of

pilot studies underway. Internet search studies show potential, but existing data are limited due to several concerns related to the conduction of such studies .

Problem Statement

One of the most important issues concerning the use internet search data for mental healthcare purposes is privacy[10]. Uncertainties about user willingness to share data are especially critical across diverse groups and countries, raising concerns about the feasibility of large-scale data collection and potential biases in the collected data[10]. Also, since the recent privacy violation lawsuits that got much media attention in the United States, a certain atmosphere of distrust has been created in the last few years, a challenge that needs to be acknowledged and overcome[11]. Still in this regard, addressing consent issues for vulnerable populations, such as children and older adults, is complex, requiring careful consideration of ethical guidelines and legal frameworks.

The second problem entails researchers' integration of internet search data with other research data, such as electronic health records. Besides the privacy concerns mentioned, this process poses technical challenges, as does interpreting search patterns, requiring advanced data analysis methods and interdisciplinary collaboration.

Third, further general technical challenges need to be addressed. For instance, user-friendliness with procedures to give permission to access data; potential changes in search behavior due to awareness of monitoring, impacting data quality and reliability; ownership determination of shared resource data, especially in regions where phones are communal; digital and mental health literacy, the perceived risk around stigma, and adherence to digital mental health initiatives.

Fourth, existing studies have been limited by small sample size, on the order of 100 participants or fewer. However, for accurate detection of psychopathology, language models have to be complex and training of such models requires thousands of observations[6]. Moreover, model generalizability must be evaluated in new samples.

At last, as scarce evidence exist for the use of internet search data for mental healthcare purposes, the clinical accuracy and utility of identifying people at risk, diagnosing, and reliably monitoring changes in symptoms remains to be determined. Retrospective data might have limitations in predicting future outcomes effectively (data leakage, model overfitting, e.g.) raising questions about the reliability of the data for certain types of analyses and predictions. Research should acknowledge that clinical utility of information that can be extracted from internet searches is currently unclear.

Proposed Solutions and Implementation

As for privacy and distrust, some actions are needed. First, thorough ethical analysis of research initiatives tied with clear and up-to-date forms of participant consent are warranted[12]. In this sense, it is important to reinforce participants that the data that will be collected is for health research and not for commercial purposes[12]. Modern data security techniques and deidentification of data should be carried out to ensure privacy. Also, to improve trust in data sharing, initiatives should be led by reputable academic institutions, with big tech companies acting as facilitators or in other supporting roles.

Regarding the second issue of integration of data, we propose to conduct research in large hospitals and universities where considerable amount of clinical data will be available to support large sample size. Inclusivity and diversity should be sought, involving academic institutions not only based in North America and Europe—where most of the scientific evidence is generated—but also those in low-and-middle-income countries as well. Enough funding should be destinated to such projects to build a team of experts and to acquire enough hardware and software to deal with large and heterogeneous datasets. Consortia to thrive the exchange of experience and know-how between teams in the academic and commercial fields should be stimulated.

Technical challenges mentioned as the third problem need to be addressed by rigorous study designs. Most importantly, evaluation of the final model in previously unseen samples and rigorous assessment of overfitting given the large quantities of data. User-friendliness of the data sharing procedures needs to be investigated by studies on usability and acceptability. Hawthorne effect needs to be considered in study designs and possible use of ecological (or retrospective) data is warranted. Participant's consent for sharing such sensitive data requires a detailed discussion with researchers, especially about data ownership. At last, as for all mental health initiatives, stigma needs to be fought by increasing mental health literacy and providing accurate information on mental illnesses. This will ultimately stimulate participation in the studies and in theory decrease possible biases.

Table 1 — Summary of problems and proposed solutions for digital technology

in mental health research

Problems	Proposed solutions
I. Privacy concerns	Ethical analysis, participant consent, data securing, research led by Academia
II. Data integration	Enroll large clinical datasets, funding for multidisciplinary team
III. Technical challenges	Sufficient training data (usually thousands of observations) and evaluation in new datasets
IV. Clinical accuracy and utility	More diverse samples, different sociocultural settings, include patient samples and outcomes of clinical interest (e.g., prognosis, response to specific treatment)

The challenge of stigma and privacy from a broader perspective will be partly addressed by the actions narrated in the first paragraph above. Additionally, other actions aiming at regulating the use of internet search data at the clinical level should be proposed. Such legal actions should protect society against inadequate commercial exploitation of the technology, technology misuse, breach of confidentiality, and related issues. Together with campaigns to improve mental health literacy and to decrease distrust in big data collection and use for medical purposes, public actions to support the consortium between the academia and companies such as Google—and government sectors—should be stimulated and publicized. This environment of multiple cooperation between diverse sectors shall encourage population participation and technology adoption. Also, enrollment of lived experience in all processes of research (from design until governance) shall further improve public adoption and acceptability of new technologies.

Finally, diversity of samples is needed to address clinical accuracy. Also, cutting-edge, and ever-evolving data analysis techniques will increasingly improve the performance of technologies. As for clinical utility, models need to be tested in real-world scenarios, to demonstrate their superiority over currently used screening strategies and clinical evaluations.

Conclusion

Internet search data that are embedded in daily life gather large amounts of individual data and have the potential to revolutionize mental healthcare. Summarizing, individual search data might be of great value in: I) the prediction of mental health problems using search history; II) real-time identification and alert mechanism to target timely interventions; and III) as adjunct to clinical management. But to do so, assuring participant's trust and confidentiality is essential. Consortia between academic institutions and tech companies should be stimulated, aiming at increasing population's reliance, knowledge exchange

between different sectors, and big dataset handling. Further research is strongly encouraged to address feasibility and acceptability. Existing studies are too small to ensure replicable results and sample sizes should be an order of magnitude larger. New initiatives should be inclusive and diverse, enrolling the Global South to assure universality of findings.

7. References

1 Bower M, Smout S, Donohoe-Bales A, *et al.* A hidden pandemic? An umbrella review of global evidence on mental health in the time of COVID-19. *Frontiers in Psychiatry* 2023; **14**. https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2023.1107560 (accessed Feb 29, 2024).

2 Torous J, Bucci S, Bell IH, *et al.* The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry* 2021; **20**: 318–35.

3 Gianfredi V, Provenzano S, Santangelo OE. What can internet users' behaviours reveal about the mental health impacts of the COVID-19 pandemic? A systematic review. *Public Health* 2021; **198**: 44–52.

4 Zhang B, Zaman A, Silenzio V, Kautz H, Hoque E. The Relationships of Deteriorating Depression and Anxiety With Longitudinal Behavioral Changes in Google and YouTube Use During COVID-19: Observational Study. *JMIR Mental Health* 2020; **7**: e24012.

5 Insel TR. Digital phenotyping: a global tool for psychiatry. *World Psychiatry* 2018; **17**: 276–7.

6 Eichstaedt JC, Kern ML, Yaden DB, *et al.* Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychol Methods* 2021; **26**: 398–427.

7 Areán PA, Pratap A, Hsin H, *et al.* Perceived Utility and Characterization of Personal Google Search Histories to Detect Data Patterns Proximal to a Suicide Attempt in Individuals Who Previously Attempted Suicide: Pilot Cohort Study. *J Med Internet Res* 2021; **23**: e27918.

8 Birnbaum ML, Kulkarni P 'Param', Meter AV, *et al.* Utilizing Machine Learning on Internet Search Activity to Support the Diagnostic Process and Relapse Detection in Young Individuals With Early Psychosis: Feasibility Study. *JMIR Mental Health* 2020; **7**: e19348.

9 Pretorius C, Chambers D, Coyle D. Young People's Online Help-Seeking and Mental Health Difficulties: Systematic Narrative Review. *Journal of Medical Internet Research* 2019; **21**: e13873.

10 Rendina HJ, Mustanski B. Privacy, Trust, and Data Sharing in Web-Based and Mobile Research: Participant Perspectives in a Large Nationwide Sample of Men Who Have Sex With Men in the United States. *Journal of Medical Internet Research* 2018; **20**: e9019.

11 Cohen IG, Mello MM. Big Data, Big Tech, and Protecting Patient Privacy. *JAMA* 2019; **322**: 1141–2.

12 Mittelstadt BD, Floridi L. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. In: Mittelstadt BD, Floridi L, eds. The Ethics of Biomedical Big Data. Cham: Springer International Publishing, 2016: 445–80.