A Systematic Review of **Ethics Disclosures in Predictive Mental Health** Research

Leah Ajmani, **Stevie Chancellor,** Bijal Mehta, Casey Fiesler, Michael Zimmer, Munmun De Choudhury

People



Leah Ajmani*
Univ of Minnesota



Stevie Chancellor*
Univ of Minnesota



Bijal Mehta
Georgetown University



Casey Fiesler
CU Boulder



Michael Zimmer
Marquette University



Munmun De Choudhury
Georgia Tech

* Both authors contributed equally to the paper



Why 1,320 Therapists Are Worried About Mental Health in America Right Now

By Tara Parker-Pope, Christina Caron and Mónica Cordero Sancho

Dec. 17, 2021

The New York Times, 2021

Why 1,320 Therapists Are Worried About Mental Health in America Right Now

By Tara Parker-Pope, Christina Caron and Mónica Cordero Sancho

Dec. 17, 2021

The New York Times, 2021

Declining Mental Health

The pandemic's toll on our mental health is showing up in our medicine cabinets.



Why 1,320 Therapists Are Worried About Mental Health in America Right Now

By Tara Parker-Pope, Christina Caron and Mónica Cordero Sancho

Dec. 17, 2021

The New York Times, 2021



Declining Mental Health

The pandemic's toll on our mental health is showing up in our medicine cabinets.



'It's Life or Death': The Mental Health Crisis Among U.S. Teens

Depression, self-harm and suicide are rising among American adolescents. For one 13-year-old, the despair was almost too much to take.

Characterizing and Predicting Postpartum Depression from Shared Facebook Data

Munmun De Choudhury Scott Counts Eric J. Horvitz Aaron Hoff
Microsoft Research, Redmond WA 98052
{munmund, counts, horvitz, aaron.hoff}@microsoft.com

ABSTRACT
The birth of a child is a major milestone in the life of parents. We leverage Facebook data shared voluntarily by 165 new mothers as streams of evidence for characterizing their postnatal experiences. We consider multiple measures including activity, social capital, emotion, and linguistic style in participants' Facebook data in pre- and postnatal periods. Our study includes detecting and predicting onset of post-partum depression (PPD). The work complements recent work on detecting and predicting significant postpartum changes in behavior, language, and affect from Twitter data. In contrast to prior studies, we gain access to ground truth on postpartum experiences via self-reports and a common psychometric instrument used to evaluate PPD. We develop a series of statistical models to predict, from data available before childbirth, a mother's likelihood of PPD. We corroborate our quantitative findings through interviews with mothers experiencing PPD. We find that increased social isolation and lowered availability of social capital on Facebook, are the best predictors of PPD in mothers.

Author Keywords
childbirth; emotion; health; language; postpartum; social media; Twitter; wellness

ACM Classification Keywords
H.3.4; H.5.2; H.5.3

INTRODUCTION
Childbirth is a major life event that changes the lives of parents. Beyond the joy and happiness of childbirth, a significant portion of new mothers experience changes in mood. Such postpartum changes include the "baby blues" called *postpartum depression* (http://www.cdc.gov/reproductivehealth/Depression/). PPD is marked by symptoms such as sadness, fatigue, changes in sleeping and eating patterns, reduced libido, crying episodes, anxiety, and irritability [2]. Beyond negative influence on mothers, there is evidence that PPD can bear short and long-term negative effects on the child's behavior, emotional, and social wellbeing [16].

Despite these ramifications, PPD is underreported, with estimates that as many as 50% of cases of PPD go undetected [29]. Underreporting is believed to be due to factors such as the fear of the social stigma of depression, lack of social support, or assumptions that the mood changes are a normal part of the overwhelming nature of new motherhood [7]. Additional barriers to help-seeking include lack of knowledge about predisposition symptoms, the availability of appropriate remedial/prevention services, or even privacy concerns [7].

What makes this form of medical condition even more challenging is that there are no adequate ways for prevention or early detection [39], unless a woman has had a history of depression in the past. Clinical diagnosis of PPD is based on the mother's self-reported experiences (e.g., DSM-IV criteria based on the mood module from the PRIME-MD—Primary Care Evaluation of Mental Disorders: the Diagnostic and Statistical Manual of Mental Disorders (DSM) published by the American Psychiatric Association provides a common language and widely relied upon standard criteria for the classification of mental disorders), prior depression history, life stress, behaviors reported by

Towards Suicide Ideation Detection Through Online Conversational Context

Ramit Sawhney*† rsawhney31@gatech.edu Georgia Institute of Technology USA
Shivam Agarwal*† shivama2@illinois.edu University of Illinois at Urbana-Champaign USA
Atula Tejaswi Neerkaje* atulatejaswi@gmail.com Conversational AI and Social Analytics (CAISA) Lab, University of Marburg Germany
Nikolaos Aletras n.aletras@sheffield.ac.uk University of Sheffield United Kingdom
Preslav Nakov pnakov@hbku.edu.qa Qatar Computing Research Institute, HBKU Doha, Qatar
Lucie Flek lucie.flek@uni-marburg.de Conversational AI and Social Analytics (CAISA) Lab, University of Marburg Germany

ABSTRACT
Social media enable users to share their feelings and emotional struggles. They also offer an opportunity to provide community support to suicidal users. Recent studies on suicide risk assessment have explored the user's historic timeline and information from their social network to analyze their emotional state. However, such methods often require a large amount of user-centric data. A less intrusive alternative is to only use conversation trees arising from online community responses. Modeling such online conversations between the community and a person in distress is an important context for understanding that person's mental state. However, this is not trivial since comments can have diverse influence on a user in distress. Typically, a handful of posts receive a significantly high number of replies, which results in scale-free dynamics in the conversation tree. Moreover, psychological studies suggested that it is important to capture the fine-grained temporal irregularities in the release of vast volumes of comments, since suicidal users react quickly to online community support. Building on these limitations and psychological studies, we propose HCN, a Hyperbolic Conversation Network, a less user-intrusive method for suicide ideation detection. Through extensive quantitative, qualitative, and ablative experiments on real-world Twitter data, we find that HCN outperforms state-of-the art methods, while using 98% less user-specific data, and while maintaining a 74% lower carbon footprint and a 94% smaller model size. We also find that the comments within the

CCS CONCEPTS
• **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → **Social media**.

KEYWORDS
suicide ideation, conversation trees, social media

ACM Reference Format:
Ramit Sawhney, Shivam Agarwal, Atula Tejaswi Neerkaje, Nikolaos Aletras, Preslav Nakov, and Lucie Flek. 2022. Towards Suicide Ideation Detection Through Online Conversational Context. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3477495.3532068

1 INTRODUCTION
Social media enable people with suicidal thoughts to share their mental struggles with a large number of users online [84]. As a result, the connection between social media and suicide has become a public health concern [65], e.g., a study has shown that eight out of ten people who exhibit suicide ideation tend to disclose their suicidal thoughts online [39]. As a result, research efforts have increasingly shifted to analyzing mental health using social media for identifying individuals at risk [53], with the aim of potentially extending appropriate care [22, 113]. Recent advances in computa-

A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media

Stevie Chancellor
Georgia Tech
Atlanta, GA, US
schancellor3@gatech.edu

Michael L Birnbaum
Northwell Health
Glen Oaks, NY, US
mbirnbaum@northwell.edu

Eric D. Caine
University of Rochester
Rochester, NY, US
Eric_Caine@urmc.rochester.edu

Vincent M. B. Silenzio
University of Rochester
Rochester, NY, US
vincent.silenzio@rochester.edu

Munmun De Choudhury
Georgia Tech
Atlanta, GA, US
munmund@gatech.edu

Conference on Fairness, Accountability, and Transparency (FAT '19)*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287587

ABSTRACT

Powered by machine learning techniques, social media provides an unobtrusive lens into individual behaviors, emotions, and psychological states. Recent research has successfully employed social media data to predict mental health states of individuals, ranging from the presence and severity of mental disorders like depression to the risk of suicide. These algorithmic inferences hold great potential in supporting early detection and treatment of mental disorders and in the design of interventions. At the same time, the outcomes of this research can pose great risks to individuals, such as issues of incorrect, opaque algorithmic predictions, involvement of bad or unaccountable actors, and potential biases from intentional or inadvertent misuse of insights. Amplifying these tensions, there are also divergent and sometimes inconsistent methodological gaps and under-explored ethics and privacy dimensions. This paper presents a taxonomy of these concerns and ethical challenges, drawing from existing literature, and poses questions to be resolved as this research gains traction. We identify three areas of tension: ethics committees and the gap of social media research; questions of validity, data, and machine learning; and implications of this research for key stakeholders. We conclude with calls to action to begin resolving these interdisciplinary dilemmas.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; **Social media**; • **Applied computing** → *Psychology*;

KEYWORDS

mental health; ethics; machine learning; algorithms; social media

1 INTRODUCTION

Last year, Facebook unveiled automated tools to identify individuals contemplating suicide or self-injury [75, 62]. The company claims that they "use pattern recognition technology to help identify posts and live streams as likely to be expressing thoughts of suicide," which then can deploy resources to assist the person in crisis [75]. Reactions to Facebook's suicide prevention artificial intelligence (AI) are mixed, with some concerned about the use of AI to detect suicidal ideation as well as potential privacy violations [86]. Other suicide prevention AIs, however, have been met with stronger public backlash. Samaritan's Radar, an app that scanned a person's friends for concerning Twitter posts, was pulled from production, citing concerns for data collection without user permission [54], as well as enabling harassers to intervene when someone was vulnerable [4].

Since 2013, a new area of research has incorporated techniques from machine learning, natural language processing, and clinical psychology to categorize individuals' moods and expressed well-being from social media data. These algorithms are powerful enough to infer with high accuracy whether an individual might be suffering from disorders such as major depression [28, 19, 84, 73, 78], postpartum depression [26, 27], post-traumatic stress [21], schizophrenia [60, 6], and suicidality [15, 22]. These algorithms can also reveal symptomatology linked to psychiatric challenges, such as self-harm [89], severity of distress [13], or cognitive distortions [82]. Together, we use the term predicting *mental health status* to describe these mental disorders and related symptomatology.

“Decisions fall into the realm of ethics when they pertain to things within our control that can show respect to human being”

- WEB DuBois

What is the current landscape of ethics in predictive mental health research?

The stakes

- **Harms to users**
  - Deanonymization of mental health diagnosis
  - Data misuse
- **Inadvertently propagate unethical practices**
- **Perpetuate the Ought-Is gap [1]**
  - Space between what we *should* be doing and what *is*

[1] Bryan A. Sisk, et al. 2020. *The "Ought-Is" Problem: An Implementation Science Framework for Translating Ethical Norms Into Practice*. The American journal of bioethics

Operationalizing Ethics with Disclosure

Ethics disclosure
Any mention of decision-making that relates to ethics, potential harms, or broader impacts

IRB Approval
Data source
Interaction with subjects

"**Approval from the institutional review board was not sought** because these **data were freely available in the public domain** and researchers had **no interaction with the users**." [2]

[2] Birnbaum et al. 2017. *A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals.* JMIR

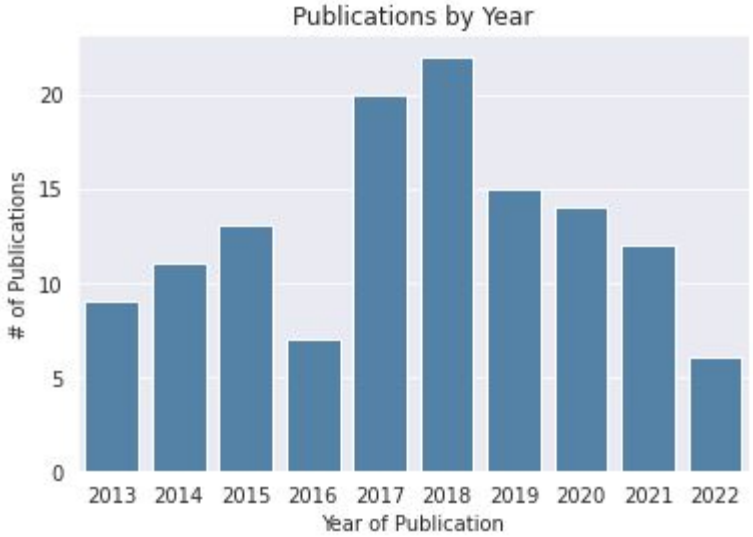
Methods

Dataset (n = 129)

2013 - 2022

Chancellor & De Choudhury corpus (n=79)

- 41 interdisciplinary venues
- 16 keyword pairs
- Filtered using PRISMA guidelines
- Snowball sampled



Methods

Human-Subjects
Data Source (Public v. Private)
Interaction with Subjects
IRB/Ethics board Mentioned
Consent
Compensation
Plan to inform subjects or gatekeepers
Data Centered
Subject inclusion/exclusion criteria
Modification of personal data
Data de-identified
Data sharing protocol
Data Storage
Impact
Harms consideration
Ethics disclosure section

Coding Schema (13 practices)

Inductively designed rubric

- Started with rough concepts
- Iteratively updated concepts as coded papers
- Formalized rubric once reached saturation

Analysis

- Three coding authors
- Cross-coded 10 papers
- Coded “Yes/No” if disclosure anywhere
- Memo for qualitative analysis
- Consultation from non-coding expert authors

Findings: Lack of standardization

	% Disclosed	Definition
Human-Centered		
Data source (public vs. private)	70.54	Origin of the data from and the visibility of this data to the general public
Interaction with Subjects	53.49	How researchers engaged with the account holders in the dataset, if at all
Ethics board mentioned	38.76	Whether and how an ethics review board was mentioned
Consent	24.03	Whether and how participants consented into the research, or logic about why consent was not sought
Compensation	9.30	Money received for participating in research, if any
Plan to inform subjects/gatekeepers	7.75	Intention to relay research results back to participants or communities
Data-Centered		
Subject inclusion/exclusion criteria	96.12	Standards to include participants in datasets
Modification of personal data	81.40	Concealing identifying information in the paper if used quotes or examples
Data de-identified	33.33	Anonymizing or changing data in the research process
Data sharing protocol	33.33	Guidelines for sharing resources (datasets, models, etc.)
Data storage	9.30	Methods of storing data
Impact		
Harms consideration	50.39	Direct or indirect mention of negative implications or impacts
Ethics consideration section	47.29	Specific written structure that discusses ethics concerns

Most disclosed (96%)
Subject inclusion/exclusion criteria

Least disclosed (8%)
Plan to inform subjects, moderators or other gatekeepers

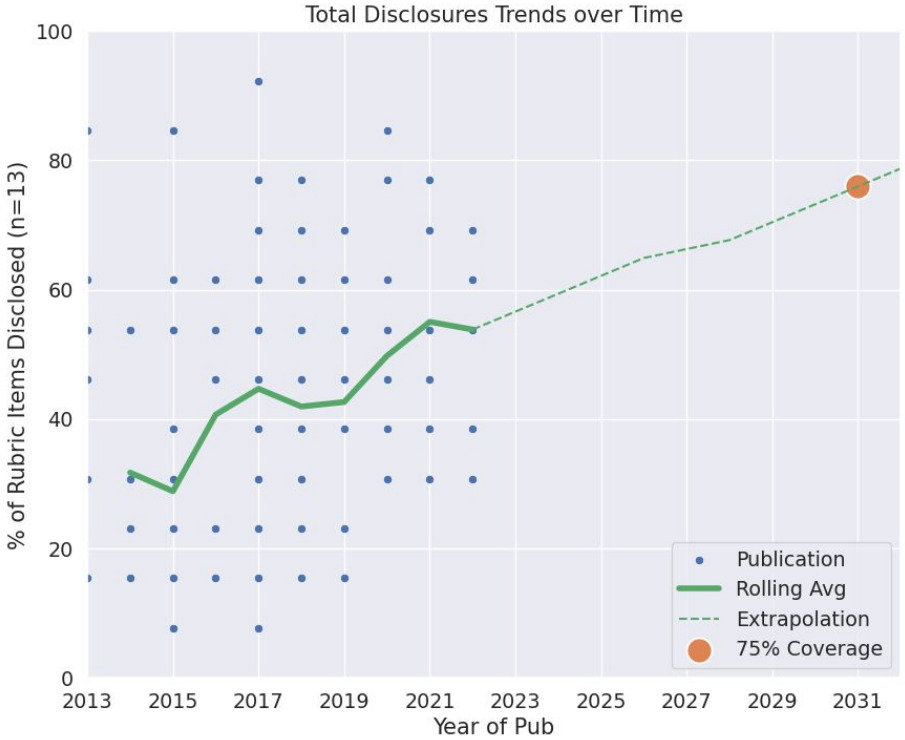
Lack of standardization

Disclosures present in super majority (>66%) of papers

- Three practices
- Methodological concerns
- Do “double disclosure duty”

	% Disclosed	Definition
Human-Centered		
Data source (public vs. private)	70.54	Origin of the data from and the visibility of this data to the general public
Interaction with Subjects	53.49	How researchers engaged with the account holders in the dataset, if at all
Ethics board mentioned	38.76	Whether and how an ethics review board was mentioned
Consent	24.03	Whether and how participants consented into the research, or logic about why consent was not sought
Compensation	9.30	Money received for participating in research, if any
Plan to inform subjects/gatekeepers	7.75	Intention to relay research results back to participants or communities
Data-Centered		
Subject inclusion/exclusion criteria	96.12	Standards to include participants in datasets
Modification of personal data	81.40	Concealing identifying information in the paper if used quotes or examples
Data de-identified	33.33	Anonymizing or changing data in the research process
Data sharing protocol	33.33	Guidelines for sharing resources (datasets, models, etc.)
Data storage	9.30	Methods of storing data
Impact		
Harms consideration	50.39	Direct or indirect mention of negative implications or impacts
Ethics consideration section	47.29	Specific written structure that discusses ethics concerns

Do practices improve over time?



Disclosure density:
Average % of disclosures in a paper

Year over year rate of change

Community is improving

8 years to reach 75% disclosure density

Sticky Practices: qualitative evidence

Secondary datasets

"[Previous work] built a data set for depression detection based on Reddit, which was named the Reddit data set. The samples in the Reddit data set are collected from the Reddit platform" - [paper in dataset]

Consent procedures

- Arguments that asking users to participate in their research could be coercive
- First appeared in 2015
- Used in multiple papers after (2019 - 2022)

Implicit Norms - Unwritten rules that are adopted through peer usage

Implications

- No common way of disclosing ethical decision-making
- We can't afford the status quo
- Papers propagate practice

Closing the Ought-Is Gap



Open, interdisciplinary conversations



Dedicated space in papers for ethics reasoning



Context documents to increase transparency

Thank you!

ajman004@umn.edu

Pdf of paper at **leahajmani.com**

A Systematic Review of Ethics Disclosures in Predictive Mental Health Research

Leah Ajmani*
ajman004@umn.edu
University of Minnesota
Minneapolis, MN, USA

Stevie Chancellor*
steviec@umn.edu
University of Minnesota
Minneapolis, MN, USA

Bijal Mehta
bm1126@georgetown.edu
Georgetown University
Washington, D.C., USA

Casey Fiesler
casey.fiesler@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Michael Zimmer
michael.zimmer@marquette.edu
Marquette University
Milwaukee, Wisconsin, USA

Munmun De Choudhury
munmund@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

Applied machine learning (ML) has not yet coalesced on standard practices for research ethics. For ML that predicts mental illness using social media data, ambiguous ethical standards can impact peoples' lives because of the area's sensitivity and material consequences on health. Transparency of current ethics practices in research is important to document decision-making and improve research practice. We present a systematic literature review of 129 studies that predict mental illness using social media data and ML, and the ethics disclosures they make in research publications. Rates of disclosure are going up over time, but this trend is slow moving – it will take another eight years for the average paper to have coverage on 75% of studied ethics categories. Certain practices are more readily adopted, or "stickier", over time, though we found prioritization of data-driven disclosures rather than human-centered. These inconsistently reported ethical considerations indicate a gap between what ML ethicists believe *ought* to be and what actually *is* done. We advocate for closing this gap through increased transparency of practice and formal mechanisms to support disclosure.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models*; • **General and reference** → *Surveys and overviews*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

ethics, mental health, systematic literature review, social media

ACM Reference Format:
Leah Ajmani, Stevie Chancellor, Bijal Mehta, Casey Fiesler, Michael Zimmer, and Munmun De Choudhury. 2023. A Systematic Review of Ethics Disclosures in Predictive Mental Health Research. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3593013.3594082

*Both authors contributed equally to the paper

Permission to make digital or hard copies of all or part of this work for personal or

1 INTRODUCTION

Mental illness is a pressing global health crisis [69], with nearly one in eight people worldwide living with a mental disorder [70]. This urgent issue has inspired researchers to identify innovative avenues for mental health prevention and treatment. As individuals turn to online social platforms to find and share their experiences with mental illness [27], researchers are exploring how and whether this data can identify mental illness with machine learning (ML) and artificial intelligence (AI) technologies. This area has grown into a thriving field of study since its origin in the early 2010s [28, 42]. The state-of-the-art can predict illnesses such as depression [29, 95, 109], anxiety [41, 90, 92], and suicidality [24, 59, 94] with social media data, as well as related experiences such as stress [83]. The impacts of this research could meaningfully change global mental health through earlier identification of illness, supporting quicker and more holistic triage in healthcare, and, ideally, informing digitally-delivered interventions [30].

As the field has matured, there are growing concerns about the research ethics of applied ML systems for predicting mental illness. Mental health prediction creates a ripe environment for possible harm. The domain uses *sensitive* data from a *vulnerable* population to communicate a possible *medical* diagnosis. This research often leverages public online communities, such as Reddit [71] or Instagram [78], but the data collected can include personal information [22, 33]. Experts have raised further concerns about issues around problem framing [15]; methods, such as data handling to maintain data subject privacy [114]; and the need for normative ethical research oversight from ethics review boards [3]. Inadver-