Ethical issues in text mining for mental health

Joshua August Skorburg & Phoebe Friesen

Introduction

Extending Moore's Law, which predicts that the number of transistors on a microchip doubles roughly every two years, the technology ethicist James Moor (2005) proposed Moor's Law: As technological revolutions increase their social impact, ethical problems increase.

There is little doubt that we are in the midst of such a technological revolution in the psychological sciences today, given the convergence of massive amounts of readily available data and rapid advances in computational tools and methods. We agree with Moor that as the social impact of this revolution increases, ethical issues will become more varied and more pressing. One of the areas where this is most apparent, and thus, the focus of the present chapter, is the intersection of text mining and mental health. Indeed, a recent systematic review of Machine Learning (ML) approaches to health data, containing over 100 studies, found that the most investigated problem was mental health (Yin et al., 2019). Relatedly, recent estimates suggest that between 165,000 and 325,000 health and wellness apps are now commercially available, with over 10,000 of those designed specifically for mental health (Carlo et al., 2019). The American Psychiatric Association has even developed rating systems and evaluation models for these apps (APA, 2019).

In light of these trends, the present chapter has three aims: (1) provide an informative overview of some of the recent work taking place at the intersection of text mining and mental health so that we can (2) highlight and analyze several pressing ethical issues that are arising in this rapidly growing field and (3) suggest productive directions for how these issues might be better addressed within future interdisciplinary work to ensure the responsible development of text mining approaches in psychology generally, and in mental health fields, specifically.

Here, then, is the plan for the chapter. In Section 1, we review some of the recent literature on text-mining and mental health in the contexts of traditional experimental settings, social media, and research involving electronic health records. Then, in Section 2, we introduce and discuss ethical concerns that arise before, during, and after research is conducted. Finally, in Section 3, we offer several suggestions about how ethical oversight of text-mining research might be improved in order to be more responsive to the concerns mapped out in Section 2.

Section 1. Text mining and mental health: A brief review

The first and perhaps most important feature to note about the literature on text mining and mental health is that it is vast and growing at a breakneck pace. Indeed, systematic reviews and meta-analyses from the past decade invariably remark that many of the reviewed publications seem to be clustered around the last few years in which the review was conducted. For example, in a 2019 review of 33 studies involving Machine Learning (ML) and Natural Language Processing (NLP) in the context of mental health, Le Glaz et al. (2019) note that the majority (63%) were published after 2016.

Almost out of necessity, then, the present review is focused less on tracking the cutting edge of this research and more on painting a picture of the field which can help to tease out the ethical issues that need to be addressed. Even with these caveats, there is still a wide variety of approaches, methods, and applications. So in the first place, it will be helpful to clarify some terminology.

For the purposes of this chapter, we use "text mining" as a catch-all term to cover a wide range of approaches which use computational methods on a variety of text corpora to investigate different aspects of mental health. While there is a broad array of research taking place under this heading, we will not consider here a number of related non-text based approaches to mental health, such as digital phenotyping, voice or facial recognition, the analysis of physiological or behavioral data from wearables, medical imaging, etc. We should note that our engagement with the mental health literature is also narrowly focused. Reflecting trends in the literature, we pay particular attention to cases involving schizophrenia, depression, and suicide, but it is worth noting that there are many other mental health conditions we do not engage with in detail. Some, but not all, of our analyses will extend to other conditions.

We can mark three further distinctions to guide our review. Our aim here is not to provide an exhaustive or historical review. Instead, we want to shed light on the ethical issues in this domain, and so we will focus on three main areas where such ethical issues are salient: (1) traditional experimental settings, (2) social media, and (3) electronic health records.[1] In what follows, we will highlight a few research papers from each of these areas which will then guide our ethical analyses in Section 2.

1.1 Traditional experimental settings

Text mining is being taken up in traditional experimental settings for mental health research, often involving a corpus comprised of written observations or patient narratives, such as responses in interviews or surveys. These approaches have been used broadly in research related to depression, anxiety, suicide, Alzheimer's, autism, addiction, Parkinson's and other disorders (for a review, see Shatte, Hutchinson, & Teague, 2019).

A representative and impressive research program in this domain explores how various semantic and syntactic features of language are associated with schizophrenia. Some early and influential work in this vein used Latent Semantic Analysis (LSA) techniques on interview transcripts or text messages to quantify language incoherence, the measure of which could then be used to distinguish patients with a diagnosis of schizophrenia from healthy controls (Elvevåg et al., 2007). A similar approach was also used to successfully distinguish between first-degree relatives of schizophrenia patients and unrelated healthy controls (Elvevåg et al., 2010).

A more recent paper by Bedi et al. (2016) used similar methods to predict, at the individual level, the later onset of psychosis in high-risk youths. In this study, the researchers conducted hour-long, open-ended narrative interviews with 34 help-seeking, clinically high-risk youths (aged 14-27) and transcribed the interviews into text. Participants were clinically assessed using the Structured Interview for Prodromal Syndromes/Scale of Prodromal Symptoms (SIPS/SOPS) at baseline, and every 3 months following, up to a period of 2.5 years. The researchers then trained a convex hull classifier on features derived from semantic coherence analyses (i.e., discontinuity between adjacent sentences within the sampled text, or average coherence between adjacent sentences), along with Part-of-Speech (POS) tagging to discriminate among participants who did transition to psychosis ($N = 5$) and those who did not transition to psychosis ($N = 29$). Cross-validation was performed by sequentially excluding participants to later be used for testing.

[1] While these three categories – traditional research settings, social media, and the electronic health record - represent a significant amount of mental health research involving text analysis, it is worth noting that these categories are not exclusive, and researchers working in this domain often utilize two or more in conjunction. For example, Merchant et al. (2019) linked medical records and social media data of consenting participants in order to evaluate their ability to predict medical conditions from user generated content on social media.

Impressively, using three parameters: minimum semantic coherence (e.g. increased discontinuity between phrases), normalized use of determiners (e.g. 'that', 'what', 'whatever', 'which') and maximum phrase length (known as "poverty of speech"), the classifier yielded 100% accuracy in predicting transition to psychosis. This automated text analysis method out-performed the predictions generated from standard clinical ratings (e.g. SIPS/SOPS), which yielded only 79% accuracy (Bedi et al., 2016).

More recently, Corcoran et al. (2018) replicated and extended these findings using a different protocol (a prompt-based rather than narrative-based speech assay) in a larger sample. In this study, the logistic regression classifier achieved 83% accuracy in predicting psychosis onset.

Though the sample sizes in these studies are relatively small, the potential implications are large. Taken together, they suggest that individuals likely to transition to psychosis could be reliably identified earlier and more accurately using automated text analysis methods, compared to traditional clinical measures. Indeed, the authors claim that their methods and others like them "may present an opportunity to move psychiatry beyond reliance on self-report and clinical observation toward more objective measures of health and illness in the individual patient" (Bedi et al., 2015, 6). Moreover, when these automated text-based methods perform as well as, or better than, traditional approaches, they offer time- and cost-effective alternatives to advance both research and clinical practice.

1.2 Social Media

Here, we will review a few text-mining approaches where the relevant corpus is comprised of the thoughts, feelings, and behaviours of patients from more "naturalistic" settings (as opposed to controlled, experimental settings). The most prominent and increasingly well-studied examples of this approach involve data from social media. One of the oft-touted strengths of these approaches is that they can fill in gaps in the "clinical whitespace," or the time between structured, formal interactions with healthcare systems (Coppersmith et al., 2017).

Some early and influential work in this area was conducted by De Choudhury and colleagues. For example, De Choudhury et al. (2013) used linguistic features (among other kinds features, such as behavioral attributes, time of day, social network structure, etc.) in Twitter posts to predict the onset of Major Depressive Disorder (MDD). In this study, the authors first used a crowdsourcing approach (Amazon Mechanical Turk or mTurk) to construct the "ground truth" regarding the presence of MDD. Participants on mTurk were given standardized clinical depression surveys, including the Center for Epidemiologic Studies Depression Scale (CES-D) and the Beck Depression Inventory (BDI), along with questions about depression history and standard demographics. Participants were also asked to share their Twitter usernames, so that the researchers could mine their Twitter posts.

476 users provided useable data and two classes of users were constructed: An MDD positive class of 171 participants scoring high on the CES-D and BDI, and a negative class of 305 participants who scored low. Across both classes, the researchers were able to mine 2,157,992 total tweets with an average of over 4,500 posts per user over the course of one year. With the MDD positive class, the authors report that individuals with depression show lowered social activity, greater negative emotion, high self-attentional focus, increased relational and medicinal concerns, and heightened expression of religious thoughts. De Choudhury et al. then built a Support Vector Machine (SVM) classifier to predict the onset of depression at the individual level. The best-performing model yielded 70% accuracy overall, with increased performance among models that incorporated linguistic style information alone. This leads the authors to conclude that "social media activity provides useful signals that can be utilized to classify and predict whether an individual is likely to suffer from depression in the future" (De Choudhury et al., 2013, 136).

Related work from this group has used similar methods to predict post-partum depression using Twitter posts (De Choudhury & Counts, 2013) and Facebook posts (De Choudhury, Counts, Horvitz, Hoff, 2014), as well as to construct a more general social media depression index (De Choudhury, Counts, & Horvitz, 2013).

Perhaps the most prominent (and as we will see below, ethically fraught) use cases involve text mining on social media platforms to predict suicide risk. For example, Coppersmith et al. (2018) utilize data from two sources: 1) text from 186 users of OurDataHelps.org[2] who attempted suicide and 2) publicly available posts on Twitter. This combined dataset contained 418 users with at least 6 months of prior social media posting history who attempted suicide, along with an equal number of age- and gender-matched controls whose posts were scraped from social media. Each user had an average of 473 social media posts, contributing to a total of 197,615 posts from those who would go on to attempt suicide in the next 6 months, and an equal number for controls, for a total of 395, 230 posts analyzed. Using a deep-learning approach (Long Short-Term Memory), the researchers trained a text classification model to aggregate risk scores from individual posts to predict a given user's suicide risk.

Their results indicate that, for models which tolerate a 10% false positive rate (e.g. categorizing a user as high-risk for suicide when, in fact, they are not), the true positive rate (e.g. correctly categorizing a high-risk user) ranged from 70 – 85%. To anticipate some of the themes we will consider below, the authors conclude that these results "demonstrate that signals exist within social media data that are quantifiable and relevant to suicide risk," and that their algorithms were able to "identify people at risk for suicide from the analysis of the language of their social media posts, at levels of precision that suggest clinical utility, and at the period early enough to permit reasonably scalable and durable interventions" (Coppersmith et al., 2018, 6).

1.3 Medical records

In this section, we will review some recent text-based approaches where the relevant corpus is comprised of patient medical information captured within electronic health records (EHR). Such records include data related to patients' diagnoses, medical history, treatments, test results, clinical notes, discharge summaries, etc. There is a significant amount of text-based research which utilizes EHR data to predict or categorize patients according to distinct psychiatric diagnoses or suicidal risk. Here, however, we will focus on another body of research, concerned with identifying in advance which psychiatric patients are most likely to place a substantial burden on the health care system, in terms of both costs and clinical outcomes. Drawing on information 'hidden' in the EHR, this research seeks to make inferences related to questions concerning service use, such as how often a patient is likely to return or how long until they are readmitted, or medical outcomes, such as whether their condition is likely to be resistant to treatment.

An example of this stems from work by Roysden et al. (2015), who developed a machine learning algorithm to predict which patients referred to mental health services[3] were likely to be either high or low service utilizers. Relying on the use of structured administrative data, lab results, and free text notes from 12,759 patients who had seen a mental health provider for the first time, Roysden et al. sorted patients into those that decreased service use after their mental health encounter, and those that used services frequently after their first encounter. Frequent users were defined as those utilizing services in the 95th or 99th percentile (the percentile was based on data from a control group). The researchers then trained a random forest classifier in order to predict which

[2] The authors describe this data as "from a set of users who have graciously donated their data to support research in this area through OurDataHelps.org. Users of this platform sign up and authorize access to data from their digital life—social media (eg, Facebook, Twitter, Instagram, Reddit, Tumblr), wearable (eg, Fitbit, Jawbone), and other technology (eg, Strava, Runkeeper). Users also fill out questionnaires asking for basic demographic data as well as for information about their history with various mental health conditions" (Coppersmith et al., 2018, 3). For more information, see https://ourdatahelps.org/

[3] Roysden et al. use the term 'behavioral health' as opposed to 'mental health'; for consistency within this chapter, we use mental health here.

patients would fall in either category. Unfortunately, the classifier was "only modestly able to outperform the extremely simplistic prediction that patients will use the same amount of health care before and after the intervention" (p. 2070).

Looking at a different, but often overlapping, measure, Perlis et al. (2012) developed a classifier using clinical notes from 127,504 patients diagnosed with major depressive disorder, in order to predict cases of treatment resistant depression. While the classifier outperformed billing data, which is often relied upon as a predictor of treatment outcomes, the authors warned that relying on clinical narrative notes may not always be a good indicator of a patient's medical status.

A closely related project in this domain is that of predicting which patients are most likely to be readmitted to the hospital after discharge. McCoy et al. (2015) sought to predict these clinical events by performing a sentiment analysis on discharge notes within the EHR. Using Pattern, an open source tool developed outside of clinical settings in order to mine text for polarity, subjectivity, intensity, and negativity, the team classified discharge summaries according to both polarity and subjectivity and examined whether these features could be used to predict readmission. They found that greater positive sentiment in a discharge note predicted a lower likelihood of readmission (a decreased risk of approximately 12% per standard deviation above the mean) for both patients admitted for psychiatric or general medical care.

Similarly, Rumshisky et al. (2016) examined discharge notes in order to predict readmission, focusing on patients receiving mental health care in particular. Using topic modeling, which uses word co-occurrence patterns to learn latent topics in a text, their analysis revealed that notes including topics such as psychiatric comorbidities (e.g. eating disorders, substance use), medical illness (dementia, infection), or severe depression, predicted readmission at an earlier date. The authors reported that their model performed with a sensitivity of 75% but a specificity of 63%.

Another example of text analysis relying on EHR data involves the detection of unreported or unrecorded medical information through the examination of unstructured clinical notes. Wu et al. (2013) used EHR data from 5,588 patients with a diagnosis of schizophrenia, schizoaffective disorder, or bipolar disorder to compare their smoking status as recorded in their medical record with their likely smoking status based on an analysis of the text included in their medical record. Written assessments, progress notes and correspondence were examined through an application developed by the authors. This application, CRIS-IE-Smoking, involves a shallow parsing, rule-based approach to analyze keywords within the EHR, in order to determine the smoking status of the patient. Rules were developed in order to identify cases in which the word smoke could be present, but not indicate smoking status (e.g. 'smoke alarm' or 'smoke weed') or cases in which they indicate a high likelihood of a positive smoking status (e.g. 'the patient smokes 20 cigarettes a day'), and were finetuned until a precision (positive predictive value) of 93% (based on 100 random documents) was reached. The medical records alone indicated that only 10.7% of the patients had a positive smoking status. The NLP model, however, indicated that 64.1% of patients ought to be designated as having a positive smoking status.

Section 2. Ethical issues in text mining and mental health

With these examples from three major areas of text mining research in the mental health context (experimental settings, social media, EHR) in mind, we are now in a position to tease out and analyze the ethical issues. While there are many ways to divide up the ethical landscape, for present purposes we have selected three broad categories to guide our discussion. First, we will discuss what takes place (or should take place) before the research begins. Second, we will discuss ethical issues during the research process. Third, we will discuss issues of translation and application which are based on the findings of the research.[4]

2.1 Before Research: Ethics Review

2.1.1 Gaps in Oversight

As any academic researcher knows, before carrying out a research project, it must be approved by an Institutional Review Board (IRB)[5] [6]. IRBs are composed of various stakeholders with different areas of expertise who are well-positioned to identify different ethical considerations, and generally, they will require researchers to specify things like the overall research question guiding the data collection, the type of data collected, the rationale and purpose of collecting such data, the limits on how long the researchers will have access to the data, the potential harms and benefits of the research, the methods used to ensure confidentiality, etc.

However, research utilizing text mining, and various other forms of research involving novel computational techniques, are often not required to undergo ethics review by an IRB. In the first place, formal ethical review provided by IRBs has a relatively narrow ambit: only institutions receiving federal funds are required to abide by the Federal Policy for the Protection of Human Subjects, better known as the "Common Rule," which, among other things, mandates IRBs. One example of an oversight gap, then, is that text mining research projects that do not receive federal funding will not be required to be reviewed by an IRB. Though some states and institutions extend requirements related to IRB review of research protocols, many research projects still fall outside of the scope of required review. Similarly, some research groups may voluntarily choose to submit their research proposals to various kinds of ethics committees, but there is variability with respect to the independence of these committees, the extent of their oversight, etc. Indeed, in many of the studies we reviewed in Section 1.2, it is difficult to discern whether the research was approved by an ethics committee, or whether any kind oversight was present. In cases where the stakes are high and vulnerable populations are involved (e.g. suicide research), it should not be difficult to appreciate the value of formal ethical review by an independent committee.[7]

However, even when research utilizing text mining receives federal funding or takes place within an institution that requires IRB review, it is still possible that it will fall through additional gaps in oversight. This is because the paradigmatic case of medical research, from which our current processes of research ethics review are derived, is the randomized control trial (RCT). This type of research involves identifiable participants who are able to participate in an informed consent process, with clear expectations about when the research will begin and end, what information will be derived from it, and how it will be used. Much data-driven biomedical research, including text mining, turns this traditional model on its head and disrupts the system of research ethics review built upon it. Consider several examples of triggers for ethics review that are now outdated.

[4] Another category where significant ethical issues arise is, of course, in the context of data collection and processing. However, to date, work in this area has received the lion's share of attention, with familiar themes such as biases encoded in word embeddings (e.g., Caliskan et al. 2016) or issues of over- and under-representation of minority groups in data mining (e.g., Barocas & Selbst, 2016). For the purposes of this chapter, we will instead focus most of our efforts on the issues of ethics review and clinical application, which we think are equally pressing, but have received less attention in the literature. We do not mean to suggest that the issues we raise here are exhaustive though.

[5] Also known as Research Ethics Committees (RECs) and Research Ethics Boards (REBs) outside of the United States.

[6] The history of how this came to pass is important in its own right, though beyond the scope of the present chapter (but for a review, see Stark, 2011).

[7] To be clear, our aim is decidedly not to single out specific authors or research teams. The oversight asymmetries noted in this section will be present so long as there are different standards for publicly- and privately-funded research, and these very same ethical issues arise for other research involving suicide, mental health, etc.

First, the types of data often used in text analysis pose problems for how we tend to think about 'human participants' in research. Rather than having research participants who will show up in person and can consent to a clearly defined set of research activities, text analysis often involves large sets of de-contextualized data, raising the question of whether there are human subjects involved, and what obligations a researcher has to them. This is all much hazier than in a clinical trial. Furthermore, the risks that arise in these forms of research often apply at the level of populations and communities and are not easily understood at the level of an individual (Metcalf & Crawford, 2016).

Second, 'research' and 'practice' have traditionally been thought of as distinct activities within medicine and medical ethics. In the United States, research is defined as that which is intended to contribute to generalizable knowledge, in contrast with practice, which is intended to contribute to patient well-being. Within insitutions, a project falling into the category of 'research' triggers review by an IRB (Ryan et al., 1979). But these domains are often difficult to disentangle in data-driven mental health research because the processes of collecting data and delivering an intervention are often linked within a feedback loop. This is exemplified by social media-based suicide prediction and prevention technologies that use supervised learning. While data is being collected, categorizing users into low risk and high risk classes, interventions (e.g. chatbots) are being deployed, and data regarding the impact of these interventions is then fed back into the model (e.g. a false positive may signal that a particular phrase should be assigned less predictive weight) (Barak-Corren et al., 2016). This means that drawing a sharp distinction between research and practice for the purpose of triggering ethical review becomes very difficult in the domain of text mining.

Third, traditional ethics review is triggered by research that involves an 'intervention' involving human participants. However, as Metcalf & Crawford explain,

> Data analytics techniques rarely appear as a direct 'intervention' in the life or body of an individual human being, which is one of the key requirements for research to be regulated in the USA. The action of Big Data analytics happens mostly at a remove from the point of data collection, which is the most plausible analog for an 'intervention.' Instead, it is focused on data sets that likely have a long lifespan and may be continuously updated and re-analyzed (Metcalf & Crawford, 2016, pp. 2-3).

These novel data lifecycles challenge traditional notions of research, which had a clear timeline for both data collection and analysis. In text mining, the 'intervention' often comes much later and can be hard to anticipate during early research phases.

Finally, research that is conducted with publicly available data sets, such as those scraped from social media, is typically exempt from ethics review, because the participants are thought to have no reasonable expectation of privacy. However, the potential to create novel medical data from unrelated user data posted online raises questions about what users might reasonably expect, and whether 'publicly available' provides a useful demarcation anymore. Furthermore, research that creates medical data (e.g. a suicide risk score) from publicly available data (e.g. social media data) is also not protected by HIPAA (the Health Insurance Portability and Accountability Act), which outlines data privacy protections for all medical data collected in clinical settings (HIPAA, 2004).

The foregoing suggests that there are a number of oversight gaps into which text mining research might fall. Projects that are not federally funded are not always subject to formal ethical review. But even if text mining research is federally funded, it may well be exempt from ethical review, due to the outdated triggers for ethical review noted above. It seems likely that a significant amount of text mining research falls into at least one of these gaps.

However, even when ethics review is required, the format and frameworks taken up by IRBs may still be insufficient for capturing the ethical issues that arise within these novel forms of research. Metcalf & Crawford

(2016) note how research using big data moves ethical inquiry "away from traditional harms such as physical pain or a shortened lifespan to less tangible concepts such as information privacy impact and data discrimination." In some cases, early research and subsequent interventions are developed so far apart that it is difficult to foresee the potential implications or uses of a tool developed in the earlier stages. So even if research ethics review takes place, a thorough analysis of the possible risks and benefits related to the research cannot be conducted.

Consider a new technology company, Receptiviti, which, among many other things, uses natural language processing to monitor employees' emails, voice, or Slack messages, in order to keep employers informed of their employees' happiness at work (Werber, 2019). The company is based on James Pennebaker's research, which has found correlations between linguistic patterns (e.g. use of pronouns) and individuals' moods. While Receptiviti makes clear they currently only intend to use the data for group level analysis (e.g. the women in finance are unhappy) and not to generate mood reports on individual employees, such an application may not be far off. It is unlikely that Pennebaker or the ethics committee that examined his earlier work foresaw this future application. Importantly, it would be unreasonable to ask them to. However, the introduction of mental health screening tools such as this in the workplace introduces novel risks and raises many ethical issues related to consent, information sharing, and the ability of employees to opt out. In our current system of research ethics review, these risks are unlikely to be captured.

These considerations highlight some of the gaps in oversight that exist at the pre-research phase, and they underscore the importance of developing new models of research ethics review which are responsive to both the robust principles of classical research ethics and the cutting-edge developments in data science. In Section 3 below, we will advance some modest suggestions to deal with these issues.

2.2 During Research: Consent and Data Sources

2.2.1 Consent

Informed consent is often treated as the core of research ethics. The most straightforward reason for this is that the procedure of disclosing necessary information to a research participant such that they can understand the information and voluntarily decide to participate (or not) in the research is one of the best ways to ensure the protection of participants' interests, health, and well-being, while respecting them as autonomous individuals.

Though not without their own problems, the concepts and procedures related to informed consent are relatively well-worked out in the biomedical ethics literature. In the case of text-mining, however, things are much murkier. In the first place, as we noted above, it is less clear what counts as "research," "human subjects," or "interventions" in text-mining. Second, it is not clear that existing consent procedures often used in text-mining fulfil the spirit (much less the letter) of informed consent mechanisms meant to protect the interests of research subjects.

Nowhere were these issues more apparent than Facebook's (in)famous emotional contagion study (Kramer, Guilory, & Hancock, 2014). Much ink has been spilled about the case.[8] Our interest here is not to deliver a verdict on ethical merits of the case, but rather, to highlight the specific role of informed consent in research in this vein, which uses text-mining methods at scale. By way of review, the researchers in the study systematically manipulated the content of nearly 700,000 Facebook users' News Feeds in two experiments in 2012. In both experiments, the researchers deployed the widely used software Linguistic Inquiry and Word

[8] See, for example, the special issue of the journal *Research Ethics* dedicated to the emotional contagion experiment (for the editorial introduction, see Hunter & Evans, 2016).

Count (LIWC) to determine if a given Facebook post contained positive or negative content. Then, on the basis of this designation, in one experiment, users saw slightly fewer negative posts (compared to a control) and in the other experiment, users saw slightly fewer positive posts (compared to a control). In both experiments, the key dependent variable was the percentage of all words produced by a given user that were positive or negative following the manipulation. As the title of the paper suggests, the researchers found (limited) evidence for the "emotional contagion" hypothesis. When users saw slightly less negative content, their subsequent posts tended to be more positive, and when users saw slightly less positive content, their subsequent posts tended to be more negative.

As is well known, the publication of this study provoked wide-ranging backlash. For present purposes, we want to focus on the role of informed consent, specifically, though of course there are many other ethical dimensions to this case. In the paper, the authors write that:

> LIWC was adapted to run on the Hadoop Map/Reduce system and in the News Feed filtering system, such that no text was seen by the researchers. As such, it was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research (Kramer, Guilory, & Hancock, 2014, p. 2).

Let us assume for the sake of argument that Facebook's Data Use Policy did indeed contain all of the relevant information about user participation in such experiments. Would merely clicking "agree" to the terms and conditions of creating a Facebook account constitute informed consent? According to Benbunan-Fich (2017), at the time of the emotional contagion study, Facebook's Terms of Service (TOS) was around 6,700 words. Despite the fact that empirical evidence shows that people spend, on average, about half a minute reading TOS, it would have taken the average reader over 30 minutes to read Facebook's TOS.

In light of these kinds of claims, there is a growing consensus that while such "notice and consent" practices may pass for *implied* consent and suffice for legal liability, these practices fall well short of the standard of *informed* consent and are inadequate with respect to ethical liability (Nissenbaum, 2009). In other words, when users are confronted with essay-length, abstruse legalese in a dialogue box on the screen of their smartphone, it is difficult to see how the average user could be said to be making a fully informed decision about being involved in experiments when pressing "agree." If these critiques - that users of social media platforms cannot be said to provide fully informed consent to participating in research – hit their mark, then notice and consent mechanisms will not provide an adequate ethical justification for text-mining. And there are, of course, a broader range of issues related to the ethics of social media platforms conducting social science-style experiments on their users in the first place (see e.g., Grimmelman, 2015).

This issue of whether (and to what extent) consent mechanisms can adequately protect individuals are especially pointed in the social media context for the reasons noted above. But as before, by pointing this out, we do not mean to suggest that research that does not involve social media data will be free of these worries. One of the features which has driven the explosion of text-based research for mental health is also the source of some of the thorniest ethical issues: the text data can be almost infinitely re-purposed, re-combined, re-analyzed, and shared. Even in some of the most prototypical settings – such as a face-to-face encounter with a clinician who enters notes into a medical record – it will often be practically impossible to inform patients or users about the possible future uses to which their data will be put. For example, do participants in a feasibility study using NLP to predict psychosis know that their data might then be combined with other datasets and re-analyzed? Are participants using social media aware that the content of their posts could be published in an academic article describing research in which they were participants? To the extent that questions like these are answered in the negative, then researchers using text-mining methods at scale will have to take additional measures (beyond notice and consent) to insure the responsible use of their tools.[9]

[9] All of that said, we do not wish to imply that fully informed consent is always required in text mining for mental health. In many cases, de-identified data will be used for a short-term project that will help to develop a tool, and after

2.2.2 Data Sources: From Subjective to Objective

Mental health research has long revolved around self reports from patients, as more 'objective' data (biological or otherwise) is often lacking within the field, compared to the rest of medicine. With the introduction of new data-driven techniques, including text mining, digital phenotyping, and genomic research, a lot of excitement is being generated about the potential to collect such 'objective' data. While these methodologies bring significant promise to the field of psychiatry, it is important to also consider the potential impact of research which is able to bypass the first-person experiences of patients in this way.

For example, research programs seeking to make predictions or classify patients on the basis of EHR data can have the effect of distancing the participant from the research process, by drawing on clinical notes, composed of indirect observations of patients written by health care professionals. This process can serve to erase the voice of those being researched, replacing it with the views and values of those offering treatment, and leading to potential increases in stigma, violations of privacy, and the translation of clinical biases.

As evidenced by the examples of text mining research involving EHR data discussed above, clinical notes are often used to develop predictions related to costs, comorbidities, and service use. While these are important indicators of relevance to clinicians and those organizing the delivery of clinical care, they can also serve to exacerbate stigmatizing views of psychiatric patients as expensive and non-compliant. Patients in mental health crises, and especially those who are known to use services frequently, are already among the most stigmatized in health care, with some referring to providing care to this population as 'dirty work' (Shaw, 2004). Health care providers often respond to such patients by delaying care, as they believe therapeutic success is unlikely, or with hostility, as a result of their frustrations. Offering practitioners data which predicts which patients are likely to be high service utilizers (Roysden & Wright, 2015) or readmitted to the hospital (McCoy et al., 2015) may contribute to these issues, flagging those who are likely to be the most burdensome on the healthcare system, and in doing so, offering further grounds for distinguishing in advance between 'good' and 'bad' patients.

Additionally, predictive tools developed from clinical notes often seek to identify information that patients have not, or are unlikely to, provide on their own (e.g. likelihood of adherence, frequency of engagement in 'bad habits' such as smoking or drug use) (Lyons, Aksayli, & Brewer, 2018; Wu et al., 2013). While this can be of clinical utility, the implementation of such technologies raises concerns related to patient privacy. Patients may not want to discuss their smoking habits, the possibility that they may stop taking their prescribed psychotropic medications, or the fact that they are thinking about suicide, because they have had negative clinical experiences in the past (e.g. forced treatment) and wish to avoid similar experiences. To develop tools that predict the likelihood of these behaviors, without prior consent, can constitute a violation of these patient's right to privacy.

A reliance on clinical notes as one's primary data set also introduces risks related to the inscription of clinician biases in the tools being developed. As McCoy et al. reported in discussing their analysis of how clinician sentiment might predict readmission, "among psychiatric patients, public insurance was associated with less positive sentiment" (McCoy et al., 2015, p. 6). This aligns with a significant body of research documenting biases in both physician attitudes towards and treatment of marginalized patients. For example, van Ryn and Burke found that clinicians tended to rate black patients as less likely to adhere to medical advice, as less intelligent, and less likely to be 'the kind of person they could be friends with' in comparison to patients who were white (van Ryn & Burke, 2000).

development, the data will be deleted. In these cases, the risks to participants are low and the potential benefits are significant, and given these features, consenting each individual participant may be overly burdensome. However, in cases in which the risks of a research project are larger, the data lifecycle is longer, the applications are significant or unknown, or the data being collected are especially sensitive, informed consent becomes more important.

While the use of EHR data to make predictions about patient behavior can often lead to ethical worries such as these, text mining in mental health research that seeks to incorporate patient voices and inform patients of information relevant to them is also occurring. For example, Hart et al. examined discussions on social media related to psychotropic medications to determine what information patients were being exposed to online and discovered that a lot of medical information related to psychotropics was negative (Hart, Perlis, & McCoy Jr, 2019). Additionally, Iqbal et al. examined clinical records to predict which patients would be most likely to experience adverse events, such as movement disorders, as a result of using antipsychotics (Iqbal et al., 2015).

2.3 After Research: Clinical Applications

2.3.1 Clinical Applications

A fundamental, and yet often overlooked, feature to be kept in mind when developing text-based tools for mental health is the clinical utility that such tools are likely to produce. While it may seem obvious that the ability to predict the occurrence of a mental health issue will lead to clinical benefits, this is only the case when treatments and resources are available to support those identified through such predictive processes. Unfortunately, despite significant hype related to neuroscience and genomics in the past several decades, we have seen remarkably little improvement in the treatments available to those living with diagnoses of mental disorders; this should serve as a warning to those engaged in psychiatric research involving computational technologies (Torous & Baker, 2016).

This pessimism is particularly appropriate in relation to psychosis-related conditions, such as schizophrenia. Treatments for schizophrenia have not improved significantly since antipsychotics were discovered in the 1950s, despite the introduction of a second generation of antipsychotics in the 1990s (Friesen, 2019). Rates of recovery in those diagnosed with schizophrenia have remained at 13% for decades (Jaaskelainen et al., 2013), and on average, individuals with a diagnosis of schizophrenia have a life expectancy 14.5 years shorter than their un-diagnosed counterparts (Hjorthøj, Stürup, McGrath, & Nordentoft, 2017). In recent years, concerns related to the efficacy and risks of antipsychotics, the standard first line treatment for psychosis-related disorders, have grown substantially, and some evidence indicates that the use of antipsychotics may lead to an increase in psychotic symptoms over time (Harrow, Jobe, & Faull, 2014; Wunderink, Nieboer, Wiersma, Sytema, & Nienhuis, 2013). Increasingly, experts in the field disagree over whether these medications are doing more good than harm (Gøtzsche, Young, & Crace, 2015; Harrow & Jobe, 2013; Sohler et al., 2016).

A state of clinical uncertainty such as this raises questions regarding the clinical utility and application of computational tools designed to help identify those who are currently experiencing or are likely to experience psychosis in the future, such as those by Bedi et al. and Corcoran et al., discussed above. Where there is no consensus regarding what the best course of treatment is for someone at risk, more information regarding who is at risk is not always easy to translate into clinical benefits. This means that knowing who is likely to develop psychosis might not always mean better treatment for those identified, and potentially, such treatment could be harmful in the long run (e.g. if it turns out that antipsychotics are doing more harm than good). In other cases, the development of predictive tools in health care (e.g. to support screening for melanoma) are likely to contribute to clear and immediate clinical benefits for patients.

Efforts should be made early on, then, to think about the potential clinical utility or application of tools in this domain. After outlining how they developed a tool using natural language processing to support the prediction of psychiatric readmission, Rumshisky et al. acknowledge that "beyond prediction, it is also possible that identifying particularly high-risk topics will facilitate the development of interventions to reduce readmission risk in particular subgroups" (Rumshisky et al., 2016, p. 3). While this would be an excellent outcome, it is unclear how the predictive tool developed by Rumshisky et al. is likely to support the development of such interventions. There is already substantial data pertaining to which groups are at highest risk of psychiatric readmission. While this tool may provide additional predictive power, no information or discussion is given to suggest to how the team's efforts can be translated into better clinical care. This is not atypical in the literature, and it points to the obvious if understated importance of thinking through the likely impact of projects in this domain before significant resources are spent. Given the wealth of available data and the speed at which analytic techniques are being developed, it is understandable that a lot of enthusiasm has been generated. However, beyond the development of new technologies to identify, predict, and classify mental disorders, it is important to consider whether there are tools and resources available to support those identified as at risk, and how these new technologies will contribute to patient wellbeing.

Section 3. Guiding frameworks and future directions

In response to the issues we raised above, we want to conclude with three suggestions. First, given the gaps in oversight outlined in Section 2.1, we will advance a modest proposal that could provide a firmer basis for ethical review mechanisms than we currently have. Second, and relatedly, we will make a recommendation with regards to how non-academic ethical review mechanisms (such as those used in mental health start-ups) might incorporate more rigorous ethical analyses. Third, we will offer a guiding ideal for future work on the ethics of text-mining, where ethical considerations are not simply offloaded to a review board (and thereby perceived as mere compliance), but rather are integrated within and central to the practice of text-mining. Finally, we will speak briefly to the turn towards participatory research in psychiatry and how mental health research involving text mining could benefit from following suit.

3.1 From "research" to risk

As discussed above, a significant amount of text-based mental health research is not required to undergo research ethics review, because it is not federally funded and therefore not subject to restrictions and regulations that are attached to such funding. However, we also showed that even in cases in which this research is federally funded and takes place within academic settings, it is also often exempt from review, because it fails to fit within the traditional mold of medical research. Returning to our discussion from Section 2.1, we want to suggest here that traditional methods of ethics review, which follow from a distinction between 'research' and 'practice', are inadequate in light of advances in data science broadly construed. Thus, we recommend moving away from our current system of research ethics oversight and introducing different criteria for determining when research ethics review should take place.

Rather than trying to draw a line between 'research' and 'practice', or to determine in which cases patients or users are involved enough to be considered 'participants', or what constitutes an 'intervention' in text mining, a better route may be to require ethics oversight and support for projects that *contain significant, unpredictable, or novel risks* (Friesen, Kearns, Redman, & Caplan, 2017). While there is much conceptual work required to flesh out the relevant notions of risk, this proposal is nonetheless poised to help fill in the gaps in oversight that currently exist. Such gaps exist, we have shown, both in projects taking place in industry (which often fall outside of federal funding or do not involve HIPAA protected data), and also within academic or collaborative settings, due to our outdated definitions of human participants and research. Moving ethics review to this risk-based framework can also help to reduce the regulatory burden faced by many projects, providing a larger overlap between where risk appears and where oversight is required. In such a system, for example, Pennebaker's early work on the relationship between text and mood would be unlikely to require much oversight (so long as the relevant kinds of risks were acceptably low), but the introduction of Receptiviti's monitoring system in the workplace may well trigger an independent ethics examination.

3.2 Consumer Subject Review Boards

In an influential paper in the *Stanford Law Review*, Ryan Calo introduced the following thought experiment:

> the Federal Trade Commission, Department of Commerce, or industry itself commissions an interdisciplinary report on the ethics of consumer research. The report is thoroughly vetted by key stakeholders at an intensive conference in neutral territory…As with the Belmont Report [one of the foundational documents in human subjects research ethics], the emphasis is on the big picture, not any particular practice, effort, or technology. The articulation of principles is incorporated in its entirety in the Federal Register or an equivalent. In addition, each company that conducts consumer research at scale creates a small internal committee comprised of employees with diverse training (law, engineering) and operated according to predetermined rules (Calo, 2013, 102).

In short, the idea here is to develop what Calo calls Consumer Subject Review Boards (CSRB), which aim to suitably translate and update some of the principles from canonical research ethics to address the gaps in oversight between academic vs industry contexts. Calo is quick to stress that he is not merely advocating for IRB-type structures to be copied over into industy settings. There are reasons to think that such efforts could be misguided.[10]

Instead, a more realistic approach recognizes that there will not be an over-arching, one-size-fits-all framework to guide ethical analyses of research in industry settings. But despite this, abstract guiding ethical principles can still be useful. This is not a small task. But we agree with danah boyd, who argues that "we need a socio-technical model of ethical oversight that creates a conversation between those doing the manipulation and those who are producing the data being manipulated. Such a model must be co-constructed by companies and researchers, not simply imposed by outsiders who think that they understand corporate decision-making" (boyd, 2015, 12).

Still, the devil is in the details and some recent work which has explored the plausibility of Calo's CSRB has pointed to a deep tension in any such approach. A fundamental question is whether CSRBs should be organized as internal corporate entities or as external oversight bodies. As Polonetsky, Tene, and Jerome (2015) point out, on the one hand, firms may worry that sharing sensitive business information with an external ethical review board could compromise competitive advantage. But on the other hand, many critics are unlikely to be satisfied by a wholly internal process, where the ethical review would likely amount to little more than a non-binding, advisory role. This has played out in practice, where some companies, including Facebook, have set up internal processes of research ethics review, but worries about the independence and authority of those involved in these processes remain (Jackman & Kanerva, 2016).

While these are difficult questions, there are nonetheless clear advantages to exploring the viability of CSRBs in the context of the text-mining research reviewed in this chapter. Calo argues that the processes involved in creating CSRBs (e.g. elaborating ethical principles to guide research, analogous to the Belmont Report) would obviously benefit consumers and users by increasing protections from their current levels. But these proposed forms of hybrid ethical oversight also stand to benefit technology companies doing mental health research. For example, Calo argues that CSRBs could help to guard against the kind of media frenzies that characterized the response to Facebook's emotional contagion study. A more structured and formalized ethical framework would also help policy managers to assess different kinds of risks. Similarly, they could reduce regulatory uncertainty. And perhaps, if the protections were sufficiently robust and transparent, and pathways of accountability were sufficiently reliable (and concerns about "ethics washing" were adequately addressed) then CSRBs could help to "furnish a measure of legitimacy" to industry-led research (Calo, 2013, 102).

[10] Calo mentions reasons such as: the need for firms to operate at different scales than academic institutions, the need for firms to move more quickly, to satisfy investors, as well as fundamental differences such as firms often being aimed primarily at profit while academic institutions are often aimed primarily at knowledge production. It could also be argued that present regulations for industry research are sufficient, or that self-regulation in industry would suffice without the need for formal oversight structures (see also Schroeder 2016 for further discussion of these issues).

3.3 Integrated Ethics

Given the gaps in oversight we have been discussing, there is space for important progress to be made. But focusing only on questions about oversight and formal ethical review does not tell the whole story. And in fact, this narrow focus might even be counterproductive. In short, the worry about focusing narrowly on oversight is that it might lead to what Leonelli (2016) describes as *externalization*. Though her remarks concern data science more generally, they are especially apt for the kinds of text-mining approaches we have been discussing here. Externalization, according to Leonelli, "is the idea that ethics is extraneous to technical concerns and constitutes an add-on to scientific research that is imposed and governed by outside forces, rather than an unavoidable and constructive part of daily scientific decision-making (2016, 3). Thus construed, ethical consideration is relegated to "a 'tick-box' exercise, which researchers often view as a drag on their research time, and which has provided an excuse to delegate away any potential concerns with the ethical implications of research work" (Leonelli 2016, 7).

For example, a researcher developing a model to predict various dimensions of mental health might think: "I just develop the tools. It's not up to me how people use them." Or, "the compliance team will handle any privacy issues, that's their job." Or, "My background is in computer science and statistics, not medical ethics." In all of these cases, downstream ethical implications are outsourced, and the can is kicked down the road. This is especially worrisome because, in many cases, the researchers themselves will be the most familiar with the technical details and nuance which are intimately connected with – and if Leonelli is correct, inextricable from - the social and ethical implications.

While the standardized ethical principles of a CSRB envisioned by Calo and others certainly could represent an important advancement, such proposals alone will not guarantee the responsible development of text-mining technologies. What is needed in addition is a conceptual shift whereby technical considerations are viewed as entangled with ethical considerations. A catalyst for such conceptual shifts will likely involve ethical trainings *integrated* within the computational sciences, rather than, as is too often the case, stand-alone modules treated as an afterthought or add-on. A model that could be useful in moving in this direction is that of research ethics consultations, in which ethicists provide support to, or are embedded on research teams, in order to help alert researchers to ethical issues that are likely to appear in their work early on. Cho et al. (2008) describe the role of research ethicists as "providing real-time advice to scientists on the conduct and dissemination of research to help identify and incorporate ethical and societal considerations into their research."

3.4 Participatory Research

Along with the rest of medicine, the field of psychiatry has taken a significant participatory turn in recent years. A growing emphasis is being placed on the importance of the involvement of service users in psychiatric policy, research, and practice. This turn can be seen most clearly in mental health research, where user-led research is increasingly taking place, and in some countries, such as the United Kingdom, where the ability to receive mental health research funding is often dependent on one's ability to demonstrate that patient involvement will be built into a research project (Department of Health, 2009; Sweeney, Beresford, Faulkner, Nettle, & Rose, 2009). Justifications for the democratization of psychiatric research are both epistemic and ethical. The involvement of service users in the research process has been shown to contribute to epistemic aims by making research outputs more relevant to the populations being studied, to positively impact recruitment and retention, and to help researchers identify potential hurdles or assumptions in their research (Brett et al., 2012; Domecq et al., 2014). However, many also emphasize the importance of including service users in psychiatric research because they have a right to contribute to knowledge related to them and because of the harms that have historically taken place in psychiatric research (Beresford, 2002; Jones, Harrison, Aguiar, & Munro, 2014).

Interestingly, a lot of the research in text-based computational psychiatry seems to be moving in the opposite direction. While there are initiatives towards involvement in the field, the majority of these appear to take place during the translational phase of research (Friedman, Kahn, Borning, & Huldtgren, 2013; Pelletier, Rowe, François, Bordeleau, & Lupien, 2013; Thieme et al., 2013). In research involving text mining, such as those described above, participants are often quite distant from the research process, as a result of the commonality of consent waivers, de-identified data sets, and the lack of direct communication with those whose data is under study. Given these tendencies, which lead to conditions in which the "familiar human subject is largely invisible or irrelevant to data science", it is unsurprising that patient involvement is not more common within the field (Metcalf & Crawford, 2016).

Involving patients in the design and implementation of text mining research in mental health could counteract some of the ethical issues flagged in this chapter. Rather than moving further away from patient's experiences, as in the examples drawing on EHR data, participatory research in text mining could ensure that those with lived experience can shape central decisions related to research design from the start, potentially preventing research that is likely to contribute to stigma, violate patient privacy, and incorporate biases into data sets. Furthermore, the involvement of psychiatric service users in research could contribute to better processes of informed consent and could help to fill in some of the many gaps in oversight that exist in this domain.

Conclusion

In this chapter, we hope to have provided a relevant overview of some text-mining research in the mental health context, involving traditional experimental contexts, social media research, and research relying on electronic health records. On the basis of this overview, we then introduced ethical concerns that arise before, during, and after research has taken place. We then concluded with suggestions about how ethical oversight of mental health research involving text mining might be improved through a shift towards risk-based oversight, through the development of CSRBs, through increased ethical integration within research, and through the adoption of more participatory research processes. While our efforts here have been primarily focused on text-mining in the mental health context, we hope that the concepts, distinctions, considerations, and recommendations introduced here will generalize to other research involving text analysis at scale.

References

Abbe, A., Grouin, C., Zweigenbaum, P., & Falissard, B. (2016). Text mining applications in psychiatry: a systematic literature review. *International journal of methods in psychiatric research*, *25*(2), 86-100.

American Psychiatric Association (2019). Mental health apps: https://www.psychiatry.org/psychiatrists/practice/mental-health-apps

Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., . . . Reis, B. Y. (2016). Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry, 174*(2), 154-162.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev., 104*, 671.

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., ... & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia, 1*, 15030.

Benbunan-Fich, R. (2017). The ethics of online research with unsuspecting users: From A/B testing to C/D experimentation. *Research Ethics, 13*(3-4), 200-218.

Beresford, P. (2002). User Involvement in Research and Evaluation: Liberation or Regulation? *Social Policy and Society, 1*(2), 95-105. doi:10.1017/S1474746402000222

Boyd, D. (2016). Untangling research and practice: What Facebook's "emotional contagion" study teaches us. *Research Ethics, 12*(1), 4-13.

Brett, J., Staniszewska, S., Mockford, C., Herron-Marx, S., Hughes, J., Tysall, C., & Suleman, R. (2012). Mapping the impact of patient and public involvement on health and social care research: a systematic review. *Health Expectations, 17*(5), 637-650.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186.

Calo, R. (2013). Consumer subject review boards: A thought experiment. *Stan. L. Rev. Online, 66*, 97-102.

Carlo, A. D., Ghomi, R. H., Renn, B. N., & Areán, P. A. (2019). By the numbers: ratings and utilization of behavioral health mobile applications. *npj Digital Medicine, 2*(1), 54.

Cho, M. K., Tobin, S. L., Greely, H. T., McCormick, J., Boyce, A., & Magnus, D. (2008). Strangers at the benchside: Research ethics consultation. *The American Journal of Bioethics*, 8(3), 4-13.

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights, 10*, 1178222618792860.

Coppersmith G, Hilland C, Frieder O, Leary R. (2017). Scalable mental health analysis in the clinical whitespace via natural language processing. *IEEE EMBS International Conference on Biomedical & Health Informatics* (BHI), 393–396. New York, NY: IEEE.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., ... & Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry, 17*(1), 67-75.

De Choudhury, M., Counts, S., Horvitz, E. J., & Hoff, A. (2014, February). Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 626-638). ACM.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013, June). Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media.*

De Choudhury, M., Counts, S., & Horvitz, E. (2013, April). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3267-3276). ACM.

De Choudhury, M., Counts, S., & Horvitz, E. (2013, May). Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 47-56). ACM

Department of Health. (2009). *Putting people at the heart of care.* Retrieved from https://webarchive.nationalarchives.gov.uk/20130123200554/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_106038

Domecq, J. P., Prutsky, G., Elraiyah, T., Wang, Z., Nabhan, M., Shippee, N., . . . Firwana, B. (2014). Patient engagement in research: a systematic review. *BMC Health Services Research, 14*(1), 89.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research, 93*(1-3), 304-316.

Elvevåg, B., Foltz, P. W., Rosenstein, M., & DeLisi, L. E. (2010). An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of neurolinguistics, 23*(3), 270-284.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory* (pp. 55-95): Springer.

Friesen, P. (2019). Expanding Outcome Measures in Schizophrenia Research: Does RDoC Pose a Threat? *Philosophy, Psychiatry, & Psychology, 26*(3), 243-260.

Friesen, P., Kearns, L., Redman, B., & Caplan, A. L. (2017). Rethinking the Belmont Report? *The American Journal of Bioethics, 17*(7), 15-21. doi:10.1080/15265161.2017.1329482

Gøtzsche, P. C., Young, A. H., & Crace, J. (2015). Does long term use of psychiatric drugs cause more harm than good? *BMJ, 350*, h2435.

Grimmelmann, J. (2015). The law and ethics of experiments on social media users. *Colo. Tech. LJ*, 13, 219.

Harrow, M., Jobe, T., & Faull, R. (2014). Does treatment of schizophrenia with antipsychotic medications eliminate or reduce psychosis? A 20-year multi-follow-up study. *Psychological Medicine, 44*(14), 3007-3016.

Harrow, M., & Jobe, T. H. (2013). Does long-term treatment of schizophrenia with antipsychotic medications facilitate recovery? *Schizophrenia Bulletin, 39*(5), 962-965. doi:10.1093/schbul/sbt034

Hart, K. L., Perlis, R. H., & McCoy Jr, T. H. (2019). What do patients learn about psychotropic medications on the web? A natural language processing study. *Journal of Affective Disorders.*

Hjorthøj, C., Stürup, A. E., McGrath, J. J., & Nordentoft, M. (2017). Years of potential life lost and life expectancy in schizophrenia: a systematic review and meta-analysis. *The Lancet Psychiatry, 4*(4), 295-301. doi:https://doi.org/10.1016/S2215-0366(17)30078-0

Hunter, D. & Evans, N. (2016). Facebook emotional contagion experiment controversy. *Research Ethics 12* (1), 2-3.

Iqbal, E., Mallah, R., Jackson, R. G., Ball, M., Ibrahim, Z. M., Broadbent, M., . . . Dobson, R. J. (2015). Identification of adverse drug events from free text electronic patient records and information in a large mental health case register. *PloS One, 10*(8), e0134208.

Jaaskelainen, E., Juola, P., Hirvonen, N., McGrath, J. J., Saha, S., Isohanni, M., . . . Miettunen, J. (2013). A systematic review and meta-analysis of recovery in schizophrenia. *Schizophrenia Bulletin, 39*(6), 1296-1306. doi:10.1093/schbul/sbs130

Jackman, M., & Kanerva, L. (2016). Evolving the IRB: Building robust review for industry research. Washington and Lee Law Review Online, 72(3), 442.

Jones, N., Harrison, J., Aguiar, R., & Munro, L. (2014). Transforming research for transformative change in mental health: Towards the future. *G., Nelson, B., Kloos, J. Ornelas,(Eds.), Community psychology and community mental health*, 351-372.

Kane, J. M., Robinson, D. G., Schooler, N. R., Mueser, K. T., Penn, D. L., Rosenheck, R. A., . . . Heinssen, R. K. (2015). Comprehensive Versus Usual Community Care for First-Episode Psychosis: 2-Year Outcomes From the NIMH RAISE Early Treatment Program. *American Journal of Psychiatry*, appiajp201515050632. doi:10.1176/appi.ajp.2015.15050632

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788-8790.

Le Glaz, A., & Kim-Dufor, D., Taylor, R., Devylder, J., Walter, M., Berrouiguet, S., & Lemey, C. (2019). Machine learning and natural language processing in mental health: a systematic review (Preprint). 10.2196/preprints.15708.

Leonelli, S. (2016). Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2083), 20160122.

Lyons, M., Aksayli, N. D., & Brewer, G. (2018). Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior, 87*, 207-211.

Marks, M. (2019). Artificial Intelligence Based Suicide Prediction. *Yale Journal of Health Policy, Law, and Ethics*, Forthcoming.

McCoy, T. H., Castro, V. M., Cagan, A., Roberson, A. M., Kohane, I. S., & Perlis, R. H. (2015). Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: an electronic health record study. *PloS one, 10*(8), e0136341.

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society, 3*(1), 2053951716650211.

Moor, J. H. (2005). Why we need better ethics for emerging technologies. *Ethics and information technology, 7*(3), 111-119.

Narayanan, A., & Felten, E. W. (2014). No silver bullet: De-identification still doesn't work.

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*: Stanford University Press.

Pelletier, J.-F., Rowe, M., François, N., Bordeleau, J., & Lupien, S. (2013). No personalization without participation: on the active contribution of psychiatric patients to the development of a mobile application for mental health. *BMC Medical Informatics and Decision Making, 13*(1), 78. doi:10.1186/1472-6947-13-78

Perlis, R. H., Iosifescu, D. V., Castro, V. M., Murphy, S. N., Gainer, V. S., Minnier, J., ... & Fava, M. (2012). Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychological medicine, 42*(1), 41-50.

Polonetsky, J., Tene, O., & Jerome, J. (2015). Beyond the common rule: Ethical structures for data research in non-academic settings. *Colo. Tech. LJ*, 13, 333.

Roysden, N., & Wright, A. (2015). Predicting health care utilization after behavioral health referral using natural language processing and machine learning. In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 2063). American Medical Informatics Association.

Rumshisky, A., Ghassemi, M., Naumann, T., Szolovits, P., Castro, V. M., McCoy, T. H., & Perlis, R. H. (2016). Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational psychiatry, 6*(10), e921.

Ryan, K., Brady, J., Cooke, R., Height, D., Jonsen, A., King, P., . . . Turtle, R. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research.* Washington, D.C.

Schroeder, C. (2015). Why can't we be friends: A proposal for universal ethical standards in human subject research. *Colo. Tech. LJ, 14*, 409.

Shatte, A. B., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine, 49*(9), 1426-1448.

Shaw, I. (2004). Doctors,"dirty work" patients, and "revolving doors". Qualitative Health Research, 14(8), 1032-1045.

Simon, G. E., Johnson, E., Lawrence, J. M., Rossom, R. C., Ahmedani, B., Lynch, F. L., ... & Shortreed, S. M. (2018). Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *American Journal of Psychiatry, 175*(10), 951-960.

Sohler, N., Adams, B. G., Barnes, D. M., Cohen, G. H., Prins, S. J., & Schwartz, S. (2016). Weighing the evidence for harm from long-term treatment with antipsychotic medications: A systematic review. *American Journal of Orthopsychiatry, 86*(5), 477.

Stark, L. (2011). *Behind closed doors: IRBs and the making of ethical research*: University of Chicago Press.

Sweeney, A., Beresford, P., Faulkner, A., Nettle, M., & Rose, D. (2009). *This is survivor research*: pccs Books.

Thieme, A., Wallace, J., Johnson, P., McCarthy, J., Lindley, S., Wright, P., . . . Meyer, T. D. (2013). *Design to promote mindfulness practice and sense of self for vulnerable women in secure hospital services.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Torous, J., & Baker, J. T. (2016). Why Psychiatry Needs Data Science and Data Science Needs Psychiatry: Connecting With Technology. *JAMA Psychiatry, 73*(1), 3-4. doi:10.1001/jamapsychiatry.2015.2622

United States. (2004). *The Health Insurance Portability and Accountability Act (HIPAA).* Washington, D.C.: U.S. Dept. of Labor, Employee Benefits Security Administration.

van Ryn, M., & Burke, J. (2000). The effect of patient race and socio-economic status on physicians' perceptions of patients. *Social Science and Medicine, 50*(6), 813-828.

Werber, C. (2019, September 24). Your work emails contain subtle clues about your emotional state. *Quartz*. Retrieved from https://qz.com/work/1709790/work-email-holds-clues-to-your-mental-health/

Wu, C. Y., Chang, C. K., Robson, D., Jackson, R., Chen, S. J., Hayes, R. D., & Stewart, R. (2013). Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. *PloS one, 8*(9), e74262.

Wunderink, L., Nieboer, R. M., Wiersma, D., Sytema, S., & Nienhuis, F. J. (2013). Recovery in remitted first-episode psychosis at 7 years of follow-up of an early dose reduction/discontinuation or maintenance treatment strategy: long-term follow-up of a 2-year randomized clinical trial. *JAMA Psychiatry, 70*(9), 913-920. doi:10.1001/jamapsychiatry.2013.19

Yin, Z., Sulieman, L. M., & Malin, B. A. (2019). A systematic literature review of machine learning in online personal health data. *Journal of the American Medical Informatics Association, 26*(6), 561-576.