JAMIA Open, 3(1), 2020, 9–15
doi: 10.1093/jamiaopen/ooz054
Advance Access Publication Date: 22 January 2020
Perspective

Perspective

Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence

Colin G. Walsh,[1] Beenish Chaudhry,[2] Prerna Dua,[3] Kenneth W. Goodman,[4] Bonnie Kaplan,[5] Ramakanth Kavuluru,[6] Anthony Solomonides,[7] and Vignesh Subbian[8]

[1]Biomedical Informatics, Medicine and Psychiatry, Vanderbilt University Medical Center, 2525 West End, Suite 1475, Nashville, TN, USA, [2]School of Computing and Informatics, University of Louisiana at Lafayette, Lafayette, Louisiana, USA, [3]Department of Health Informatics and Information Management, Louisiana Tech University, Ruston, Louisiana, USA, [4]Institute for Bioethics and Health Policy, University of Miami, Miller School of Medicine, Miami, Florida, USA, [5]Yale Center for Medical Informatics, Yale Bioethics Center, Yale Information Society, Yale Solomon Center for Health Law & Policy, Yale University, New Haven, Connecticut, USA, [6]Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, Lexington, Kentucky, USA, [7]Outcomes Research and Biomedical Informatics, NorthShore University HealthSystem, Research Institute, Evanston, Illinois, USA and [8]Department of Biomedical Engineering, Department of Systems and Industrial Engineering, The University of Arizona, Tucson, Arizona, USA

Corresponding Author: Colin G. Walsh, MD, MA, Biomedical Informatics, Medicine and Psychiatry, Vanderbilt University Medical Center, 2525 West End, Suite 1475, Nashville, TN, USA; Colin.walsh@vanderbilt.edu

Received 8 March 2019; Revised 29 July 2019; Editorial Decision 17 September 2019; Accepted 30 October 2019

ABSTRACT

Effective implementation of artificial intelligence in behavioral healthcare delivery depends on overcoming challenges that are pronounced in this domain. Self and social stigma contribute to under-reported symptoms, and under-coding worsens ascertainment. Health disparities contribute to algorithmic bias. Lack of reliable biological and clinical markers hinders model development, and model explainability challenges impede trust among users. In this perspective, we describe these challenges and discuss design and implementation recommendations to overcome them in intelligent systems for behavioral and mental health.

Key words: behavioral health, ethics, predictive modeling, artificial intelligence, precision medicine, health disparities, algorithms, mental health

INTRODUCTION

Artificial intelligence (AI), a rich area of research for decades, has attracted unprecedented attention in healthcare in the past few years. Academics and industry collaborators apply AI to a variety of biomedical issues ranging from clinical prediction to phenotyping complex disease states, or for guiding diagnosis, prognosis, treatment, and lifestyle change.[1–11] While public perceptions of AI center on strong or artificial general intelligence (the ability for smart agents to think as humans do), most if not all published efforts in biomedicine focus on weak or applied AI.

Applied AI (subsequent mentions of "AI" in this piece will refer to applied or weak AI), from complex multivariate models to simple clinical prediction rules, has been a mainstay in prediction of hospital readmissions,[12] acute kidney injury,[13] mortality,[14,15] and imaging (eg, retinal imaging[16,17] or radiology[18]) for over a decade. But it has more recently been applied to challenges in mental and behavioral health (eg, predicting suicide,[19] treatment resistance in depression,[20] dementia,[21] and more). Behavioral health includes emotional, mental, and social factors as well as behaviors to prevent illness (eg, avoiding substance abuse) and promote wellness (eg, ex-

ercise).[22] Since we do not yet live in a world where behavioral healthcare is simply "healthcare" as we hope will one day be the case, informaticians must be attuned to the ways in which mental and behavioral health differ from other areas in medicine. Failure to do so leads to unintended consequences and potential harms or, at best, the most common fate for published predictive models: that they are never used in clinical practice.

To help the informatics community reach the potential for AI to impact behavioral healthcare, we will discuss issues either unique to or exemplified by behavioral health. We will then share recommendations for designing and deploying intelligent systems in this domain.

CHALLENGES

Behavioral health poses uncommon challenges to designing impactful AI. Broadly, these challenges include (1) lack of data because of (i) stigma and silence (ie, under-reporting, under-coding) and (ii) lack of or unreliable biomarkers; (2) algorithmic biases; and (3) danger of inappropriate use due to gaps in interpretability or explainabilty, trust, and privacy concerns.

Lack of data

Under-reporting and under-coding

One in five adult Americans (~43.8 million) experience a mental health disorder in any given year, regardless of race, religion, gender, or socioeconomic status.[23] Behavioral health issues like abuse of tobacco, alcohol, and illicit drugs account for ~$232 billion in healthcare costs annually.[24] However, approximately two-thirds of those with mental illness suffer privately, without treatment.[25] Stigma, both self-directed and public, contributes to this dilemma. Self-stigma feeds self-discriminating and stereotyping behavior with negative professional and personal consequences.[26] Public stigma leads to restricted opportunities, coercive treatment, and reduced independence for individuals with mental and behavioral health conditions. Social stigma, for example, for opioid use disorders, can have implications for public health and punitive policy-making.[27]

Silence leads to under-reporting, but under-coding exacerbates this gap. Under-coding is particularly common in primary care and in patients presenting with multiple co-morbidities.[28,29] For example, in patients presenting with both mental illness and a chronic condition, clinicians are more likely to code and claim for just the chronic condition.[29] Even when documented, behavioral health symptoms might not be recorded in structured forms. For example, suicidal thoughts are only coded 3% of the time in primary care even when documented in notes.[30] Qualifying words such as "likely" or "suspected" soften firm diagnoses. Coding suicidal ideation or severe symptoms might raise administrative (eg, expectation of triggering alerts downstream) or liability concerns for providers,[30] even if they spend sufficient time assessing and planning an effective management plan with those patients.

Unreliable or absent biomarkers and objective measures

Unlike other illnesses such as congestive heart failure or sepsis,[31,32] mental illness or behavioral health concerns are not directly diagnosed via objective measures, laboratory reports or other quantitative biomarkers. Recent trends suggest this fact might change, such as a study linking heart rate-related metrics to post-traumatic stress disorder (PTSD).[33] Instead, diagnoses result from medical history, general physical examination findings—anxiety or nervousness and a thorough psychiatric exam—and provider impressions. Often, these potentially predictive data are either recorded only in unstructured data such as text or in covert forms, for example, as text about "trouble sleeping" without overt documentation of insomnia related to depression. As a result, algorithms reliant on readily available structured data might fail to incorporate diagnostic or prognostic risk factors.

Attempts to incorporate unstructured text via natural language processing (NLP) in behavioral health have been published.[34–40] However, the tradeoff between sensitivity and specificity is particularly challenging here, because of the well-known problems of dimensionality and negation. Adding large numbers of NLP predictors (eg, bag of words or word2vec[41]) to models adds to dimensionality and potential for overfitting. Moreover, the prevalence of clinical screening in practice and documentation of negative assertions means that basic NLP or regular expressions might fail. For instance, most documentation of suicidal thoughts in electronic health records (EHRs) describe when risk isn't present (eg, screening), not when it is.

Potential for algorithmic bias

All algorithms based in AI generally involve bias, but not in the sense familiar to the public.[42] High bias in algorithms implies missing important relationships between the input features and output variables (referred to as underfitting). A related concept, high variance, implies that models learn almost all data points in given training data but fail to perform well on new data (referred to as overfitting). Reducing bias or variance tends to increase the other quantity. This bias-variance tradeoff defines model performance.[43] A rich literature in AI includes representational and procedural bias in the interest of generalizing algorithmic learning beyond "strict consistency with individual instances."[44] The media and public perceive a broader definition of bias as prejudice against a person, group, race, or culture, which we refer to as "algorithmic bias" here, as others have.[45–49]

Health disparities contribute to algorithmic bias. Mental illnesses such as schizophrenia and affective disorders are likely to be misdiagnosed, particularly among African Americans and lower socioeconomic groups.[50] Women have higher prevalence of major depression, anxiety disorders, and PTSD,[51] whereas Native Americans have disproportionate rates of suicide-associated deaths.[52] Cultural differences in language and mannerisms, difficulties relating between patients and therapists of sufficiently different backgrounds, and prevailing societal notions about various groups' susceptibility to mental illness add to algorithmic bias.[53] While AI might be well-suited for the diagnosis and prognosis in complex, nuanced phenotypes like these, we risk producing models that incorporate bias in underlying data[54] (eg, lack of nonbinary gender categories in EHRs) and algorithmic bias in model specification. Finally, model developers rarely have "ground truth" to use for validation and model training. The dependent variables and "gold standards" might also rely on expert review or chart validation and might be flawed. A final critical issue is that of a harmful feedback loop: existing disparities may lead to unrepresentative training data. This bias may seep into predictive models, which further exacerbate disparities owing to biased predictions for certain minorities and vulnerable segments of patient populations.

Considerable scholarship discusses algorithmic bias: in data, in model specification, in deployment and use, and, if machine-learning was involved, in model training and its trainers.[55–57] Robust discussion includes the need for data sharing and re-use for

transparency in how algorithms work, their accuracy and reliability,[52] the explainability of their conclusions, and accountability for using or not using them. In health care, other incentives might also influence how data are recorded or interpreted. Reimbursement and billing, social or employment consequences, or other financial and stigma avoidance strategies could bias what information is collected and how it is recorded, as well as the output of algorithms processing patient data.

Inappropriate use, interpretability/explainability, and trust

We have learned from ethical analyses of biomedical informatics literature that appropriate users and uses of technology are often identified based on potential to improve care.[51,58,59] If an algorithm contributes positively to a patient's treatment, then that is a good reason to use it. If it harms or does not help, then we should be hard-pressed to justify its use. In some cases, empirical research can help answer these questions.

Investigating disruptions in behavioral health requires linking data from multiple levels, from cells to the individual to the environment. Furthermore, a single biological disturbance may produce two different psychological issues and, conversely, two different neurological disturbances may produce the same or similar psychological symptoms.[60] This complexity makes interpretability, "how" a model arrives at an output, and explainability, "why" a model arrives at that output, more complex. The literature in explainable AI has expanded in recent years.[61–64] We highlight that explainability and interpretability are particularly important in behavioral health because we have so few complex models in behavioral healthcare delivery unlike, for example, models predicting readmissions or sepsis. Thus, we have not reached an inflection point where users "trust" that AI provides accurate recommendations even if the process that led to them cannot be interrogated.

Generating explanations to interpret results from a model is critical for most conditions of interest. Clinicians rightly crave actionable insights at the time of decision-making in line with the "Five Rights" of decision support (the right information, delivered to the right person, in the right intervention format, through the right channel, and at the right time in workflow).[65] But models derived from large complex datasets are harder to interpret.[65] With complex nonlinear models such as deep neural networks, the task of generating explanations is non-trivial, especially in the context of EHRs that contain a mixture of structured and textual data.

Because many outcomes in behavioral and mental health might be clinically rare yet have very high stakes, end-users must also be given appropriate context in which to interpret predictions. For example, many published predictive models of suicide risk show high sensitivity at the expense of low precision.[66] Preventive efforts might be wasted on large numbers of false positives secondary to imprecise models. At the same time, false negatives might lead to loss of life from suicide and loss of trust in automated recommendations. The clinical harms of such events are further compounded by liability and legal implications.

A corollary challenge relates to relative inattention to calibration performance of predictive models in favor of discrimination performance (eg, c-statistics, sensitivity, and specificity). A recent systematic review showed 79% (56) of included studies did not address calibration in their results. If an outcome occurs 0.2% of the time in one clinic, a 20% predicted risk is quite high, but a clinician not educated in this interpretation might not prioritize this number without proper context.[67] Failure to account for and educate end-users such as clinical providers about these issues will compromise trust in algorithmic recommendations as well as uptake.

Attempts to hybridize "black box" prediction with more readily interpretable algorithms are underway.[58,68] We highlight this challenge as algorithms in behavioral health have high potential to be care-altering, career-altering (eg, employment or military deployment decisions), or life-altering. Providers might feel compelled to respond to a "high risk" designation for suicide risk beyond, for example, readmission risk. Thus, the onus remains on informaticians to forge trust with end-users (ie, clinicians and patients) in demonstrating the reasoning behind the recommendations made by algorithms.

RECOMMENDATIONS

Despite challenges in implementing AI for behavioral health, appropriate effort to overcome them supports continued innovation in this domain. Our recommendations follow to best integrate intelligent systems alongside humans to augment, and not replace, what people do best: taking a broad view, exercising judgment, and accepting moral responsibility.

Foster trust through education and transparency

The issue of trust can be addressed at both the community-level and the technology-level. Ample literature focuses on model development, validation, and implementation. Far less focuses on providing tools and knowledge to noninformatics clinical providers on how to best integrate risk prediction into practice. For providers to better judge algorithmic outputs, designers and informaticians should contextualize and educate the broader community about how to assess, integrate, and evaluate clinically applied AI. An AI-educated practitioner will also be far more likely to notice errors or potentially harmful edge cases before bad outcomes can occur.

George E.P. Box is famously paraphrased for "All models are wrong, some are useful."[69] In 2019, we might amend that statement to say, "All models are wrong, some are useful, some might be dangerous." We need to make clear to patients, providers, and healthcare leaders that predictive models will sometimes misclassify risk and that unintended consequences will result. Systems that permit transparency to this fact and to the factors that contribute to prediction are critical. At minimum, appropriate uncertainty quantification or calibration methods should deliver predictions that quantify risk in an actionable manner while accounting for changes in outcome prevalence, input data, and their relationships (a process known as "drift") over time.[59,70]

At the technology-level, systems should be designed to elicit providers' trust. Zuboff found that trust in a new technology depends on trial-and-error experience, followed by understanding of the technology's operation, and finally, faith.[71] To foster users' trust, users must have channels to disagree with recommendations such that algorithms (and their designers) can learn from these disagreements. Providers should not fear negative consequences for trusting their own clinical judgment—a safe, collaborative culture remains a key element to achieve this end. We recommend decision support systems permit users to share elements of their decision-making not included in algorithmic design or those elements that explain why users do not follow decision support recommendations.

Leverage determinants to address algorithmic bias

Behavioral and mental health conditions correlate with social determinants of health (eg, employment status; lesbian, gay, bisexual, transgender, and queer (LGBTQ) identification; and marital status), which may only be recorded in notes.[72–75] Unsupervised NLP methods can identify "homelessness" and "adverse childhood events" at scale in clinical text,[72] but few centers are able to integrate it into care delivery. Scalable NLP to make unstructured clinical text as readily available as structured diagnostic codes is needed to further catalyze behavioral health informatics research and operations. It would concomitantly increase capture of critical biopsychosocial determinants of health.

Performance of an algorithm may pass general thresholds set to be used in practice for alerts and risk assessment but may perform poorly for specific demographic segments. If so, care must be exercised when using it. Collecting additional data from these populations for retraining models might be an effective means to build fairer models without sacrificing accuracy.[71] Care should be taken to ensure that models do not discriminate in risk assessments regardless of demographic segment size or prevalence rates. We recommend providing guidance to interpretability or analytic similarity at the time of clinical decision-making to make transparent how similar a particular patient's demographics might be to the algorithmic training cohort, akin to efforts to display similarity of clinical trial participants to our communities using census health indicator data.[76]

Encourage interdisciplinary collaborations

AI in healthcare perches near the peak of the hype cycle. To speed its descent through the trough of disillusionment to the plateau of productivity, we must partner across disciplines. Unprecedented willingness to combine expertise in informatics, psychology, psychiatry, healthcare delivery, engineering, and more, have stimulated excitement around AI in healthcare and in other aspects of our lives. However, to avoid predictive models that never reach implementation or clinical use, clinical processes should be linked to the nascent stages of model development. Retrospective validation remains an accepted initial step in AI development and often relies on relatively accessible resources to complete. Transitioning to prospective use in clinical practice requires different study design, for example, pragmatic clinical trial and ongoing evaluation, as well as ongoing commitment from clinical and operational partners.

To improve interpretability, models should also be able to identify attributes in the patient records that have contributed to predictions or recommendations generated.[33] A focus on actionable, modifiable risk factors will convert prognostic models to predictive models that not only suggest risk of a future event but also those risks based on potential interventions that might be made right now.[77] For example, a predictive model might direct provider behavior in measurable and impactful ways if it suggests that reducing polypharmacy might lower downstream risk of an adverse drug event by a 10%, not just the presence of polypharmacy.

Augment the human elements

Another significant consideration relates to roles best played by humans and those by machines. Though it applies to any use of technology in healthcare, behavioral health in particular involves delicacy of interactions during times of crisis.

Since the advent ELIZA, it has been well-established that patients report symptoms more readily to digital or intelligent agents, such as chat-bots, than to humans, especially for behavioral health concerns.[25,26] On the other hand, interacting with vulnerable individuals requires skill and sensitivity not generally attributed to computers, however sophisticated. Therefore, more recently the push has been toward integrating intelligent agents with human providers. For instance, the AI algorithm of Crisis Text Line, a national nonprofit, uses two text responses of users contemplating suicide to triage them to a live counselor.[26] However, these systems are rarely linked to EHRs and routine clinical care at medical centers, so the onus remains on patients to report and providers to ask about these exchanges at the next clinical encounter.

Based on this evidence, we make the following suggestions to address under-reporting and under-coding. When designed in the context of behavioral health, AI models provide an opportunity to address issues such as under-coding, for example, through mining relationships between various data categories (eg, labs, medications, and diagnosis).[78]

Access to mental health expertise and allocation of a precious resource—consultation from mental health specialists—remain major challenges in healthcare around the world.[79] In the short-term, AI approaches to allocate such resources optimally and to queue the appropriate next patients for consultation by busy providers are key steps to begin proving clinical efficacy of intelligent systems in this area.

We emphasize the need to improve ascertainment of both predictors and phenotypes not well captured in structured or objective measurements. For example, we again note the need to augment NLP at the point-of-care. For example, basic sentiment analysis of discharge notes alone improves prediction of suicidality.[36] A behavioral health crisis might not be explicitly coded at the time of billing but might be well-described in clinical text. Intelligent agents that analyze clinical text in production, even clinical messages in patient portals, might improve our ability to identify patients in times of need in the same way we receive alerts for a creatinine lab test that has dramatically increased to surveil for signs of acute renal failure.

An over-arching question in this domain asks whether we can trust intelligent counselors or whether they should always be human. Though diagnostic criteria are debated and findings subject to interpretation and negotiation, humans remain more likely to provide better, and more humane, outcomes for patients, at least for now.[80] The lack of clearly established markers or measurable objective markers for some conditions in behavioral health further complicates this matter and reminds us about the importance of human judgment. If human comfort and sympathetic touch is called for—and would a machine be able to tell?—it still would be up to the healthcare practitioner to provide it.

WHAT'S MISSING

Many key themes were out of scope for this work. Mobile health applications for communication, activity tracking, meditation, and much more transform daily life for millions and are increasingly used in large-scale data collection, including behavioral health conditions.[81] Telemedicine has waxed and waned over the past decades but has unprecedented purchase in healthcare today. Telepsychiatry and telepsychology are potentially potent care delivery mechanisms on their own and stand to be enhanced through appropriate use of predictive models. We touched on ethical and privacy concerns, which are developed more fully elsewhere but still need further attention.[82–87] Finally, a growing body of press and literature outline concerns around commercial use and public–private partnerships involving clinical data and in particular mental health data via the app ecosystem, data aggregation, and others.[88–91] We highlight this important area that remains in need of further inquiry and empirical research.

CONCLUSIONS

The issues above help shape AI's potential in healthcare. Though consequences may be starker in behavioral and mental health, they deserve attention for all areas of AI in medicine. Inattention to them contributes to the most common fate for published predictive models—they are rarely translated into clinical practice.[92] Some models are appropriately evaluated, tested operationally, and not deployed, but many never reach that point.

Because behavioral health poses key informatics challenges, our recommendations are intended to catalyze further discussion. We have achieved our current state of predictive technology in behavioral health through close collaboration across disciplines. Rigorous, prospective evaluation is necessary to ensure outcomes improve with minimum unintended consequences. We should address these challenges to protect and improve quality of life and to improve mental and behavioral health through these same means.

FUNDING

Authors' effort were partially supported by the following grants: under grant # W81XWH-10-2-0181 and R01 MH116269-01 (CGW), the National Institute of General Medical Sciences of the National Institutes of Health under grant #P20 GM103424-17 (PD); U.S. National Center for Advancing Translational Sciences via grant #UL1TR001998 (RK); National Science Foundation under grant #1838745 (VS).

AUTHOR CONTRIBUTIONS

CGW led planning and drafting of this work. VS initiated the work and formed the team. The remaining authors contributed equally to manuscript planning, drafting, and revision, and therefore are listed in alphabetical order. This work is a collaborative effort between the AMIA Ethical, Legal, and Social Issues Working Group and the Mental Health Informatics Working Group.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Turing AM. Computing machinery and intelligence. *Mind* 1950; LIX (236): 433–60.
2. Lesaffre E, Speybroeck N, Berkvens D. Bayes and diagnostic testing. *Vet Parasitol* 2007; 148 (1): 58–61.
3. Yerushalmy J. Reliability of chest radiography in the diagnosis of pulmonary lesions. *Am J Surg* 1955; 89 (1): 231–40.
4. Garland LH. Studies on the accuracy of diagnostic procedures. *Am J Roentgenol Radium Ther Nucl Med* 1959; 82 (1): 25–38.
5. Lusted LB, Ledley RS. Mathematical models in medical diagnosis. *J Med Educ* 1960; 35: 214–22.
6. Lusted LB. Application of computers in diagnosis. *Circ Res* 1962; 11: 599–606.
7. Senders JT, Arnaout O, Karhade AV, *et al*. Natural and artificial intelligence in neurosurgery: a systematic review. *Neurosurgery* 2018; 83 (2): 181–92.
8. Librenza-Garcia D, Kotzian BJ, Yang J, *et al*. The impact of machine learning techniques in the study of bipolar disorder: a systematic review. *Neurosci Biobehav Rev* 2017; 80: 538–54.
9. Rajpara SM, Botello AP, Townend J, *et al*. Systematic review of dermoscopy and digital dermoscopy/artificial intelligence for the diagnosis of melanoma. *Br J Dermatol* 2009; 161 (3): 591–604.
10. Wongkoblap A, Vadillo MA, Curcin V. Researching mental health disorders in the era of social media: systematic review. *J Med Internet Res* 2017; 19 (6): e228.
11. van den Heever M, Mittal A, Haydock M, *et al*. The use of intelligent database systems in acute pancreatitis–a systematic review. *Pancreatology* 2014; 14 (1): 9–16.
12. Kansagara D, Englander H, Salanitro A, *et al*. Risk prediction models for hospital readmission: a systematic review. *JAMA* 2011; 306 (15): 1688–98.
13. Ohnuma T, Uchino S. Prediction models and their external validation studies for mortality of patients with acute kidney injury: a systematic review. *PLoS One* 2017; 12 (1): e0169341.
14. Fahey M, Crayton E, Wolfe C, *et al*. Clinical prediction models for mortality and functional outcome following ischemic stroke: a systematic review and meta-analysis. *PloS One* 2018; 13 (1): e0185402.
15. Gravante G, Garcea G, Ong SL, *et al*. Prediction of mortality in acute pancreatitis: a systematic review of the published evidence. *Pancreatology* 2009; 9 (5): 601–14.
16. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* 2017; 124 (7): 962–9.
17. Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; 316 (22): 2402–10.
18. Jaja BNR, Cusimano MD, Etminan N, *et al*. Clinical prediction models for aneurysmal subarachnoid hemorrhage: a systematic review. *Neurocrit Care* 2013; 18 (1): 143–53.
19. Belsher BE, Smolenski DJ, Pruitt LD, *et al*. Prediction models for suicide attempts and deaths: a systematic review and simulation. *JAMA Psychiatry* 2019; 76 (6): 642–51.
20. Perlis RH, Iosifescu DV, Castro VM, *et al*. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med* 2012; 42 (1): 41–50.
21. Dallora AL, Eivazzadeh S, Mendes E, *et al*. Machine learning and microsimulation techniques on the prognosis of dementia: a systematic literature review. *PLoS One* 2017; 12 (6): e0179804.
22. Centers for Medicare and Medicaid Services (CMS), Substance Abuse and Mental Health Services Administration (SAMHSA). A Roadmap to Behavioral Health – A Guide to Using Mental Health and Substance Use Disorder Services. *CMS.gov Consumer Resources*:25.
23. *Mental Health by the Numbers | NAMI: National Alliance on Mental Illness*. https://www.nami.org/learn-more/mental-health-by-the-numbers. Accessed March 6, 2019.
24. National Institute on Drug Abuse. *Trends & Statistics*. 2017. https://www.drugabuse.gov/related-topics/trends-statistics. Accessed March 5, 2019.
25. Henderson C, Evans-Lacko S, Thornicroft G. Mental illness stigma, help seeking, and public health programs. *Am J Public Health* 2013; 103 (5): 777–80.
26. Corrigan PW, Kleinlein P. The Impact of Mental Illness Stigma. In P. W. Corrigan ed. On the stigma of mental illness: Practical strategies for research and social change. Washington, DC, US: American Psychological Association; 2005:11–44.
27. Kennedy-Hendricks A, Barry CL, Gollust SE, *et al*. Social stigma toward persons with prescription opioid use disorder: associations with public support for punitive and public health–oriented policies. *Psychiatr Serv* 2017; 68 (5): 462–9.
28. Rost K, Smith R, Matthews DB, *et al*. The deliberate misdiagnosis of major depression in primary care. *Arch Fam Med* 1994; 3 (4): 333–7.
29. Doktorchik C, Patten S, Eastwood C, *et al*. Validation of a case definition for depression in administrative data against primary chart data as a reference standard. *BMC Psychiatry* 2019; 19 (1): 9.
30. Anderson HD, Pace WD, Brandt E, *et al*. Monitoring suicidal patients in primary care using electronic health records. *J Am Board Fam Med* 2015; 28 (1): 65–71.
31. Roberts E, Ludman AJ, Dworzynski K, *et al*. The diagnostic accuracy of the natriuretic peptides in heart failure: systematic review and diagnostic meta-analysis in the acute care setting. *BMJ* 2015; 350: h910.

32. Giannakopoulos K, Hoffmann U, Ansari U, *et al*. The use of biomarkers in sepsis: a systematic review. *Curr Pharm Biotechnol* 2017; 18 (6): 499–507.
33. Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In Advances in Neural Information Processing Systems, 2016: 3504–3512.
34. Cook BL, Progovac AM, Chen P, *et al*. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Comput Math Methods Med* 2016; 2016: 1.
35. Calvo RA, Milne DN, Hussain MS, *et al*. Natural language processing in mental health applications using non-clinical texts. *Nat Lang Eng* 2017; 23 (5): 649–85.
36. McCoy TH, Castro VM, Roberson AM, *et al*. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. *JAMA Psychiatry* 2016; 73 (10): 1064–71.
37. Wang Y, Chen ES, Pakhomov S, *et al*. Automated extraction of substance use information from clinical texts. *AMIA Annu Symp Proc* 2015; 2015: 2121–30.
38. Carrell DS, Cronkite D, Palmer RE, *et al*. Using natural language processing to identify problem usage of prescription opioids. *Int J Med Inf* 2015; 84 (12): 1057–64.
39. Abbe A, Grouin C, Zweigenbaum P, *et al*. Text mining applications in psychiatry: a systematic literature review. *Int J Methods Psychiatr Res* 2016; 25 (2): 86–100.
40. Castro VM, Minnier J, Murphy SN, *et al*. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *Am J Psychiatry* 2015;172 (4): 363–72.
41. Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. Curran Associates Inc.; 2013: 3111–19; Lake Tahoe, CA. http://dl.acm.org/citation.cfm? id=2999792.2999959. Accessed July 22, 2019.
42. Mitchell, TM. The need for biases in learning generalizations. New Jersey: Department of Computer Science, Laboratory for Computer Science Research, Rutgers University; 1980.
43. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. *Neural Comput* 1992; 4 (1): 1–58.
44. Gordon DF, Desjardins M. Evaluation and selection of biases in machine learning. *Mach Learn* 1995; 20 (1–2): 5–22.
45. Baeza-Yates R. Data and algorithmic bias in the web. In: Proceedings of the 8th ACM Conference on Web Science. ACM; 2016: 1; Hannover, Germany.
46. Danks D, London AJ. Algorithmic bias in autonomous systems. In: IJCAI. 2017: 4691–97.
47. Garcia M. Racist in the machine: the disturbing implications of algorithmic bias. *World Policy J* 2016; 33 (4): 111–7.
48. Hajian S, Bonchi F, Castillo C. Algorithmic bias: from discrimination discovery to fairness-aware data mining. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM; 2016: 2125–26.
49. Lambrecht A, Tucker CE. Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. *Management Science*; 2019; 65 (7): 2947–3448.
50. Schwartz RC, Blankenship DM. Racial disparities in psychotic disorder diagnosis: a review of empirical literature. *World J Psychiatry* 2014; 4 (4): 133–40.
51. Lehavot K, Katon JG, Chen JA, *et al*. Post-traumatic stress disorder by gender and veteran status. *Am J Prev Med* 2018; 54 (1): e1–9.
52. Leavitt RA, Ertl A, Sheats K, *et al.* Suicides Among American Indian/Alaska Natives — National Violent Death Reporting System, 18 States, 2003–2014. *MMWR Morb Mortal Wkly Rep* 2018; 67: 237–42.
53. Leong FTL, Kalibatseva Z. Cross-cultural barriers to mental health services in the United States. *Cerebrum Dana Forum Brain Sci* 2011; 5: 11. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3574791/. Accessed March 5, 2019.
54. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics* 2019; 21: 167–79.
55. Ferryman K, Pitcan M. Fairness in precision medicine. Data & Society Research Institute, 2018; 54.
56. Caplan R, Donovan J, Hanson L, *et al*. *Algorithmic Accountability: A Primer*. Data Society. Washington, DC. https://datasociety.net/output/algorithmic-accountability-a-primer/. Accessed March 5, 2019.
57. Osoba OA, Welser WI. *An Intelligence in Our Image*. 2017. https://www.rand.org/pubs/research_reports/RR1744.html. Accessed March 5, 2019.
58. Lenert MC, Walsh CG. Balancing performance and interpretability: selecting features with bootstrapped ridge regression. *AMIA Annu Symp Proc AMIA Symp* 2018; 2018: 1377–86.
59. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. New York, NY: ACM; 2005: 625–32.
60. Huys QJM, Maia TV, Frank MJ. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci* 2016; 19 (3): 404–13.
61. Doran D, Schulz S, Besold TR. What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv: 1710.00794; 2017.
62. Gunning D, Aha D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 2019; 40 (2): 44–58.
63. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain?. *arXiv preprint arXiv*:1712.09923; 2017.
64. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* 2019; 51 (5): 93
65. Osheroff, J, Teich J, Levick D, *et al*. Improving outcomes with clinical decision support: an implementer's guide. Chicago, IL: HIMSS Publishing; 2012.
66. Carter G, Milner A, McGill K, *et al*. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. *Br J Psychiatry* 2017; 210 (6): 387–95.
67. Oruch R, Pryme IF, Engelsen BA, *et al*. Neuroleptic malignant syndrome: an easily overlooked neurologic emergency. *Neuropsychiatr Dis Treat* 2017; 13: 161–75.
68. McKernan LC, Lenert MC, Crofford LJ, Walsh CG. Outpatient Engagement and Predicted Risk of Suicide Attempts in Fibromyalgia. *Arthritis Care Res (Hoboken)*. 2019;71 (9):1255–1263.
69. Box G. Science and statistics. *J Am Stat Assoc* 1976; 71 (356): 791–9.
70. Davis SE, Lasko TA, Chen G, *et al*. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc* 2017; 24 (6): 1052–61.
71. Chen I, Johansson FD, Sontag D. Why is my classifier discriminatory? In: Bengio S, Wallach H, Larochelle H, *et al*., eds. Advances in Neural Information Processing Systems 31. Curran Associates, Inc.; 2018: 3539–50; Montreal, Canada. http://papers.nips.cc/paper/7613-why-is-my-classifier-discriminatory.pdf. Accessed March 5, 2019.
72. Bejan CA, Angiolillo J, Conway D, *et al*. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc* 2018; 25 (1): 61–71.
73. Allen J, Balfour R, Bell R, *et al*. Social determinants of mental health. *Int Rev Psychiatry* 2014; 26 (4): 392–407.
74. Feller DJ, Zucker J, Don't Walk OB, *et al*. Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. *AMIA Annu Symp Proc* 2018; 2018: 422–9.
75. Dillahunt-Aspillaga C, Finch D, Massengale J, *et al*. Using information from the electronic health record to improve measurement of unemployment in service members and veterans with mTBI and post-deployment stress. *PLoS One* 2014; 9 (12): e115873.

76. Lenert MC, Mize DE, Walsh CG. X Marks the spot: mapping similarity between clinical trial cohorts and US counties. *AMIA Annu Symp Proc* 2017; 2017: 1110–9.
77. Wong HR, Cvijanovich NZ, Anas N, *et al*. Endotype transitions during the acute phase of pediatric septic shock reflect changing risk and treatment response. *Crit Care Med* 2018; 46 (3): e242.
78. Weaver C, Garies S, Williamson T, *et al*. Association rule mining to identify potential under-coding of conditions in the problem list in primary care electronic medical records. *Int J Popul Data Sci* 2018; 3(4) *Conference Proceedings for IPDLC 2018* doi:10.23889/ijpds.v3i4.622.
79. WHO. *Mental Health: Massive Scale-Up of Resources Needed if Global Targets are to be Met*. Geneva, Switzerland: WHO. http://www.who.int/mental_health/evidence/atlas/atlas_2017_web_note/en/. Accessed July 26, 2019.
80. Goodman KW, Cushman R, Miller RA. Ethics in Biomedical and Health Informatics: Users, Standards, and Outcomes. In: Shortliffe EH and Cimino JJ, eds. Biomedical Informatics—Computer Applications in Health Care and Biomedicine. New York: Springer; 2014:329–53.
81. Schmitz H, Howe CL, Armstrong DG, *et al*. Leveraging mobile health applications for biomedical research and citizen science: a scoping review. *J Am Med Inform Assoc* 2018; 25 (12): 1685–95.
82. Kaplan B. Selling health data: de-identification, privacy, and speech. *Camb Q Healthc Ethics* 2015; 24 (3): 256–71.
83. Kaplan B. How should health data be used? *Camb Q Healthc Ethics* 2016; 25 (2): 312–29.
84. Kaplan B, Litewka S. Ethical challenges of telemedicine and telehealth. *Camb Q Healthc Ethics* 2008; 17 (4): 401–16.
85. Kaplan B, Ranchordás S. Alzheimer's and m-Health: Regulatory, Privacy, and Ethical Considerations. In: Hayre CM, Muller DJ, and Scherer MH, eds. Everyday Technologies in Healthcare. 1 ed. Boca Raton, FL: CRC Press, 2019:31–52.
86. McKernan LC, Clayton EW, Walsh CG. Protecting life while preserving liberty: ethical recommendations for suicide prevention with artificial intelligence. *Front Psychiatry*, 2018; 9: 650.
87. Tucker RP, Tackett MJ, Glickman D, *et al*. Ethical and Practical Considerations in the Use of a Predictive Model to Trigger Suicide Prevention Interventions in Healthcare Settings. *Suicide Life Threat Behav*. 2019; 49 (2): 382–92.
88. The Privacy Project. N. Y. Times. 2019.https://www.nytimes.com/series/new-york-times-privacy-project. Accessed September 25, 2019.
89. Mental Health Apps are Scooping up Sensitive Data. Will you Benefit? STAT. 2019. https://www.statnews.com/2019/09/20/mental-health-apps-capture-sensitive-data/. Accessed September 25, 2019.
90. *Artificial Intelligence Can Complicate Finding the Right Therapist*. STAT. 2019. https://www.statnews.com/2019/09/20/artificial-intelligence-tool-finding-mental-health-therapist/. Accessed September 25, 2019.
91. Zuboff S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. 1 ed. New York: PublicAffairs; 2019.
92. Grossman L, Reeder R, Walsh CG, *et al*. Improving and implementing models to predict 30-day hospital readmissions: a survey of leading researchers. In: Academy Health. New Orleans, LA; 2017. https://academyhealth.confex.com/academyhealth/2017arm/meetingapp.cgi/Paper/17304.