Schizophrenia Bulletin vol. 46 no. 1 pp. 11–14, 2020
doi:10.1093/schbul/sbz105
Advance Access publication 1 November 2019

AT ISSUE

Using Machine Learning in Psychiatry: The Need to Establish a Framework That Nurtures Trustworthiness

Chelsea Chandler*[,1], Peter W. Foltz[2,3], and Brita Elvevåg[4,5]

[1]Department of Computer Science, University of Colorado Boulder, 430 UCB, 1111 Engineering Drive, Boulder, CO 80309; [2]Institute of Cognitive Science, University of Colorado Boulder; [3]Pearson PLC, London, UK; [4]Department of Clinical Medicine, University of Tromsø, Tromsø, Norway; [5]Norwegian Centre for eHealth Research, Tromsø, Norway

*To whom correspondence should be addressed; tel: 703-895-4764, fax: 303-492-7177, e-mail: chelsea.chandler@colorado.edu

Abstract The rapid embracing of artificial intelligence in psychiatry has a flavor of being the current "wild west"; a multidisciplinary approach that is very technical and complex, yet seems to produce findings that resonate. These studies are hard to review as the methods are often opaque and it is tricky to find the suitable combination of reviewers. This issue will only get more complex in the absence of a rigorous framework to evaluate such studies and thus nurture trustworthiness. Therefore, our paper discusses the urgency of the field to develop a framework with which to evaluate the complex methodology such that the process is done honestly, fairly, scientifically, and accurately. However, evaluation is a complicated process and so we focus on three issues, namely explainability, transparency, and generalizability, that are critical for establishing the viability of using artificial intelligence in psychiatry. We discuss how defining these three issues helps towards building a framework to ensure trustworthiness, but show how difficult definition can be, as the terms have different meanings in medicine, computer science, and law. We conclude that it is important to start the discussion such that there can be a call for policy on this and that the community takes extra care when reviewing clinical applications of such models..

Key words: computational psychiatry/artificial intelligence/guidelines/explainability/transparency/generalizability

Introduction

There has been a recent surge in the popularity of researching artificial intelligence (AI), and more specifically machine learning (ML), in psychiatry. Automated analyses can provide supplementary information during clinical decision making and uncover subtle trends that humans have difficulty detecting. However, although results seem successful, there lacks understanding of the underlying mechanisms and how and when it should be applied. This can engender fear that automated methods will make decisions contrary to human judgments.[1,2] Indeed, the mystery around AI does not nurture trustworthiness, which is critical when applying medical technology.

The role of AI and ML in society is a source of highly charged debates in many disciplines, notably in education, policy, and medicine. Given that a growing number of psychiatry publications incorporate ML methods,[3–9] we seek to initiate a conversation on how best to ensure that these methods are valid, provide information that is clinically useful, and are implemented and used correctly. We need to start discussing what procedures must be in place to ensure research is of high caliber, results are evaluated thoroughly and fairly, and clinical usefulness is established before application. We discuss three concepts, (1) explainability, (2) transparency, and (3) generalizability, that are critical for establishing the viability of using AI, focusing on methods which could in the near-future be employed in psychiatry to facilitate in diagnosis, monitoring, evaluation, and prognosis of illness.[3–9]

How Well Can We Trust the Empirical Basis That These Clinical Applications of AI Are Based Upon?

Some argue that for a ML model (see table 1 for definition) to be *trustworthy*, it must be *explainable*, meaning it must be possible to obtain a description of the reason a model arrived at a decision. In medicine this has never been a logical requirement. For example, the exact mechanisms of statins to lower cholesterol or neuroleptic medications

Table 1. Definition of Terms Used in This Article

Artificial Intelligence (AI)	AI is the general concept of creating expert systems to carry out various types of tasks. These systems exhibit intelligent behavior and are able to learn, explain, and advise their users.
Machine Learning (ML)	ML is a subset of AI. ML models are statistical systems that either learn *features* of the data and associated importance of each, that are highly predictive or some variable of interest, or just the associated importance of user-defined features of the data. Once the features and their weights, as well as other *hyperparameters,* are set, the model can predict some outcome or clinical classification on new, unseen data.
Features	A measurable property of the data. For example, the number of words spoken is a feature that can be measured from natural language data.
Hyperparameters	Parameters that are set before training rather than learned during the process, like the number of iterations through a training set to train for.
Classification	A type of ML model that is trained to predict a category (eg, mentally ill or healthy). A common type of classification model is a decision tree, which is a flow chart-like structure that breaks an entire dataset into subsets at each level depending on some criteria. At the bottom of the structure, the entire dataset is broken into some number of subsets which represent categories of interest in a classification problem.
Regression	A type of ML model that is trained to predict a numerical, continuous valued output (eg, a rating). A common type of regression model is a linear/polynomial regression model which simply generates a linear/polynomial line to fit values in a dataset. These equations can be made into classification problems by setting threshold(s) to bin the real valued output variables into discrete categories.
Neural Network	A system of nodes, composed in layers, where each node learns some nonlinear equation on some subset of training data and when all nodes are combined, a categorical or real valued output can be computed. Modern neural networks are deep, meaning they have hundreds to thousands of nodes and layers and are trained on large datasets.

to reduce psychiatric symptoms are not necessarily understood but are nonetheless administered to patients. However, even if truly explainable models are unattainable, they can be *transparent* in that the developer must be open about the model details, as well as *generalizable* so the model performs well on a diverse population.

Explainability

There are many recent high-profile examples of seemingly successful outcomes with AI (eg, Google AI's lung cancer screening model that performed on par with 6 radiologists[10]), as well as failures (eg, IBM Watson's cancer treatment decision where the model ignored contraindication information[11]). Both success and failure force us to consider who is ultimately responsible, how to minimize errors in the future, and how best to continue the path of AI assistance. These questions require that the conclusion arrived at by the AI be explained in terms of its decision-making process.

What Does It Mean for a Model to Be Explainable? This question will have different answers depending upon whether the person who answers is a computer scientist, clinician, or lawyer, for example. To a computer scientist, explainability might be where the feature space can be divided to create distinct classes, how the model weights different features or what a prototypical example from a class is. To a clinician, explainability might refer to whether the model produces a particular output that can be tied to clinical constructs (eg, thought disorder). To a lawyer, explainability refers to a legal justification of a decision and the rights of citizens for an explanation of an algorithmic output (eg, why a specific treatment is given).[12]

What Constitutes a Sufficient Explanation? For clinicians, a model that provides a good explanation might relate understandable features of behavior (eg, lack of vocal modulation across changing state in psychosis) to clinical constructs (an indicator of a worsening clinical state[13]). These constructs can be defined as the phenomenology or symptoms as determined by DSM-V.[14] However, ML-based modeling often does not work at this feature level, rather is more comparable to the fundamental and granular level characterization advocated by the NIMH Research Domain Criteria[15] (eg, frequency counts, word choice, time of task completion). In addition to the granularity of features, models often combine many features (sometimes hundreds or thousands) and learn how to weight these to best explain the clinical construct.

While it may sometimes not align directly with clinical constructs, it is nonetheless critical to understand the classes of features the model is using to make a prediction as models could be working for spurious reasons[16] which in medicine can have terrible consequences, as clinical decisions could be based upon a flawed foundation.

An explanation would vary by application domain and ML model type. Explanations from classification models could require a counterfactual example to specify what would need to be different about a specific case in order for the model to give an alternate label. On the other hand, regression models lend themselves to explanations based on specific features and weights, or importance, assigned to each one. With this information, one could say that with more or less of a specific variable, we would have received a different score. These variables may not be causal factors, but rather simply associated with one another.[17,18] In both cases, knowing the statistics of the

data the models were trained on, as well as distributions and probabilities assigned to features and classes, is critically important.

There is yet another distinction to be made in model explainability, namely that between highly interpretable models (eg, decision trees or linear regression/classification models) and uninterpretable models (eg, complex, deep neural networks). An interpretable model involves few features such that the model learns weights for each one. For example, a model of memory recall could learn a weight of 2.0 for the feature *number of words recalled* and 3.0 for *semantic similarity of a recall to the original story*, indicating that the more words recalled and the greater the semantic similarity, the larger the recall score. These simple models can provide rich explanations, that more complex, deep neural networks cannot, often at the cost of accuracy. This is the well-known tradeoff between model performance and model explainability.[19] The most successful ML models represent information with thousands to millions of features which do not lend themselves to explainability. While recent advances in ML do allow us to peer into the black box, the view may still be at a high level. While we believe in striving toward explainability, a more realistic goal is transparency and generalizability.

Transparency

For a ML model to be transparent, its purpose, how its architecture was chosen, the statistics of the training data, how it was trained, and any underlying assumptions made in the process must be clear. Furthermore, there should be a level of transparency on how the features used in the model align to the constructs of interest being assessed.

The above criteria are necessary to foster trust in ML models. A model advertised as being a money-saver could generate distrust, while a model that is open about its procedures and assumptions and is advertised as accurately predicting a variable of interest would engender more trust. Understanding common assumptions (eg, the data are independent and identically distributed, cross validation is used in training, or model choice [eg, choosing a linear regression classifier implies that the data are believed to be linearly separable]) can give the community more insight into model behavior.

We, as members of the scientific community, should strive for a ML model's output to support a clinician's decision-making process rather than be a final result that is subsequently questioned. In addition to being transparent about the intricacies of the model, it is also important to include clinical guidelines for use. A clinician *must* be informed of the details of the system and how best to and when to use it. We do not imply that the code must be open source as this is unrealistic in many cases, but rather that high-level methods and details are available.

Generalizability

Generalizability is the foundation of good science and its progress. Complex ML approaches require large datasets because they otherwise tend to overgeneralize. It is regretful that in psychiatry many studies are published with modest and inadequate sample sizes (eg, less than 50 training samples and generalization sets of less than 20), further compounded by a lack of standard cross validation practices.[9] It *must* be noted that the large datasets that have become standard in ML render it highly unlikely to achieve a "matched" dataset in the classic clinical sense. While the datasets are unlikely to be matched, the advantage of ML is that it is able to leverage the variability inherent in large data to provide valid characterizations.

Since ML algorithms learn from the samples of data upon which they are trained, they mirror historical forms of racial, economic, and gender discrimination primarily because of the biases and prejudices present in the data.[20–23] To guard from issues in bias and generalizability, the scientific community must work to develop large and trusted datasets that are representative of target populations in order to test algorithms. This would ensure less bias against particular groups, as well as a standard that every model must meet to be trusted for real world use. It is critical that databases comprise representative samples from the full spectrum of all intended stakeholders. This will increase the accuracy of assaying the appropriate measurement constructs and thereby increase the probability of early and accurate detection and diagnosis. Building the next generation of tools that leverage ML necessitates that the underlying methods are valid across genders, ages, ethnicity, and sociodemographics as well as provide reliable measurements within the same individual over time.

Furthermore, it is necessary to exhaustively test a ML system before it is deployed in a clinical setting, which is a general rule for any product of computer science. In testing, there exist notions of *edge cases* and *corner cases* that must be considered. The former is an example that occurs at the extremes of acceptable operating parameters (eg, minimum and maximum values of features) and the latter is an example that occurs outside of expected operating parameters (eg, the combination of features at their outer limits); the importance being that a model must be able to perform on the most extreme clinical cases. Attempting to "break" a ML model is the best way to ensure it will perform as expected in all situations.

Discussion

The aforementioned framework is needed to address common issues with the integration of ML into psychiatry. However, there are other long-term problems that need to be accounted for. Since AI tools work most accurately with large datasets, there is a need to share in a

consortium like fashion, but how is this possible with sensitive data? The use of state of the art, publicly available tools is impossible with identifiable patient data (and even if the Personally Identifiable Information is removed from the data, various sources can be triangulated to retain identification), so how do we work around this? Finally, who is to be held accountable for errors? Does accountability fall onto the engineer who created the model, the entity that reviewed and deemed it safe for use, or ultimately the clinician who used it in practice? These are the types of questions that warrant discussion across all disciplines.

Conclusion

We are still at the research stage of evaluating ML in psychiatry, but we need a principled way to translate this research to clinical application. When the accuracy gains of ML models inevitably plateau, it will be important to shift our focus to making these models as robust as possible to be applied in the real world. Rather than looking for ML models to become the ultimate decision-maker in medicine, we should leverage the things that machines do well that are distinct from what humans do well. We have presented a framework applicable for reviewers of research, as well as clinicians looking to apply ML to their practice. This framework provides a basis for the need to evaluate the explainability, assess the transparency, and ensure the generalizability of ML models.

References

1. Friend T. How frightened should we be of A.I.? *The New Yorker.* May 14, 2018 issue.
2. Tran V, Riveros C, Ravaud P. Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. *npj Digital Med.* 2019;2. doi:10.1038/s41746-019-0132-y.
3. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry.* 2018;3:223–230.
4. Bedi G, Carrillo F, Cecchi GA, et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophr.* 2015;1:15030. doi:10.1038/npjschz.2015.30.
5. Rezaii N, Walker E, Wolff P. A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophr.* 2019;5. doi:10.1038/s41537-019-0077-9.
6. Elvevåg B, Foltz PW, Rosenstein M, et al. Thoughts about disordered thinking: measuring and quantifying the laws of order and disorder. *Schizophr Bull.* 2017;43:509–513.
7. Tandon N, Rajiv T. Will machine learning enable us to finally cut the Gordian Knot of schizophrenia. *Schizophr Bull.* 2018;44:939–941.
8. Corcoran CM, Carrillo F, Fernández-Slezak D, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry.* 2018;1:67–75. doi:10.1002/wps.20491.
9. Foltz PW, Rosenstein M, Elvevåg B. Detecting clinically significant events through automated language analysis: Quo imus? *npj Schizophr.* 2016;2:15054. doi:10.1038/npjschz.2015.54.
10. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med.* 2019;25:954–961. doi:0.1038/s41591-019-0447-x.
11. Ross C, Swetlitz I. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *In STAT News.* 2018. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/. Accessed September 30, 2019.
12. European Commission. 2018 reform of EU data protection rules. https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en. Accessed September 30, 2019.
13. Cohen AS, Fedechko TL, Schwartz EK, et al. Ambulatory vocal acoustics, temporal dynamics and serious mental illness. *J Abnorm Psychol.* 2019;128:97–105. doi:10.1037/abn0000397.
14. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. Arlington, VA: American Psychiatric Publishing; 2013.
15. Insel T, Cuthbert B, Garvey M, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry.* 2010;167:748–751.
16. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, 2016, 1135-1144.
17. Pearl J, Mackenzie D. *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books, Inc.; 2018.
18. Mehler D, Anton M, Kording KP. The lure of causal statements: rampant mis-inference of causality in estimated connectivity. arXiv preprint arXiv:1812.03363. 2018.
19. Bzdok D, Nichols TE, Smith, SM. Towards algorithmic analytics for large-scale datasets. *Nat Mach Intell*;1:296–306.
20. Henrich J, Heine S, Norenzayan A. The weirdest people in the world? *Behav Brain Sci.* 2010;33:61–83.
21. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.* 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. Accessed September 30, 2019.
22. Bolukbasi T, Chang K, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain; 2016;4349–4357.
23. O'Neil C. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. New York: Crown; 2016.