Psychological Medicine

cambridge.org/psm

Computing schizophrenia: ethical challenges for machine learning in psychiatry

Georg Starke[1], Eva De Clercq[1], Stefan Borgwardt[2,3] and Bernice Simone Elger[1,4]

[1]Institute for Biomedical Ethics, University of Basel, Basel, Switzerland; [2]Department of Psychiatry, University of Basel, Basel, Switzerland; [3]Department of Psychiatry and Psychotherapy, University of Lübeck, Lübeck, Germany and [4]University Center of Legal Medicine, University of Geneva, Geneva, Switzerland

Editorial

Cite this article: Starke G, De Clercq E, Borgwardt S, Elger BS (2021). Computing schizophrenia: ethical challenges for machine learning in psychiatry. *Psychological Medicine* **51**, 2515–2521. https://doi.org/10.1017/S0033291720001683

Received: 12 July 2019
Revised: 26 January 2020
Accepted: 13 May 2020
First published online: 15 June 2020

Key words:
Artificial intelligence; bioethics; case studies; computational psychiatry; psychosis; research ethics

Author for correspondence:
Georg Starke, E-mail: georg.starke@unibas.ch

Abstract

Recent advances in machine learning (ML) promise far-reaching improvements across medical care, not least within psychiatry. While to date no psychiatric application of ML constitutes standard clinical practice, it seems crucial to get ahead of these developments and address their ethical challenges early on. Following a short general introduction concerning ML in psychiatry, we do so by focusing on schizophrenia as a paradigmatic case. Based on recent research employing ML to further the diagnosis, treatment, and prediction of schizophrenia, we discuss three hypothetical case studies of ML applications with view to their ethical dimensions. Throughout this discussion, we follow the principlist framework by Tom Beauchamp and James Childress to analyse potential problems in detail. In particular, we structure our analysis around their principles of beneficence, non-maleficence, respect for autonomy, and justice. We conclude with a call for cautious optimism concerning the implementation of ML in psychiatry if close attention is paid to the particular intricacies of psychiatric disorders and its success evaluated based on tangible clinical benefit for patients.

Introduction

The quest for objective measures of mental disorders has been a long-standing ambition of psychiatry (Kapur, Phillips, & Insel, 2012; Singh & Rose, 2009). Given the notorious difficulties of classifying mental disorders and the challenge of establishing psychiatric biomarkers, many recent advances put their hope in approaches using machine learning (ML) as a paradigm-shifting way forward (Bzdok & Meyer-Lindenberg, 2018; Janssen, Mourao-Miranda, & Schnack, 2018; Shatte, Hutchinson, & Teague, 2019). By applying ML on large-scale datasets, it seems feasible to distinguish between healthy controls and patients diagnosed with major depressive disorder or schizophrenia on an individual level – although reported diagnostic accuracies differ largely across studies (Ebdrup et al., 2019; Gao, Calhoun, & Sui, 2018; Kambeitz et al., 2015). Furthermore, ML techniques can differentiate successfully between subgroups within psychiatric categories (Drysdale et al., 2017; Dwyer et al., 2018) and predict the success of specific psychopharmacological interventions for single subjects (Chekroud et al., 2016; Webb et al., 2018). Of high clinical interest are ML applications that provide robust probabilistic estimates regarding the future onset of psychosis (Borgwardt et al., 2013; Chung et al., 2018; Koutsouleris et al., 2018) or the risk of suicide (Franklin et al., 2017; Just et al., 2017; Walsh, Ribeiro, & Franklin, 2017). However, to allow translation to current clinical practice, further multicenter imaging studies, integrating clinical measures and multivariate imaging data, are needed to replicate promising initial findings (Giordano & Borgwardt, 2019).

Currently, there is no established ML application in psychiatric clinical practice. The drastic increase of FDA approvals for medical applications of artificial intelligence (AI) in the past 2 years (Topol, 2019) suggests that some ML programs could soon be integrated into standard clinical care, improving prediction and early detection, diagnostic certainty, and individual treatment outcome in the sense of personalized psychiatry (Perna, Grassi, Caldirola, & Nemeroff, 2018). Unfortunately, the majority of ML applications in psychiatry still lack in-depth ethical analysis. With few exceptions discussing specific case studies (Martinez-Martin, Dunn, & Roberts, 2018), ethical concerns are often voiced in a general form (Char, Shah, & Magnus, 2018; Topol, 2019; Vayena, Blasimme, & Cohen, 2018), thus necessarily neglecting the particular intricacies of potential psychiatric applications.

ML is an extremely broad term, covering many distinct computational approaches for even more heterogeneous real-world problems. We aim to demonstrate that any categorical rejection of the use of ML in psychiatry would be ethically wrong given its potential benefits but that careful evaluation is needed whether a particular procedure improves clinical care or merely constitutes a nifty computational exercise. Using schizophrenia as a paradigmatic

https://doi.org/10.1017/S0033291720001683 Published online by Cambridge University Press

Table 1. Supervised, unsupervised, and reinforced ML

ML type	Required data	Typical problem	Exemplary application in schizophrenia
Unsupervised	Unlabeled training data	Clustering	Refine diagnostic criteria (case 1)
Supervised	Labeled training data	Classification and regression	Improve diagnostic accuracy (case 2)
Reinforced	Labeled and unlabeled data	Dynamic decision-making	Suggest optimal treatment regime (case 3)

case, we will first sketch some fundamental distinctions of different ML methods, before turning to three (hypothetical) case studies. To support our main claim, we will discuss these cases following the principlist framework of Beauchamp and Childress (2013), which has recently been embraced as providing suitable principles for the ethical use of AI as well (AI HLEG, 2019; Floridi et al., 2018).

Machine learning in psychiatry

The meaning of the term ‘machine learning’ is often ambiguous. In the present paper, we use ML to describe learning algorithms which improve their performance in a certain task based on prior computation (Iniesta, Stahl, & McGuffin, 2016; Mitchell, 1997). ML in this sense comprises a narrower field than AI, which includes generalized AI and incidentally describes ‘whatever hasn't been done yet’ (Hofstadter, 1980, p. 601). At the same time, ML itself entails many specific computational approaches, from deep learning (DL) using artificial neural networks to algorithms relying on support vector machines. Across the many different methods of ML, a common distinction is drawn between three types: supervised, unsupervised, and reinforced learning.

Typical tasks performed by supervised learning are problems of discriminative classification where the ML algorithm assigns a probability of belonging to a certain category *Y* based on feature *X*. To do so, supervised learning requires labeled training data, matching the training instances to labels such as ‘diseased’–‘healthy’, ‘developed psychosis’–‘did not develop psychosis’, or ‘positive treatment outcome’–‘negative treatment outcome’. After training, the ML algorithm can then assign these labels correctly to new data. Unsupervised learning, on the other hand, does not require labeled training data. Instead, it can make use of often more readily available, unlabeled data, such as whole-genome sequences or cell phone metadata, to find clusters within these data points. In real-life settings, applications may fall between these two approaches and are described as ‘semi-supervised’ or, as recently suggested by LeCun, as ‘self-supervised’ (LeCun, 2018), complementing labeled training data with large bits of unlabeled data (Chapelle, Schölkopf, & Zien, 2010). Finally, reinforcement learning denotes ML programs that optimize their interaction with an environment by trying to maximize reward over time (Mnih et al., 2015). While this approach, inspired by neuroscientific accounts of learning, does not require fully labeled data, it needs some formalization of rewards, e.g. winning an ATARI game.

The schematic distinction of these three general ML types can also be instructive for ethical debate of applied ML in psychiatry. For as we will show, differences in methodology do not only have a big impact on feasibility since labelling of data often requires cost- and labor-intensive efforts but may also account for important ethical implications (Table 1).

Before turning to the potential of ML techniques to improve clinical care, some methodological limitations of psychiatric ML need to be mentioned, recently stressed by Vieira et al. (2020). Some of these concerns, such as small sample size or publication bias, are pervasive across different research areas and neuroscientific research in particular (Button et al., 2013; Kellmeyer, 2017; Schnack & Kahn, 2016). Other methodological issues arise with specific regard to ML, e.g. regarding failure to rigorously employ nested cross-validation, testing the predictions of an ML program on a fully independent sample (Stahl & Pickles, 2018). In addition, psychiatry's high-dimensional and often noisy data demand particular consideration and may hinder adopting computational strategies popular in other medical areas. While DL is frequently considered the method of choice for medical image analysis (Shen, Wu, & Suk, 2017), some recent results suggest that for imaging-based predictions of cognitive and behavioral measures, classical kernel regression is at least as successful as DL (He et al., 2020; Mihalik et al., 2019), rendering a linear and more interpretable approach (Heinrichs & Eickhoff, 2020) potentially preferable. These methodological challenges may partially account for inconsistent results across different studies, e.g. reporting largely variable accuracies for potential biomarkers of schizophrenia based on ML and neuroimaging (Kambeitz et al., 2015).

The potentially deepest challenge for implementing ML in psychiatry lies in its long-embattled nosology though (Kendler, 2016; Kendler, Zachar, & Craver, 2011; Zachar, 2015), calling into question the choice of appropriate data for training. Given that psychiatry arguably still lacks a successful diagnostic scheme that is valid and reliable (Barron, 2019), establishing psychiatric ML programs relies on a shaky ground truth. This problem is exacerbated by fundamental concerns whether a reductionist framework, considering psychiatric disorders as mere brain diseases to be investigated with neuroimaging and genetics, is convincing (Borsboom, Cramer, & Kalis, 2019). While we largely focus on neuroimaging studies in our examples for the sake of simplicity, research should thus be careful to not restrain their input *a priori* to biological data but also include social and idiosyncratic information on individual patients. Using natural language processing on narrative electronic health records could provide a starting point for such an endeavor (Rumshisky et al., 2016).

Applications of ML for schizophrenia

Future ML applications for patients with schizophrenia may differ largely. For research purposes, using unsupervised learning to identify altered brain structures in patients with schizophrenia is common. In some of these possible approaches, which have been described as data- or discovery-oriented (Huys, Maia, & Frank, 2016; Krystal et al., 2017), the algorithm is provided with neuroimaging data of patients with schizophrenia and left to find clusters (Dwyer et al., 2018; Schnack, 2019). Hence, apart from sample choice, little human labeling determines the data. Instead, the algorithm is left to find clusters that may or may not map onto a given hypothesis and can, in some cases, correlate with clinical data. Indeed, given the manifold disputes over

https://doi.org/10.1017/S0033291720001683 Published online by Cambridge University Press

Table 2. Case vignettes

Three potential applications for ML in schizophrenia
Case 1: R is presenting with newly developed negative and positive symptoms at a university psychiatry department. Based on a clinical interview, R is diagnosed with schizophrenia by a psychiatrist. As part of a research program that aims to distinguish amongst schizophrenia subtypes, Z undergoes structural cranial magnetic resonance imaging (MRI) scanning which is analyzed by an ML algorithm trained to find commonalities and differences of brain volume in specific cortical areas across all brain scans acquired from first-episode patients with schizophrenia presenting to the university hospital. Based on his brain scan, R is assigned to a subtype of schizophrenia with a typical pattern of superior-temporal grey matter loss.
Case 2: D is presenting at a psychiatric day-clinic with mild psychotic symptoms and is diagnosed with schizophrenia after a clinical interview. Given her markedly depressed mood and further reported symptoms such as insomnia, psychomotor retardation, and strong headache, the attending psychiatrist also considers differential diagnoses such as a major depressive episode or a space-consuming intracerebral process. To exclude the latter, the attending psychiatrist refers her antipsychotic-naïve patient to a neuroradiologist to obtain a structural MRI. After segmentation of white and grey matter, the radiological data are fed to a machine learning algorithm which, based on previous training data in a comparable population, classifies the patient as suffering from schizophrenia with a probability of 70%. The psychiatrist sees her diagnosis confirmed and commences psychopharmacological treatment.
Case 3: T is diagnosed with a first episode of schizophrenia based on a clinical interview. To choose the most effective drug for his individual situation, his psychiatrist recommends a newly approved routine employing functional MRI during a reward-learning task. Based on T's brain activity and a plethora of other available information, from demographic data to his clinical records, the ML algorithm suggests one specific anti-psychotic drug as ideal for T's specific situation. Following the automated recommendation, the psychiatrist prescribes the drug to her patient.

psychiatric categorizations, some authors hope that embracing such a data-driven ML approach may provide new insights into neurobiological mechanisms of psychiatric diseases (Adams, Huys, & Roiser, 2016; Huys et al., 2016; Madsen, Krohne, Cai, Wang, & Chan, 2018; Skatun et al., 2017). A recent study that associated neuroanatomically distinct subtypes of schizophrenia with different illness duration and degrees of negative symptoms may serve as an example for this aspiration (Dwyer et al., 2018).

Also for diagnostic purposes, ML presents new opportunities for psychiatry. Based on specific changes in brain volume, several groups have shown that ML can distinguish non-medicated, first-episode patients with schizophrenia from healthy controls using volumetric MRI data (Chin, You, Meng, Zhou, & Sim, 2018; Gould et al., 2014; Haijma et al., 2013; Lee et al., 2018; Rozycki et al., 2018; Xiao et al., 2019). As noted, findings so far have been rather inconsistent and one should avoid overoptimistic interpretations of these results (Kambeitz et al., 2015; Vieira et al., 2020). Still, it seems reasonable to assume that in the future some ML techniques could assist physicians in their diagnostic process. Such applications could provide probabilistic estimates regarding one or several diagnostic labels such as schizophrenia, based on overlap with previously diagnosed patients. Arguably, most such methods would fall under the label of supervised learning since the training data need to be labelled, consisting of a vector of individual data such as brain data assigned to a category of 'diseased' *vs.* 'healthy', respectively.

Finally, recent psychiatric advances employing ML have seen a turn toward predicting certain quantifiable events beyond diagnostic labels, e.g. providing probabilities for the likelihood of an onset of psychosis (Koutsouleris et al., 2015, 2018) or for the treatment success of one certain drug (Chekroud et al., 2016; Webb et al., 2018). While the majority of these approaches draw on supervised or unsupervised ML, some also use reinforcement learning to derive recommendations for optimal dynamic treatment regimes, using e.g. longitudinal data from so-called *Sequential Multiple Assignment Randomized Trials* (SMARTs). For example, by considering the treatment success of specific anti-psychotics from the CATIE study (Stroup et al., 2003), Ertefaie, Shortreed, and Chakraborty (2016) have constructed a Q-learning approach which optimizes treatment outcome based on a patient's characteristics. Even more to the point, Koutsouleris et al. (2016) have shown that a cross-validated ML tool trained on diverse data from 334 patients could identify individuals which were more likely to benefit from treatment with amisulpride or olanzapine than with haloperidol, quetiapine, or ziprasidone. Such studies should be taken with a grain of salt though, given that there is no agreement what constitutes useful measures of treatment outcomes in psychiatry (Zimmerman & Mattia, 1999; Zimmerman, Morgan, & Stanton, 2018) – a conundrum the introduction of ML seems unlikely to solve.

To highlight the dissimilarities between different usages, we provide three schematic cases that fall within the range of possible applications, from research to diagnosis and choice of treatment (Table 2). All three cases, we hold it, touch upon important ethical concerns that can be discussed in accordance with the four principles put forth by Beauchamp and Childress: beneficence, non-maleficence, respect for autonomy, and justice (Beauchamp & Childress, 2013).

Beneficence

The principle of beneficence expresses an aspiration to further the welfare and interests of others, potentially implying particular obligations of acting (Beauchamp & Childress, 2013, p. 165–176). As our previous points and cases indicate, patients may benefit from applied ML in many different ways, both directly and indirectly.

Direct

Firstly, ML-supported diagnostic tools aim at improving diagnostic certainty. Techniques such as in the case of D (case 2) may serve as an automated second opinion, confirm a psychiatrist's judgement, and help with unclear cases. In fact, if the algorithm is trained on data of the highest quality, which are, e.g. labeled independently by several internationally leading and experienced psychiatrists, it could provide patients with a reliable diagnosis. Considering the difficulty of establishing whether schizophrenia is accurately diagnosed and given the considerable inter-rater disagreement among experts (Mokros, Habermeyer, & Kuchenhoff, 2018), a diagnostic algorithm supporting psychiatrists in their decision-making could increase the likelihood of patients receiving a correct diagnosis and hence of receiving an adequate treatment. By providing prognostic estimates concerning the future course of a disorder, such as the occurrence of psychotic episodes, or the success of specific treatments, ML applications may also

https://doi.org/10.1017/S0033291720001683 Published online by Cambridge University Press

help to reduce extraneous psychopharmacological interventions (Martinez-Martin et al., 2018) and track the progression of the disorder. This is the case for T (case 3), who may be spared an arduous trial-and-error regime of medication by an algorithm suggesting one potentially ideal medication early on. Of course, the benefits of a correct diagnosis might be infringed dramatically by additional risks, to which we turn later, if these diagnostic or predictive processes were to be left unchecked. However, at least for now, such a development seems rather unlikely, both technically and socially, in most medical specialties (Topol, 2019).

Indirect

Beyond these immediate clinical uses, patients may also benefit from research projects similar to our first case, leading to more accurate diagnostic categories. After all, most current psychiatric diagnoses as enshrined in the DSM or ICD are purely descriptive, optimized primarily for validity and inter-rater reliability, not for underlying pathophysiology – but this lack of concern for etiological underpinnings has long been of concern to many in the field (Hyman, 2011). In contrast, computational approaches based on ML aspire 'to automatically segregate brain disorders into natural kinds' (Bzdok & Meyer-Lindenberg, 2018, p. 223). Notwithstanding conceptual questions regarding the nature of psychiatric disorders (Kendler, 2016; Zachar, 2015), ML may be eminently suited to develop biologically more plausible diagnostic categories, allowing for more specific treatment options. After all, concerns of insufficiently grasping psychiatric complexity have long accompanied the development of psychiatric biomarkers (Singh & Rose, 2009). ML drawing on rich data, from detailed biological information such as (f)MRI scans or whole-genome sequences to demographic data and electronic health records, could arguably accommodate such complexity. Still, the concern remains that ML applications drawing on ML may overtly reify diagnostic categories designed as heuristic constructs (Hyman, 2010) – and thus end up harming patients.

Non-maleficence

Abstaining from harm is a bedrock of clinical practice (Smith, 2005). How does ML in psychiatry fare with regard to this crucial principle? Firstly, privacy concerns may come to mind here (Vayena et al., 2018). How is sensitive medical information disclosed to an algorithm and how can data created by the algorithm be protected appropriately? These are essential questions but only concern ML techniques indirectly, via the data used and produced by its applications. Since privacy issues of big data have been addressed extensively elsewhere (Price & Cohen, 2019), we will leave them aside here to focus on harm potentially caused by ML in psychiatry. As in the case of benefits, there are both direct and indirect ways in which its use may harm patients.

Direct

First, using an algorithm may bring about harm directly, e.g. when the diagnosis or predictions made by the ML application are erroneous. Previous shortcomings of health-related ML can be instructive here. IBM's ML-based computer system Watson, advertised as a revolutionary tool for cancer care, has been shown to recommend unsafe treatments endangering patients' safety and health (Ross & Swetlitz, 2018). Such errors are particularly worrying if recommendations of algorithms are readily accepted by medical staff, as in T's case, or if the process would become fully automated. Although an erroneous algorithm is likely to affect more patients compared to an individual mistake made by a physician, errors are far from exclusive to algorithms (McLennan et al., 2013), and these concerns could be tackled by a model of shared responsibility in which competent human agents check the ML-based suggestions (Topol, 2019). However, as opposed to human physicians, a trained ML algorithm may not be flexible enough to account for contextual changes such as the swift rise of smartphone usage or altered eating habits. Given the dependency of psychiatric conditions on contingent societal contexts, even a tested and approved program may thus require regular overhauling and retraining to avoid systematic misjudgments.

Indirect

The more intricate questions seem to arise from indirect effects of using ML in patients with schizophrenia. By potentially modifying the expectations of doctors, the result of a computationally assigned risk-category will most likely influence downstream diagnostic and therapeutic decision-making. For example, in mammography screening, risk stratification affects the detection performance of radiologists: a known BRCA mutation strongly decreases the number of missed visible breast cancer lesions in MRI scans (Vreemann et al., 2018). Timing the disclosure of ML-based computations to the physician is thus crucial: should she have to decide on one diagnosis first before being confronted with the results of ML diagnostics? Furthermore, the impact of incorporating ML in the clinical setting will require additional scrutiny regarding its effects on the therapeutic relationship. How do patients perceive the use of ML by their physicians to arrive at diagnostic judgements or prognostic estimates? Does it impair their trust in health care professionals and if so, could it harm their compliance and the therapeutic outcome? These questions are of particular importance in the case of psychiatric patients who are particularly vulnerable to so-called 'diagnostic overshadowing', i.e. health care professionals falsely attributing somatic symptoms to known mental health issues (Callard, Bracken, David, & Sartorius, 2013; Jones, Howard, & Thornicroft, 2008; Shefer, Henderson, Howard, Murray, & Thornicroft, 2014). These challenges merit ongoing attention and require accompanying efforts of clinical ML implementation with corresponding empirical bioethical research to explore the potential negative impact.

Patients' autonomy and clinicians' judgement

Respect for autonomy demands conveying sufficiently detailed and understandable information to patients about planned medical procedures and asking for their consent (Manson & O'Neill, 2007). Such disclosure may be particularly challenging in cases of applied ML, used by medical practitioners who may themselves not fully understand the mathematical underpinnings of an algorithm. Does the, to some extent, unavoidable opacity of ML, commonly discussed as 'black box'-problem, clash with the requirement to appropriately inform patients? And should one ask patients for their explicit consent when using (existing) data before providing it to the algorithm at all? After all, obtaining informed consent for the use of predictive analytics is not legally mandatory at the moment (Cohen, Amarasingham, Shah, Xie, & Lo, 2014). One could wonder whether discussing ML algorithms with a group as vulnerable as patients at risk of psychosis or paranoid symptoms might not exacerbate their situation and cause severe additional psychological stress (Martinez-Martin et al., 2018).

https://doi.org/10.1017/S0033291720001683 Published online by Cambridge University Press

Questions of autonomy also stretch to the domain of medical doctors' discernment, and respecting clinicians' judgement is vital in the context of modern health care systems (Faden et al., 2013). Much depends on the conceptualization of the relation between human expert and ML algorithm. One analogy, recently proposed by Topol (2019), suggests that we conceptualize the relation of clinician and algorithm similarly to assisted driving and increasingly autonomous cars. While the machine may take over some tasks, the drivers or physicians need to remain in charge as a backup, checking the machine's output by comparing it to their own judgements. This would facilitate attributing degrees of responsibility to health care personnel, clarifying important issues of accountability and liability. It implies that human agents need to remain able to weigh ML recommendations and potentially decide against them. Ideally, as a safeguard against bad judgements by single individuals, one could envision provisions in which disagreements between physicians and ML application lead to consultations with other clinicians, e.g. during departmental meetings, providing an opportunity to sharpen the clinical skills of everyone involved. Furthermore, an institutional framework may be needed to test and approve ML applications in a similar fashion as pharmaceutical products (Paulus, Huys, & Maia, 2016).

Fair allocation and systematic biases

Finally, using ML in psychiatry also raises important issues concerning justice, from financial aspects to systematic biases. Does increased diagnostic certainty justify the allocation of scarce financial means to additional computational efforts and vindicate even highly expensive exams such as (f)MRI? Integrating the data from examinations such as MRI into psychiatric routines may pose additional serious challenges for equal treatment if certain patients cannot undergo scanning due to limited availability or contraindications such as claustrophobia. Arguably, any new technique needs to establish a measurable clinical benefit over a conventional psychiatric assessment to vindicate its cost (Iwabuchi, Liddle, & Palaniyappan, 2013), or show that it can avoid costs elsewhere. With regard to discerning different diagnostic entities, research based on ML could also lead to issues commonly known as salami slicing: even without understanding the underlying pathophysiological mechanisms, lobbying by pharmaceutical companies might have an interest to split psychiatric disorders into many distinct categories to gain advantages in the approval of new drugs. On the other hand, we should not forget that in many countries, only a very limited amount of the overall healthcare budget is allocated to mental health (World Health Organization, 2018). More precise diagnoses and better treatments might convince policymakers to overcome this health disparity, ultimately empowering psychiatric patients.

Of further concern are systematic biases, easily induced by poor training data and particularly worrisome in diagnostic contexts (Vayena et al., 2018). The example of schizophrenia is a case in point, with its long-standing disproportionate number of diagnoses in African-Americans and Latin-Americans, arguably influenced by stereotypes, the clinician's own ethnicity, or the under-diagnosis of other psychiatric diseases (Schwartz & Blankenship, 2014). ML trained on data with these or other biases could further purport and reify misconceptions (Tandon & Tandon, 2018). If training data are less than carefully curated, ML applications might hence not constitute an independent diagnostic tool for enhancing diagnostic accuracy, undermining the endeavor's very aim. To avoid perpetuating pathophysiologically misleading biases, developing appropriate supervision strategies for the ML algorithm thus seems key to a successful clinical implementation. Such supervision should (1) track which parameters are taken into account by the algorithm to arrive at its recommendations and (2) compare the results of algorithms trained on different databases. Such strategies would also help to foster explicability which the initially mentioned AI4people initiative rightly suggests as a fifth principle for ethical AI use, enabling the other four (Floridi et al., 2018). The implementation of such safety measures will be critical for minimizing biases in decision-making but it is not yet clear how ML algorithms will nonetheless capitalize on existing biases in the data.

Conclusion

A plethora of context-specific ethical issues might arise in applied ML in psychiatry and the treatment of schizophrenia. For now, ML remains in the domain of research and should be accompanied by exploring its ethical aspects as there is no standard rule to determine when an application is ethically permissible given the complexity of each singular case. Further, empowering psychiatric patients can only happen with the help of important support systems such as family, peer, and community members. Still, if some of the vast potential benefits of psychiatric ML can indeed lead to tangible improvements for patients, we believe it is not only permissible but it may in fact be a moral obligation to pursue them further and aim at their successful clinical implementation.

Acknowledgements. The authors would like to thank Andrea Martani and Christopher Poppe for their helpful comments as well as the three anonymous reviewers for substantial improvements on a previous draft.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of interest. None.

References

Adams, R. A., Huys, Q. J., & Roiser, J. P. (2016). Computational psychiatry: Towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery, and Psychiatry*, *87*(1), 53–63. doi: 10.1136/jnnp-2015-310737

Barron, D. (2019, February 19). Should mental disorders have names? [Blog post]. *Scientific American*. Retrieved from https://blogs.scientificamerican.com/observations/should-mental-disorders-have-names

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7th ed.). New York: Oxford University Press.

Borgwardt, S., Koutsouleris, N., Aston, J., Studerus, E., Smieskova, R., Riecher-Rössler, A., & Meisenzahl, E. M. (2013). Distinguishing prodromal from first-episode psychosis using neuroanatomical single-subject pattern recognition. *Schizophrenia Bulletin*, *39*(5), 1105–1114. doi: 10.1093/schbul/sbs095

Borsboom, D., Cramer, A., & Kalis, A. (2019). Brain disorders? Not really… Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, *42*(e2), 1–63. doi: 10.1017/S0140525X17002266.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi: 10.1038/nrn3475

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230. doi: 10.1016/j.bpsc.2017.11.007

https://doi.org/10.1017/S0033291720001683 Published online by Cambridge University Press

Callard, F., Bracken, P., David, A. S., & Sartorius, N. (2013). Has psychiatric diagnosis labelled rather than enabled patients? *BMJ*, *347*, f4312. doi: 10.1136/bmj.f4312

Chapelle, O., Schölkopf, B., & Zien, A. (2010). *Semi-supervised learning*. Cambridge, Mass: MIT Press.

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care – addressing ethical challenges. *New England Journal of Medicine*, *378*(11), 981–983. doi: 10.1056/NEJMp1714229

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., … Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *Lancet Psychiatry*, *3* (3), 243–250. doi: 10.1016/S2215-0366(15)00471-X

Chin, R., You, A. X., Meng, F., Zhou, J., & Sim, K. (2018). Recognition of schizophrenia with regularized support vector machine and sequential region of interest selection using structural magnetic resonance imaging. *Scientific Reports*, *8*(1), 13858. doi: 10.1038/s41598-018-32290-9

Chung, Y., Addington, J., Bearden, C. E., Cadenhead, K., Cornblatt, B., Mathalon, D. H., … Genetics Study, C. (2018). Use of machine learning to determine deviance in neuroanatomical maturity associated with future psychosis in youths at clinically high risk. *JAMA Psychiatry*, *75*(9), 960–968. doi: 10.1001/jamapsychiatry.2018.1543

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., & Lo, B. (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs*, *33*(7), 1139–1147. doi: 10.1377/hlthaff.2014.0048

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., … Liston, C. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, *23*(1), 28–38. doi: 10.1038/nm.4246

Dwyer, D. B., Cabral, C., Kambeitz-Ilankovic, L., Sanfelici, R., Kambeitz, J., Calhoun, V., … Koutsouleris, N. (2018). Brain subtyping enhances the neuroanatomical discrimination of schizophrenia. *Schizophrenia Bulletin*, *44*(5), 1060–1069. doi: 10.1093/schbul/sby008

Ebdrup, B. H., Axelsen, M. C., Bak, N., Fagerlund, B., Oranje, B., Raghava, J. M., … Glenthoj, B. Y. (2019). Accuracy of diagnostic classification algorithms using cognitive-, electrophysiological-, and neuroanatomical data in antipsychotic-naive schizophrenia patients. *Psychological Medicine*, *49* (16), 2754–2763. doi: 10.1017/s0033291718003781.

Ertefaie, A., Shortreed, S., & Chakraborty, B. (2016). Q-learning residual analysis: Application to the effectiveness of sequences of antipsychotic medications for patients with schizophrenia. *Statistics in Medicine*, *35*(13), 2221–2234. doi: 10.1002/sim.6859

Faden, R. R., Kass, N. E., Goodman, S. N., Pronovost, P., Tunis, S., & Beauchamp, T. L. (2013). An ethics framework for a learning health care system: A departure from traditional research ethics and clinical ethics. *Hastings Center Report*, *43*(s1), S16–S27. doi: 10.1002/hast.134.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People – an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. doi: 10.1007/s11023-018-9482-5

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., … Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, *143*(2), 187–232. doi: 10.1037/bul0000084

Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics*, *24*(11), 1037–1052. doi: 10.1111/cns.13048

Giordano, G. M., & Borgwardt, S. (2019). Current goals of neuroimaging for mental disorders: A report by the WPA section on neuroimaging in psychiatry. *World Psychiatry*, *18*(2), 241–242. doi: 10.1002/wps.20652

Gould, I. C., Shepherd, A. M., Laurens, K. R., Cairns, M. J., Carr, V. J., & Green, M. J. (2014). Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: A support vector machine learning approach. *Neuroimage-Clinical*, *6*, 229–236. doi: 10.1016/j.nicl.2014.09.009

Haijma, S. V., Van Haren, N., Cahn, W., Koolschijn, P. C., Hulshoff Pol, H. E., & Kahn, R. S. (2013). Brain volumes in schizophrenia: A meta-analysis in over 18 000 subjects. *Schizophrenia Bulletin*, *39*(5), 1129–1138. doi: 10.1093/schbul/sbs118

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., … Yeo, B. T. T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage*, *206*, 116276. doi: 10.1016/j.neuroimage.2019.116276.

Heinrichs, B., & Eickhoff, S. B. (2020). Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping*, *41*(6), 1435–1444. doi: 10.1002/hbm.24886.

High-Level Expert Group on Artificial Intelligence (AI HLEG) (2019). *Ethics guidelines for trustworthy AI*. Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation

Hofstadter, D. R. (1980). *Gödel, Escher, Bach: An eternal golden braid*. New York: Vintage Books.

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19* (3), 404–413. doi: 10.1038/nn.4238

Hyman, S. E. (2010). The diagnosis of mental disorders: The problem of reification. *Annual Review of Clinical Psychology*, *6*, 155–179. doi: 10.1146/annurev.clinpsy.3.022806.091532

Hyman, S. E. (2011). Diagnosing the DSM: Diagnostic classification needs fundamental reform. *Cerebrum*, *2011*, 6.

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, *46*(12), 2455–2465. doi: 10.1017/S0033291716001367

Iwabuchi, S. J., Liddle, P. F., & Palaniyappan, L. (2013). Clinical utility of machine-learning approaches in schizophrenia: Improving diagnostic confidence for translational neuroimaging. *Frontiers in Psychiatry*, *4*, 95. doi: 10.3389/fpsyt.2013.00095

Janssen, R. J., Mourao-Miranda, J., & Schnack, H. G. (2018). Making individual prognoses in psychiatry using neuroimaging and machine learning. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(9), 798–808. doi: 10.1016/j.bpsc.2018.04.004

Jones, S., Howard, L., & Thornicroft, G. (2008). 'Diagnostic overshadowing': Worse physical health care for people with mental illness. *Acta Psychiatrica Scandinavica*, *118*(3), 169–171. doi: 10.1111/j.1600-0447.2008.01211.x

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D., Cha, C., Nock, M. K., & Brent, D. (2017). Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature Human Behaviour*, *1*, 911–919. doi: 10.1038/s41562-017-0234-y

Kambeitz, J., Kambeitz-Ilankovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., … Koutsouleris, N. (2015). Detecting neuroimaging biomarkers for schizophrenia: A meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology*, *40*(7), 1742–1751. doi: 10.1038/npp.2015.22

Kapur, S., Phillips, A. G., & Insel, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, *17*(12), 1174–1179. doi: 10.1038/mp.2012.105

Kellmeyer, P. (2017). Ethical and legal implications of the methodological crisis in neuroimaging. *Cambridge Quarterly of Healthcare Ethics*, *26*(4), 530–554. doi: 10.1017/S096318011700007X

Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry*, *15* (1), 5–12. doi: 10.1002/wps.20292

Kendler, K. S., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychological Medicine*, *41*(6), 1143–1150. doi: 10.1017/S0033291710001844

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., … Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: A machine learning approach. *Lancet Psychiatry*, *3*(10), 935–946. doi: 10.1016/S2215-0366(16)30171-7

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D. B., … Consortium, P. (2018). Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: A multimodal, multisite machine learning analysis. *JAMA Psychiatry*, *75*(11), 1156–1172. doi: 10.1001/jamapsychiatry.2018.2165

Koutsouleris, N., Riecher-Rössler, A., Meisenzahl, E. M., Smieskova, R., Studerus, E., Kambeitz-Ilankovic, L., … Borgwardt, S. (2015). Detecting the psychosis prodrome across high-risk populations using neuroanatomical biomarkers. *Schizophrenia Bulletin*, *41*(2), 471–482. doi: 10.1093/schbul/sbu078

Krystal, J. H., Murray, J. D., Chekroud, A. M., Corlett, P. R., Yang, G., Wang, X. J., & Anticevic, A. (2017). Computational psychiatry and the challenge of schizophrenia. *Schizophrenia Bulletin*, *43*(3), 473–475. doi: 10.1093/schbul/sbx025

LeCun, Y. (2018). The power and limits of deep learning. *Research-Technology Management*, *61*(6), 22–27. doi: 10.1080/08956308.2018.1516928

Lee, J., Chon, M. W., Kim, H., Rathi, Y., Bouix, S., Shenton, M. E., & Kubicki, M. (2018). Diagnostic value of structural and diffusion imaging measures in schizophrenia. *Neuroimage-Clinical*, *18*, 467–474. doi: 10.1016/j.nicl.2018.02.007

Madsen, K. H., Krohne, L. G., Cai, X. L., Wang, Y., & Chan, R. C. K. (2018). Perspectives on machine learning for classification of schizotypy using fMRI data. *Schizophrenia Bulletin*, *44*(suppl_2), S480–S490. doi: 10.1093/schbul/sby026

Manson, N. C., & O'Neill, O. (2007). *Rethinking informed consent in bioethics*. Cambridge; New York: Cambridge University Press.

Martinez-Martin, N., Dunn, L. B., & Roberts, L. W. (2018). Is it ethical to use prognostic estimates from machine learning to treat psychosis? *AMA Journal of Ethics*, *20*(9), E804–E811. doi: 10.1001/amajethics.2018.804

McLennan, S., Engel, S., Ruhe, K., Leu, A., Schwappach, D., & Elger, B. (2013). Implementation status of error disclosure standards reported by Swiss hospitals. *Swiss Medical Weekly*, *143*, w13820. doi: 10.4414/smw.2013.13820

Mihalik, A., Brudfors, M., Robu, M., Ferreira, F. S., Lin, H., Rau, A., ...Oxtoby, N. P. (2019). ABCD Neurocognitive Prediction Challenge 2019: Predicting Individual Fluid Intelligence Scores from Structural MRI Using Probabilistic Segmentation and Kernel Ridge Regression. In Pohl, K., Thompson, W., Adeli, E., & Linguraru, M. (Eds.), *Adolescent Brain Cognitive Development Neurocognitive Prediction* (pp. 133–142). Cham: Springer.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. doi: 10.1038/nature14236

Mokros, A., Habermeyer, E., & Kuchenhoff, H. (2018). The uncertainty of psychological and psychiatric diagnoses. *Psychological Assessment*, *30*(4), 556–560. doi: 10.1037/pas0000524

Paulus, M. P., Huys, Q. J., & Maia, T. V. (2016). A roadmap for the development of applied computational psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(5), 386–392. doi: 10.1016/j.bpsc.2016.05.001

Perna, G., Grassi, M., Caldirola, D., & Nemeroff, C. B. (2018). The revolution of personalized psychiatry: Will technology make it happen sooner? *Psychological Medicine*, *48*(5), 705–713. doi: 10.1017/S0033291717002859

Price, II. W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, *25*(1), 37–43. doi: 10.1038/s41591-018-0272-7

Ross, C., & Swetlitz, I. (2018). IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. Stat News. Retrieved from https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

Rozycki, M., Satterthwaite, T. D., Koutsouleris, N., Erus, G., Doshi, J., Wolf, D. H., … Davatzikos, C. (2018). Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. *Schizophrenia Bulletin*, *44*(5), 1035–1044. doi: 10.1093/schbul/sbx137

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. doi: 10.1038/tp.2015.182

Schnack, H. G. (2019). Improving individual predictions: Machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophrenia Research*, *214*, 34–42. doi: 10.1016/j.schres.2017.10.023.

Schnack, H. G., & Kahn, R. S. (2016). Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters. *Frontiers in Psychiatry*, *7*, 50. doi: 10.3389/fpsyt.2016.00050

Schwartz, R. C., & Blankenship, D. M. (2014). Racial disparities in psychotic disorder diagnosis: A review of empirical literature. *World Journal of Psychiatry*, *4*(4), 133–140. doi: 10.5498/wjp.v4.i4.133

Shatte, A. B. R., 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. doi: 10.1017/s0033291719000151

Shefer, G., Henderson, C., Howard, L. M., Murray, J., & Thornicroft, G. (2014). Diagnostic overshadowing and other challenges involved in the diagnostic process of patients with mental illness who present in emergency departments with physical symptoms-a qualitative study. *PLoS ONE*, *9*(11), e111682. doi: 10.1371/journal.pone.0111682

Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, *19*, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Singh, I., & Rose, N. (2009). Biomarkers in psychiatry. *Nature*, *460*(7252), 202–207. doi: 10.1038/460202a

Skatun, K. C., Kaufmann, T., Doan, N. T., Alnaes, D., Cordova-Palomera, A., Jonsson, E. G., … Westlye, L. T. (2017). Consistent functional connectivity alterations in schizophrenia spectrum disorder: A multisite study. *Schizophrenia Bulletin*, *43*(4), 914–924. doi: 10.1093/schbul/sbw145

Smith, C. M. (2005). Origin and uses of primum non nocere – above all, do no harm!. *Journal of Clinical Pharmacology*, *45*(4), 371–377. doi: 10.1177/0091270004273680

Stahl, D., & Pickles, A. (2018). Fact or fiction: Reducing the proportion and impact of false positives. *Psychological Medicine*, *48*(7), 1084–1091. doi: 10.1017/S003329171700294X

Stroup, T. S., McEvoy, J. P., Swartz, M. S., Byerly, M. J., Glick, I. D., Canive, J. M., … Lieberman, J. A. (2003). The National Institute of Mental Health Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) project: Schizophrenia trial design and protocol development. *Schizophrenia Bulletin*, *29*(1), 15–31. doi: 10.1093/oxfordjournals.schbul.a006986

Tandon, N., & Tandon, R. (2018). Will machine learning enable us to finally cut the Gordian knot of schizophrenia. *Schizophrenia Bulletin*, *44*(5), 939–941. doi: 10.1093/schbul/sby101

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. doi: 10.1038/s41591-018-0300-7

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, *15*(11), e1002689. doi: 10.1371/journal.pmed.1002689

Vieira, S., Gong, Q. Y., Pinaya, W. H. L., Scarpazza, C., Tognin, S., Crespo-Facorro, B., … Mechelli, A. (2020). Using machine learning and structural neuroimaging to detect first episode psychosis: Reconsidering the evidence. *Schizophrenia Bulletin*, *46*(1), 17–26. doi: 10.1093/schbul/sby189.

Vreemann, S., Gubern-Merida, A., Lardenoije, S., Bult, P., Karssemeijer, N., Pinker, K., & Mann, R. M. (2018). The frequency of missed breast cancers in women participating in a high-risk MRI screening program. *Breast Cancer Research and Treatment*, *169*(2), 323–331. doi: 10.1007/s10549-018-4688-z

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, *5*(3), 457–469. doi: 10.1177/2167702617691560

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., … Pizzagalli, D. A. (2018). Personalized prediction of antidepressant v. placebo response: Evidence from the EMBARC study. *Psychological Medicine*, *49*(7), 1118–1127. doi: 10.1017/S0033291718001708

World Health Organization (2018). *Mental health atlas 2017*. Geneva, Switzerland: WHO.

Xiao, Y., Yan, Z., Zhao, Y., Tao, B., Sun, H., Li, F., … Lui, S. (2019). Support vector machine-based classification of first episode drug-naive schizophrenia patients and healthy controls using structural MRI. *Schizophrenia Research*, *214*, 11–17. doi: 10.1016/j.schres.2017.11.037.

Zachar, P. (2015). Psychiatric disorders: Natural kinds made by the world or practical kinds made by us? *World Psychiatry*, *14*(3), 288–290. doi: 10.1002/wps.20240

Zimmerman, M., & Mattia, J. I. (1999). Psychiatric diagnosis in clinical practice: Is comorbidity being missed? *Comprehensive Psychiatry*, *40*(3), 182–191. doi: 10.1016/s0010-440x(99)90001-9

Zimmerman, M., Morgan, T. A., & Stanton, K. (2018). The severity of psychiatric disorders. *World Psychiatry*, *17*(3), 258–275. doi: 10.1002/wps.20569

https://doi.org/10.1017/S0033291720001683 Published online by Cambridge University Press