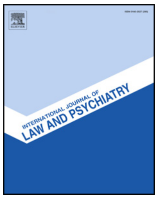
Contents lists available at ScienceDirect

International Journal of Law and Psychiatry

journal homepage: www.elsevier.com/locate/ijlawpsy



The risks of risk. Regulating the use of machine learning for psychosis prediction

Paolo Corsico*

Centre for Social Ethics and Policy, Department of Law, School of Social Sciences, The University of Manchester, United Kingdom

ARTICLE INFO

Keywords:
Psychosis
Machine learning
Risk
Prediction
Regulation

ABSTRACT

Recent advances in Machine Learning (ML) have the potential to revolutionise psychosis prediction and psychiatric assessment. This article has two objectives. First, it clarifies which aspects of English Law are relevant in order to regulate the use of ML in clinical research on psychosis prediction. It is argued that its lawful implementation will depend upon the legal requirements regarding the balance between potential harms and benefits, particularly with reference to: (i) any additional risks introduced by the use of ML for data analysis and outcome prediction; and (ii) the inclusion of vulnerable research populations such as minors or incapacitated adults. Second, this article investigates how clinical prediction via ML might affect the practice of risk assessment under mental health legislation, with reference to English Law. It is argued that there is a potential for virtuous applications of clinical prediction in psychiatry. However, reaffirming the distinction between psychosis risk and risk of harm is paramount. Establishing psychosis risk and assessing a person's risk of harm are discrete practices, and so should remain when using artificial intelligence for psychiatric assessment. Evaluating whether clinical prediction via ML might benefit individuals with psychosis will depend on which risk we try to assess and on what we try to predict, whether this is psychosis transition, a psychotic relapse, self-harm and suicidality, or harm to others.

1. Introduction[1]

David Reynolds was diagnosed with schizophrenia in 1998 at the age of 29. On 16 March 2005, Mr. Reynolds contacted his care coordinator at a local National Health Service (NHS) mental health team: he was hearing voices ordering him to kill himself. The care coordinator told him that he could have a crisis bed in a local intensive support unit. A clinical assessment was conducted and he told the psychiatrist that he did not want to kill himself. He had no history of self-harm or attempted suicide. Mr. Reynolds was assessed to be a low suicide risk and was admitted as a voluntary inpatient. At around 10.30 pm, he broke a window in his room and fell from the sixth floor to his death. In March 2012, the European Court of Human Rights (ECtHR) ruled that there had been a violation of Article 2(1) of the European Convention of Human Rights (ECHR), which provides, "everyone's right to life shall be protected by law.[2]" Did Mr. Reynolds want to kill himself? What if we had reliable measures of suicide risk that do not rely on self-reporting? What would happen if we could predict whether someone will develop a psychotic episode before he or she eventually slips into full-blown psychosis?

The advent of countless forms of Artificial Intelligence (AI) is transforming the way in which we design and deliver health care (Alpaydin, 2016; Dadich, 2016; Shatte, Hutchinson, & Teague, 2019). Legislation, however, evolves slower than technology. Whilst the hype is great, numerous are the questions: how will technology transform the way in which we diagnose mental illness? How do we regulate the myriad of AI applications that are being developed in mental health care? Most importantly, how can AI applications be developed to be 'forces for good' (Taddeo & Floridi, 2018), and to not become the biased actors of injustice and State control that Cathy O'Neil has eloquently described as 'weapons of math destruction' (O'Neill, 2016)?

The present article provides a contribution to this emerging field of ethical and legal theory by trying to answer a narrower question: What legal challenges arise from the use of Machine Learning (ML)—a specific form of AI—in the context of psychosis prediction? More specifically, the present contribution has two objectives. First, it wishes to clarify which aspects of English Law are relevant in order to regulate the use of ML in *clinical research* on psychosis prediction. Second, it

* The University of Manchester, Williamson Building, Oxford Road, M13 9PL Manchester, United Kingdom.
E-mail address: paolo.corsico@postgrad.manchester.ac.uk.

[1] The following non-standard abbreviations are used consistently in this article: AI = Artificial Intelligence; HR = High-Risk; ML = Machine Learning

[2] *Reynolds v. The United Kingdom* (2012) 55 EHRR 35.

https://doi.org/10.1016/j.ijlp.2019.101479
Received 27 February 2019; Received in revised form 21 July 2019; Accepted 24 July 2019
Available online 17 August 2019

wishes to investigate how ML might affect the practice of *psychiatric assessment* in the context of psychosis. Technological innovation opens up the possibility to identify psychosis risk, predict psychosis transition or relapse, and foresee harm to self or others. This article explores the notions of 'psychosis risk' and 'risk of harm'. It investigates how these notions relate to the issue of diagnostic uncertainty in psychiatric assessment, and the implications of such concepts for legal theory. It is argued that there is a potential for virtuous applications of AI-mediated prediction in mental health. However, maximizing this potential will require a careful evaluation of how we interpret different notions of 'risk'. This article refers to the jurisdiction of England and Wales. However, the reflections presented here may be relevant to other jurisdictions.

2. Machine learning and psychosis prediction(s)

"Psychiatry is not an exact science".[3] Nonetheless, technology is reshaping the way in which we understand, diagnose, and treat psychotic illness. This article does not address the question of whether it is ethically acceptable to use ML to predict psychotic illness. Rather, it provides some clarifications on how to regulate a number of recent advances in medical technology in the context of psychosis. It follows a simple principle: in order to understand how to regulate the use of medical technologies, it is essential to understand how such technologies are used, and how they are likely to be used in the near future. According to Jiang et al., AI systems are being designed in health care with three aims: (i) to reduce diagnostic uncertainty by helping clinicians to classify individuals in different populations; (ii) to support identification of at-risk status for specific conditions, and (iii) to help to predict health outcomes, thus supporting clinical prevention (Jiang et al., 2017). Classification, risk identification, and prediction are the ways in which AI—mostly ML applications—is currently used in the context of psychosis.

Let us unfold this further. First, what is ML? The term ML identifies a number of techniques used to analyse data and to make outcome predictions. The Information Commissioner's Office (ICO) has borrowed a definition from iQ, Intel's tech culture magazine, which defines ML as the "set of techniques and tools that allow computers to 'think' by creating mathematical algorithms based on accumulated data".[4] ML is a sub-field of narrow AI—that is, AI which is designed for a specific application, and not to fully resemble human intelligence (Turner, 2018). It is used for data analysis, whereby software is trained from a dataset to classify the data, identify patterns, and make predictions. As Alpaydin writes, "[…] it is easy to collect data, and now the idea is to learn the algorithms for these [applications] automatically from the data, replacing programmers with learning programmes. This is the niche of machine learning."[5] ML algorithms can be described as *supervised*, *unsupervised*, or *semi-supervised* depending on the degree of preclassification of the datasets used to train and develop the algorithm (Shatte et al., 2019).

Second, how is ML currently used in mental health? Shatte et al. (2019) have recently provided a very useful scoping review. The authors identify four domains of application: (i) detection and diagnosis; (ii) prognosis, treatment and support; (iii) public health; and (iv) research and clinical administration. The vast majority of studies reviewed falls within the first two domains. The first domain includes studies that attempt to *classify* clinical groups and *identify risk* status for mental health conditions, among which are psychosis and schizophrenia. The second domain includes attempts to *predict* clinical outcomes again for a number of conditions including psychosis and schizophrenia. Interestingly, the second domain also includes studies whose purpose was to identify suicidal ideation and to predict self-harm and suicide. Indeed, self-harm and suicidal ideation do not represent discrete diagnostic categories. Rather, they are occurrences which characterise mental disorders across the diagnostic spectrum (Harris & Barraclough, 1997), and which have been reported also in the early stages of psychosis (Xu et al., 2016).

Third, which types of data sources are fed to ML algorithms to accomplish these tasks? Data sources used to develop ML algorithms vary considerably. They include—but are not limited to—neuroimaging data, clinical assessment and clinical record data, digital health data collected via wearables and digital phenotyping (Jain, Powers, Hawkins, & Brownstein, 2015), and speech data. Shatte et al. (2019) highlight that "the majority of studies investigating the detection and diagnosis of mental health conditions used neuroimaging data with supervised classification techniques".[6] Neuroimaging has traditionally been used to investigate the neural correlates of psychosis and schizophrenia.[7] In the past decades, early intervention services have started promoting psychosis prevention by targeting people at clinical High-Risk (HR) of psychosis (Fusar-Poli et al., 2013). Neuroimaging data are currently used to investigate the psychosis HR state, to support the classification of clinical groups (Kempton & McGuire, 2015; Valli et al., 2016), and to predict psychosis transition in HR individuals (Gifford et al., 2017). For instance, Koutsouleris et al. were recently able to predict transition outcomes in 80% of individuals using MRI data (Koutsouleris et al., 2015). Clinical assessment and clinical record data are being used to predict psychosis transition in HR individuals (Mechelli et al., 2017), and to identify suicidal ideation and predict suicide attempts (Fernandes et al., 2018; Walsh, Ribeiro, & Franklin, 2017). The use of digital health data in digital phenotyping holds promise to revolutionise prediction of psychosis *relapse* in individuals with schizophrenia (Barnett et al., 2018; Torous et al., 2018). In addition, also speech data are being used to predict psychosis transition in HR individuals (Bedi et al., 2015; Corcoran et al., 2018; Rezaii, Walker, & Wolff, 2019) and to identify suicide ideation and suicidal attempts (Fernandes et al., 2018).

What applications are relevant for the purpose of our analysis? In other words, what is relevant when we discuss ML-mediated psychosis prediction? We can identify three main areas of interest: (i) the use of ML with different datasets to enhance the identification of psychosis *risk* and the prediction of psychosis *transition* in HR individuals; (ii) the use of ML for prediction of psychosis *relapse* in individuals who suffer from a psychotic disorder, including schizophrenia; and (iii) the use of ML for the identification of risk of *harm*, as well as for prediction of self-harm and suicide in the context of psychosis. If the relevance of the first two areas stems from the importance of diagnostic and prognostic prediction in the context of psychosis, the importance of the third area originates from the fact that mental health legislation—at least in England and Wales—places great emphasis on the assessment of risks associated with mental disorders (Bartlett & Sandland, 2014; Fanning, 2016; Glover-Thomas, 2011). Legal scholarship has traditionally targeted the use of neuro-technology for the evaluation of criminal responsibility or for establishing the presence of a neurological condition (Meynen, 2013; Spranger, 2012). Little attention has been dedicated to the use of AI for psychosis prediction and civil admission of the mentally ill. In this paper, I argue that legal challenges emerge within two domains: the conduct of clinical research and the practice of psychiatric

[3] *Regina v. Ashworth Hospital Authority* [2005] UKHL 20, at para 31.

[4] See Information Commissioner's Office (2017), p. 7.

[5] Alpaydin (2016), "Preface", p. X.

[6] Shatte et al. (2019), p. 1434.

[7] 'Neuroimaging' is used here as an umbrella term to indicate a number of techniques used to study brain structures (structural Magnetic Resonance Imaging, or MRI), brain functions (functional Magnetic Resonance Imaging, or fMRI), or neurotransmitter dysfunction and other molecular processes (various imaging techniques such as Single Photon Emission Tomography, or SPET, and Positron Emission Tomography, or PET). For an interesting overview of neuroimaging techniques used in psychosis studies, see McGuire, Howes, Stone, and Fusar-Poli (2008).

assessment.

3. Regulating clinical research

In their scoping review of ML applications in mental health, Shatte et al. state that: "Very little research was found that demonstrated the use of ML techniques in real-world settings, suggesting that further research is required to test clinical utility."[8] Psychosis prediction via ML has yet to become a clinical reality. In addition, as Shatte et al. have noted, even though the majority of ML studies investigating detection and diagnosis have until now used neuroimaging data, it seems problematic to foresee widespread access to imaging services for diagnostic purposes.[9] Thus, we might argue that the implementation of ML which uses other data sources seems, to date, more likely. Nonetheless, cases in which ML is used in psychiatric practice seem to be still rare. Before this translation happens, ML will make its way into the realm of clinical research. For this reason, I believe it is important to investigate how research ethics restrictions might influence later downstream applications. The debate on how to regulate AI in clinical research is still in its infancy. Efforts to regulate AI with statutory instruments are currently underway at a national and international level (Turner, 2018). As AI makes its way into the practice of psychiatry, it falls within the scopes of medical research regulation.[10] Therefore, the lawful implementation of ML in medical research will depend upon jurisdictional research governance frameworks. This article refers to the regulatory framework in England and Wales.

Balancing the duty to generate new knowledge with the interests of research participants has been at the core of the efforts to regulate medical research since the WMA Declaration of Helsinki.[11] This is particularly relevant when those who are targeted by research programmes are in a condition that may increase their vulnerability: young people, clinical populations, and people with enduring and severe mental illness. In order to ensure the lawful conduct of ML research in the context of psychosis prediction, I argue that two areas of regulation are of particular relevance: (1) protection of research participants, and (2) privacy and data protection.

3.1. Risks, benefits, and protection of research participants

Protection of participants in clinical research is generally intended as protection from potential harms that may be disproportionate to the benefits of the research. Informed consent is meant to ensure that participants are aware of the objectives the research, and of potential harms and benefits. Within the UK regulatory framework, protection of research participants and informed consent procedures depend on two elements: (i) *who* are the research subjects—with reference to the legal age of competence and to the capacity to consent—and, (ii) what *type* of research is performed—whether this is general healthcare research or a clinical trial.

In England and Wales, Research Ethics Committees (RECs) are responsible for evaluating the risk-benefit ratio of research studies.[12] It is important to highlight one principle. When RECs evaluate the risk-benefit ratio of a research study—especially in the case of vulnerable populations—such evaluation is performed with reference to research *procedures*; in other words, to what will happen to research subjects as a result of their participation. This is evident in the discussion around the "minimal risk" threshold in research with children, which is often interpreted as the risk that is "[…] ordinarily encountered in daily life or during the performance of routine physical and psychological examinations or tests".[13] The use of ML constitutes in itself a research procedure. However, ML is a set of techniques for data *analysis*. Therefore, the assessment of risks and benefits must take into consideration whether any additional risks to participants derive directly from the use of ML for data analysis and outcome prediction, in addition to the risks posed by data acquisition procedures. Let us try to unfold what the implications of this principle might be in the jurisdiction of England and Wales.

First, let us consider the case of adults who retain capacity to consent. The case of David Reynolds, which I introduced earlier, may provide a good example of the legal challenges of implementing ML-mediated prediction of harm in individuals with psychosis. David Reynolds was an adult and retained his capacity to consent to hospital admission for medical treatment—consent which he, in fact, had given.[14] In principle, there appear to be no reasons why he should not be allowed to consent to (at least) therapeutic medical research. However, few would deny that Mr. Reynolds was in a vulnerable condition: he was actively psychotic, had ongoing suicidal ideation, and was seeking help. Should we consider including someone like Mr. Reynolds in a research study in order to track his behaviour and establish his risk to commit a suicide attempt? From an ethical point of view, we may recognise the presence of two moral duties: the duty to protect the subject's life,[15] and the duty to conduct research with vulnerable individuals in order to improve suicide prevention. At the same time, the primary aim of clinical research regulation is to protect participants from disproportionate risks and burdens. How do we minimize *risk*, when the aim of a research programme is to investigate someone's risk (likelihood) to commit self-harm, or even a suicide attempt?

It can be argued that the use of wearables, digital phenotyping, or the collection of speech data poses minimal risks in terms of data acquisition. As Martinez-Martin et al. argue: "the collection of digital data is ostensibly of relatively low risk, as it consists of the same activities an individual would otherwise engage in."[16] The same could be said with reference to the use of data collected via clinical assessment or clinical records. It is less clear whether data acquisition via neuroimaging poses significant risks in the context of psychosis, though it has been argued that the use of neuroimaging with vulnerable populations, such as children, can be classified as minimal risk (Holland et al., 2014). Does the use of ML for data analysis pose any *additional* risks? Recent literature has highlighted two relevant sets of risks related to the use of ML in health care: risks to privacy and data protection, and the risk of algorithmic bias[17] (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016; Vayena, Blasimme, & Cohen, 2018). I shall briefly address the issue of data protection later in this article. Here, to answer our original question, I argue that in order to minimize risk in ML research on self-harm in psychosis, researchers should: (1) employ low-risk data acquisition techniques; (2) ensure that appropriate data protection protocols are in place; and (3) minimize algorithmic bias in ML design. Further, I argue that such principles should not be confined to the ML research on self-

[8] Shatte et al. (2019), p. 1438.

[9] ibid., p. 1434.

[10] See Turner (2018). Building a Regulator. In *Robot Rules. Regulating Artificial Intelligence* (pp. 207–262): Palgrave Macmillan.

[11] As expressed in Article 8 of the Declaration, "while the primary purpose of medical research is to generate new knowledge, this goal can never take precedence over the rights and interests of individual research subjects", World Medical Association (1964), Declaration of Helsinki. Ethical Principles for Medical Research Involving Human Subjects, VII revision, 2013.

[12] Where there is no investigation of medical products or devices—which in England and Wales is regulated by the Medicines and Healthcare Products Regulatory Agency (MHRA)—RECs are the *only* agencies that evaluate the risk-benefit ratio of a research study.

[13] This wording is taken from the US Federal Regulations as reported in Kopelman (2004), p. 360. See also Shah, Whittle, Wilfond, Gensler, and Wendler (2004).

[14] *Reynolds v. The United Kingdom* (2012) 55 EHRR 35.

[15] As recognised also by the ECtHR in *Reynolds v. The United Kingdom* (2012) 55 EHRR 35.

[16] Martinez-Martin, Insel, Dagum, Greely, and Cho (2018), p. 68.

[17] See also "Accuracy", in Information Commissioner's Office (2017), pp. 43–45.

harm and suicidality. They may be applied to ML research which aims to identify psychosis risk or to predict psychosis transition or relapse in adults who retain capacity to consent.

Second, let us consider the case of minors and children. Early intervention services in England and Wales target people aged 14 to 65.[18] It is thus a possibility that minors are asked to participate in ML research, particularly with reference to prediction of psychosis *transition* in HR individuals, or prediction of self-harm. English Law has established clear guiding principles for the conduct of research with minors.[19] In the case of healthcare research, *Gillick* competent minors—and legal representatives of other minors— may consent to research that produces no direct benefits if this is not against the child's best interests.[20] Unless ML procedures are shown to be against a child's best interest, the consent of Gillick competent minors (and of legal representatives of other minors) should be sufficient. The situation is more complicate with regard to clinical trials. Clinical trials conducted in England and Wales fall under the Medicines for Human Use (Clinical Trials) Regulations 2004.[21] However, this regulation may soon be replaced by the Clinical Trials Regulation EU No 536/2014.[22] The new general rule will be as follows: the informed consent of the minor's legal representative must be obtained[23]; "the clinical trial either relates directly to a medical condition from which the minor concerned suffers or is of such a nature that it can only be carried out on minors"[24]; a *direct benefit* for the minor can be expected, *or* some benefit for the *population* represented can be expected *and* the trial only poses *minimal risk.*[25]

We can argue that clinical trials that use ML to enhance the identification of psychosis risk or to ameliorate prediction of psychosis transition *should*—in theory—be able to offer the prospect of a *population benefit*; their rationale being precisely to ameliorate diagnostic procedures and prediction in the context of psychosis. Should they be classified as minimal risk? Let us look at data acquisition. Again, there is evidence to believe that neuroimaging poses minimal risks to children (Holland et al., 2014); the same is valid for data collection via clinical assessment, clinical records, and digital health applications. What about ML data analysis procedures? As I argued above, it will be important to assess whether the use of ML for data analysis might pose any *additional* risks to the minors involved in the clinical trial. In England and Wales, it will be RECs' responsibility to establish whether ML procedures expose minors to risks that are more than minimal, again especially with regard to data protection and the risk of algorithmic bias.

Third, let us consider the case of adults who lack mental capacity. It should not be assumed that individuals who suffer from psychosis lack the capacity to consent to research (Spencer, Gergel, Hotopf, & Owen, 2018). At the same time, it must be acknowledged that a clinical history of severe mental illness might affect capacity (Appelbaum, 2005). Therefore, it is not unlikely that ML research on prediction of psychosis relapse or self-harm in individuals with schizophrenia might involve people who lack capacity to consent. Here, English Law is more complicated than in the case of minors. Healthcare research with incapacitated subjects is governed in England and Wales by the Mental Capacity Act 2005 (MCA). Under the MCA, the research must have the potential to *benefit* the subject without imposing a disproportionate burden, *or* be intended to provide knowledge about the condition from which the incapacitated subject is affected.[26] In the latter case, the *risk* must be *negligible,*[27] and research procedures should not be *unduly* invasive or restrictive.[28] With regard to the criterion of 'benefit' to the subject, the MCA Code of Practice specifies that potential benefits include "developing more effective ways of treating a person or managing their condition" or "reducing the risk of the person being harmed […]"[29] In theory, it seems that ML research that aims to improve prediction of psychotic relapse and self-harm could satisfy this requirement. With regard to the criterion of 'negligible risk', we can refer to the reflections presented above regarding the minimal risk threshold in minors.

In the case of clinical trials, the new EU Clinical Trials Regulation will establish the same risk-benefit criteria discussed above for minors.[30] However, until the new Regulation becomes applicable, clinical trials with incapacitated subjects in England and Wales will fall under the Medicines for Human Use (Clinical Trials) Regulations 2004. Under these Regulations, informed consent must be obtained from the subject's legal representative,[31] and there must be 'grounds' to think that the trial will produce *direct benefit* to the individual, or *no risk at all.*[32] RECs will have to consider whether a prospect of direct benefit can be established from ML-mediated prediction of psychosis relapse or self-harm. At the same time, it might be more difficult to demonstrate that a clinical trial poses 'no risk at all' for incapacitated participants, considering the risks to data protection and the risk of algorithmic bias mentioned above.

To summarise: what additional risks may result from the use of ML for data analysis and outcome prediction? In relation to which populations? How should researchers and RECs try to minimize these risks and maximize the benefits of ML-mediated prediction in the context of psychosis? Answering these questions could ensure that research studies prioritize the rights and interest of participants over the duty to produce new knowledge, as established by the WMA Declaration of Helsinki.

3.2. *Privacy and data protection*

The development of AI in healthcare poses significant challenges to privacy and data protection (Information Commissioner's Office, 2017; Vayena et al., 2018). It is beyond the scope of this paper to provide a comprehensive account of these challenges. However, it is important to mention some recent developments in the European data protection framework. The General Data Protection Regulation 2016/679 (GDPR),[33] which was incorporated into English Law with the Data Protection Act 2018, can be considered the first European Regulation to deal explicitly with data processing by means of AI. The GDPR is

[18] See NHS England (2016), Implementing the Early Intervention in Psychosis Access and Waiting Time Standard: Guidance.

[19] See Brazier and Cave (2016). Young people in medical research programmes. In *Medicine, Patients and the Law* (VI ed.) (pp. 485–489). Manchester: Manchester University Press.

[20] See Brazier and Cave (2016), ibid.

[21] According to the 2004 Regulations, the age of consent is set at 16 years; informed consent by a parent or a legal representative is required for minors (Part 4, para 4); and some direct benefit for the group of patients involved in the trial must be obtained (Part 4, para 10).

[22] The new Regulation will be applied after the development of a fully functional EU clinical trials portal and database, which is currently estimated to occur in 2020. At the time of writing, it is difficult to predict what will be the consequences of Brexit on the application of the Regulation in the United Kingdom. Also, note that the 'direct benefit for the group of patients involved' criterion established by the 2004 English Regulations is more restrictive than the '*population* benefit' required under the new EU Regulation.

[23] Article 32 (1) (a), Clinical Trials Regulation EU No 536/2014.

[24] ibid., article 32 (1) (f).

[25] ibid., article 32 (1) (g).

[26] Mental Capacity Act 2005, section 31(5).

[27] ibid., section 31(6)(a).

[28] ibid., section 31(6)(b)(ii).

[29] Department of Constitutional Affairs (2013), Mental Capacity Act 2005 Code of Practice, chap. 11.14, p. 207.

[30] See Article 31, Clinical Trials Regulation EU No 536/2014.

[31] Medicines for Human Use (Clinical Trials) Regulations 2004, schedule 1, part 5.

[32] ibid., condition (1)9.

[33] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

relevant to psychosis prediction for two reasons. First, it sets out the legal framework for processing health-related data. Second, it establishes some important limitations to *automated* data processing.

Under GDPR, the level of data protection depends upon which type of data is being processed. At article 4, the GDPR distinguishes among: (i) genetic data, (ii) biometric data, and (iii) data concerning health. Neuroimaging data could be classified as either biometric or as data concerning health, though it seems likely that data obtained for the purpose of *clinical* prediction shall be classified as data concerning health.[34] How will clinical assessment data and behavioural data collected via wearables and digital phenotyping be classified? Article 4 specifies that: “data concerning health means personal data related to the physical and mental health of a natural person, […] which reveal information about his or her health status”.[35] Therefore, we can assume that clinical assessment and behavioural data collected for clinical prediction constitute data concerning health. Article 9 clarifies that genetic data, biometric data, and data concerning health constitute ‘special-category personal data’, and are thus subject to specific rules for processing. Processing of special-category personal data is permitted, among other conditions, in the case of: (a) explicit consent of the data subject, (c) to protect the *vital interest* of the data subject or another natural person, where the data subject is not able to give consent, (g) for reasons of ‘substantial public interest’, (h) for the purposes of ‘preventive medicine’ and ‘medical diagnosis’, and for (j) research purposes.[36]

Under article 8(2) ECHR, the right to respect for private life is not absolute as exceptions are permitted, including “for the prevention of disorder or crime” and “for the protection of health and morals”. The GDPR, however, establishes some important limitations to the use of automated decision-making in this regard. First, the data subject must be informed of the existence of automated decision making regardless of whether or not the data are collected from him or her.[37] Second, article 22(1) recognizes “[…] the right not to be subject to a decision based *solely* on automated processing including profiling, which produces legal effects [concerning the data subject]”. Even though exceptions to this rule are possible, in the case of special-category personal data these exceptions are admissible only with the explicit consent of the data subject, *or* “for reasons of substantial public interest”.[38]

Some considerations can thus be drawn on how ML will be used in the context of psychosis prediction. First, it appears that data collected for the purpose of clinical prediction—be it prediction of psychosis transition, relapse, or harm—will be classified as ‘special-category personal data’ and be subject to specific rules for processing, both in research and in clinical contexts. Second, the data subject will retain the right to be informed of the existence of ML procedures, as these constitute automated processes. Third, and most importantly, it appears that the data subject will retain the right to object to decisions based *solely* on the outcomes of ML algorithmic applications. It might not be possible for clinicians, for instance, to delegate decisions following confirmation of risk status or outcome prediction to fully automated ML applications. Clinicians and public authorities will retain legal responsibility for the decisions they take concerning a data subject, following identification of risk status or outcome prediction via ML applications.

4. Psychiatric assessment and ML-mediated prediction

Even though ML applications are currently mostly confined to clinical research (Shatte et al., 2019), it is important to anticipate how they might shape mental health care in the near future. Many civil and common-law jurisdictions recognise *risk* criteria for compulsory admission or treatment of the mentally ill (de Stefano & Ducci, 2008; Ryan, Nielssen, Paton, & Large, 2010). As we move towards clinical prediction in mental health, the use of AI will likely affect the practice of *psychiatric assessment* under mental health legislation. In the next pages, I investigate how ML might affect the practice of psychiatric assessment and determination of risk in the context of psychosis, with reference to English Law.

The *logic of risk* is apparent under English Law (Fanning, 2016; Glover-Thomas, 2011). Compulsory hospital admission and treatment for the mentally ill is governed in England and Wales by the Mental Health Act 1983 (MHA).[39] Individuals who suffer from a mental disorder may be detained in hospital for assessment followed by treatment if, inter alia, this is “necessary for the health or safety of the person or for the protection of other persons […]”.[40] Risk assessment is particularly relevant at patient discharge. Under the MHA, individuals who have been detained under Section 3, for example, may be subject to a community treatment order if this is necessary for “preventing risk of harm to the patient's health or safety[41]” or for “protecting other persons[42]”. As elegantly phrased by the ECtHR, mental health legislation deals with the “unpredictability of human conduct[43]”. English Law recognizes *risk of harm* to self or others as the main criterion for civil detention or compulsory treatment of the mentally ill. The discriminatory nature of such practice is being contended in light of the United Nations Convention on the Rights of Persons with Disabilities (Szmukler, 2017). Researchers are calling for a removal of risk criteria from mental health legislation.[44] This article does not address the issue of whether risk criteria are discriminatory. To date, such criteria remain part of our statutes. Rather, this article investigates how prediction via AI might affect our ability to manage the ‘unpredictability of human conduct’ in people who suffer from psychosis, within the context of current mental health law.

4.1. On risk

Until now, we have discussed the use of ML around three areas: (i) the identification of psychosis risk and the prediction of psychosis *transition* in HR individuals; (ii) the prediction of psychosis *relapse* in individuals who suffer from a psychotic disorder; and (iii) the prediction of self-harm and suicidality in the context of psychosis. As we have seen, Shatte et al. (2019) significantly include prediction of harm in their scoping review of ML applications in mental health. Some nuanced distinctions are necessary. The field of clinical prediction and the practice of risk assessment under mental health law have something in common. This is the concept of *risk* and the attempt to predict future behaviour. Nonetheless, establishing psychosis risk and assessing a person's risk of harm are discrete practices, though they may occasionally overlap. Here, I argue that the notion of *psychosis risk* and the notion of *risk of harm* are different in nature.

Psychosis risk is a clinical notion. In the past twenty years, this notion has emerged to identify (young) individuals who have not yet experienced a first episode of psychosis, but whose behavioural deterioration has reached a threshold that warrants clinical attention. The concept of psychosis risk has been operationalised in several

[34] For instance, the UK Biobank GDPR notice classifies data collected, including neuroimaging data, as health-related data, see https://www.ukbiobank.ac.uk/gdpr/, last accessed 6 July 2019.

[35] Regulation (EU) 2016/679, article 4(15).

[36] ibid., article 9(2).

[37] ibid., article 13(2)(f) and article 14(2)(g).

[38] ibid., article 22(4).

[39] As amended by the Mental Health Act 2007.

[40] Mental Health Act 1983, Part II, section 3(2)(c).

[41] Mental Health Act 1983, section 17B(2)(b).

[42] Mental Health Act 1983, section 17B(2)(c).

[43] *Osman v. United Kingdom* (2000) 29 EHRR 245 at para 116.

[44] See also the recent Report of the Special Rapporteur on the rights of persons with disabilities A/HRC/40/54, released by the United Nations General Assembly Human Rights Council on 11 January 2019, available at https://undocs.org/A/HRC/40/54, last accessed 6 July 2019.

denominations, including the HR state (Fusar-Poli et al., 2013). Moreover, amid a rather heated debate (Yung, Nelson, Thompson, & Wood, 2010), the latest edition of the DSM-5 eventually included the 'attenuated psychosis syndrome', whose criteria strongly resemble the ones of the psychosis HR state.[45] As a clinical notion, the concept of psychosis risk represents someone's likelihood to transition to—that is, to develop—a psychotic episode. Linked to the notion of psychosis risk is the occurrence of further psychotic episodes after a period of remission, which is usually called *psychosis relapse* (Emsley, Chiliza, Asmal, & Harvey, 2013).

Risk of harm is a broader, complex ethico-legal notion. It refers to the harm that may derive from a mental disorder. The MHA Code of Practice defines risk of harm as "risk of suicide, self-harm, self-neglect, […] jeopardising one's own health or safety accidentally, recklessly, or unintentionally, or […] otherwise put one's health or safety at risk[46]". In addition, according to the Code, risk of harm includes potential harm to other people.[47] Whilst psychosis risk can be conceptualised as the risk to develop a certain mental state (or illness), risk of harm can be conceptualised as the risk that harmful *consequences* may occur because of (the presence of) a mental state. In other words, the concept of risk of harm represents someone's likelihood to cause harm to herself or to a third party.

Using ML to establish psychosis risk is *not* the same as using ML to establish risk of harm. Interestingly, the MHA does not define 'risk' (Fanning, 2016). As outlined above, the definition provided by the Code of Practice is sufficiently loose to include any risks to the health of a person. It could be argued that psychosis transition in HR individuals, or psychosis relapse in individuals who suffer from schizophrenia might indeed 'jeopardize their health'. This is particularly true in the context of psychiatric assessment under the MHA. However, would this be sufficient to establish the presence of 'risk of harm'? The extent to which psychosis risk and risk of harm may overlap depends on whether we can consider psychosis as harmful per se. Does psychosis risk constitute a risk of harm to self/others and therefore qualify as a basis for intervention? Even though there are good reasons to believe that intervening early in the clinical course of psychotic illness is beneficial, there are also good reasons to believe that psychosis and psychotic illness are not the same (Read & Dillon, 2013). People who have psychotic experiences—especially auditory hallucinations—greatly outnumber the ones who develop a psychotic disorder.[48] Therefore, we cannot consider psychosis as harmful per se. As AI ameliorates our ability to predict a psychotic episode, using psychosis risk as a criterion to justify coercion—if coercion can ever be justified—without any indication that harmful behaviour is likely to occur, could potentially have the *paradoxical* effect of harming the individual while trying to prevent harm. Psychosis risk and risk of harm are discrete entities, and should remain so. This principle may be valid regardless of the tools used to predict psychosis—whether this is done via clinical interview or via ML. However, it will be important to reaffirm this principle when translating ML applications into psychiatric assessment, in order to minimize the risk of algorithmic bias.

4.2. On prediction, algorithms, and coercion

The use of ML in psychiatry opens up the possibility to predict certain events and clinical outcomes. However, what exactly are we predicting when we use ML in the context of psychosis? Let us focus on four possible scenarios: (i) psychosis transition; (ii) psychosis relapse; (iii) self-harm and suicide; and (iv) harm to others.

First, predicting *psychosis transition* means that we are predicting that (young) high-risk individuals, who have never had psychosis, will develop a first psychotic episode (Koutsouleris et al., 2015; Rezaii et al., 2019). It has long been established that individuals who experience a first psychotic episode should be offered participation in youth-friendly, community-based early intervention services (Corsico, Griffin-Doyle, & Singh, 2018). In this case, coercion would most likely *generate harm*, not prevent it.

Second, predicting *psychosis relapse* means that we are predicting that individuals, who already suffer from a psychotic disorder including schizophrenia, will relapse into psychosis, with potentially harmful consequences for their health. Digital phenotyping is particularly promising in this regard (Barnett et al., 2018; Torous, Barnett, et al., 2018). Prediction of psychosis relapse could help to improve prognosis of severe mental illness. There is little doubt that reducing diagnostic uncertainty in psychiatry could produce clinical benefits. Yet, the issue of coercion remains unsolved. The case of *Winterwerp v. The Netherlands* has established that, in order for a deprivation of liberty to be lawful, "objective medical expertise" is needed to determine whether a person is of 'unsound mind', as well as to demonstrate the nature and degree of a "true mental disorder".[49] Relapse prediction could in fact support 'objective medical expertise' by providing a reliable estimate of who is likely to relapse into psychosis—and potentially, when. However, the issue of whether psychosis relapse may be per se harmful remains open. Risk of harm to self or others would still have to be established in order to authorize a deprivation of liberty.

Third, predicting *self-harm and suicide* means predicting harm to self or, in extreme cases, a suicide attempt (Just et al., 2017; Torous et al., 2018). In a recent article, Walsh et al. reported that they were able to improve the accuracy of suicide-attempt prediction from 720 days to 7 days before the suicide attempt, by applying a ML algorithm to electronic health records (Walsh et al., 2017). Again, there is little doubt that improving the accuracy of prediction would be beneficial to suicide prevention. The case of Carol Savage,[50] a woman who died by suicide while detained for treatment under Section 3 of the MHA, established that there is an *operational duty* to protect the right to life of a detained patient under Article 2(1) of the ECHR where the hospital knows, or ought to know, of a real and immediate risk to life. The UK Supreme Court has further held in *Rabone v. Pennine Care NHS Foundation Trust*[51] that this operational duty may be owed also to informal patients (Allen, 2013). Prediction of self-harm and suicidality—whether or not it is achieved via AI—if possible, timely, and accurate would support the operational duty to protect life, which is owed to detained and informal patients by the clinical institutions responsible for their care.

Lastly, predicting *harm to others* means predicting that a person will cause harm to someone else because of their mental illness. In *Osman v. United Kingdom*, the ECtHR held that public authorities must know of a "real and immediate risk to life" of a person by a third party in order to proceed with preventive measures.[52] Among others, Large et al. argue that prediction of dangerousness has very modest scope in preventing violence, and that it unfairly discriminates against the mentally ill (M.

[45] See American Psychiatric Association (2013). Other Specified Schizophrenia Spectrum and Other Psychotic Disorder. In *Diagnostic and Statistical Manual of mental disorders (DSM-5)* (s. II, p.122).

[46] Department of Health (2015). *Mental Health Act 1983: Code of Practice* (p. 114).

[47] ibid., p. 115.

[48] See for instance Bentall, R. (2013). Understanding psychotic symptoms. Cognitive and integrative models. In J. Read, & J. Dillon, *Models of madness. Psychological, social and biological approaches to psychosis* (pp. 220–237) (2nd ed.). London & New York: Routledge.

[49] *Winterwerp v. The Netherlands* (A/33) (1979–80) 2 EHRR 387, at para 39.

[50] *Savage v. South Essex Partnership NHS Foundation Trust* [2010] EWHC 865 (QB).

[51] *Rabone v. Pennine Care NHS Foundation Trust* [2012] UKSC 2, at para 22: "[…] the operational duty will be held to exist where there has been an assumption of responsibility by the state for the individual's welfare and safety (including by the exercise of control)." See also *Fernandes de Oliveira v. Portugal* [2019] ECHR 106.

[52] *Osman v. United Kingdom* (2000) 29 EHRR 245, at para 116.

M. Large, Ryan, Nielssen, & Hayes, 2008).[53] The authors argue that even the best violence prediction tool has very limited utility. They make a strong case about the discriminatory nature of dangerousness criteria:

> Those accused of a violent crime are deemed innocent until proven guilty, and the state must prove their guilt beyond reasonable doubt. "Better that ten guilty persons escape than that one innocent suffers." Very few statutes permit the incarceration of innocents merely because they might harm others in the future.[54]

Large et al. argue that there are reasons to doubt that accurate prediction of harm to others is even possible. Yet, we cannot predict what AI-mediated prediction might look like in the future. It is indicative that in their scoping review of current ML applications in mental health, Shatte et al. did not report any category which refers to prediction of 'harm to others' (though they did include self-harm and suicide). This could mean that efforts to predict harm to others via ML are *not* currently underway; or, it could mean that prediction of harm others via ML is perceived to transcend the scope of clinical prediction. As 'harm to others' is a *social* occurrence, its prediction might intrinsically differ from efforts to predict psychosis transition, psychosis relapse, or self-harm via ML. Whether prediction of harm to others via ML might be possible depends on the level of surveillance that our societies will decide to accept. Its role in psychiatric coercion will depend on the broader issue of whether risk of harm to others can (ever) justify preventive detention of the mentally ill.

5. David Reynolds and the risks of risk

Could AI have saved David Reynolds? Mr. Reynolds had been diagnosed with schizophrenia seven years before his death. It was clear that he was having a psychotic relapse on the day of his death. However, he had no history of suicide attempt. He had told the psychiatrist that he did not want to kill himself and was assessed to be a low suicide risk. Was he really at low risk of self-harm or suicide? It is not clear what Mr. Reynolds' intentions were when he fell from the window.[55] It is also not possible to say whether using AI to monitor his behaviour could have saved his life. However, the ECtHR recognised that "[…] an operational duty arose to take reasonable steps to protect him from a real and immediate risk of suicide and that that duty was not fulfilled".[56] Should reliable ML applications that can predict a suicide attempt be available in the future, would it be a 'reasonable step'—using the language of the ECtHR—for a clinical team to use such applications with psychotic inpatients? If no, for what reasons this would be unreasonable? Potential issues in ML design, such as privacy concerns and the risk of algorithmic bias described above, might constitute some of those reasons. However, it is difficult to answer such questions at this stage. Nonetheless, we can argue that digitally tracking Mr. Reynolds' behaviour to assess his risk of suicidality might have helped the clinical team to fulfil their *operational duty* to protect Mr. Reynolds' right to life. ML-mediated prediction of suicidality, if possible, timely, and accurate, could have produced real benefits for him, in light of the operational duty to protect his right to life as a voluntary inpatient.

The case of David Reynolds suggests that there are situations where ML-mediated prediction may be beneficial to psychiatric assessment. At the same time, there are risks involved in the logic of risk.

Most of our mental health law frameworks recognise—or, some might say, establish (Rose, 1998)—a dialectic tension between *freedom* and *security*. Freedom here is intended as the liberty of individuals who suffer from mental illness not to be subject to compulsory detention and treatment. Security is intended as the protection from potential harm, which (allegedly) justifies State action in enforcing surveillance, coercion, and detention of the mentally ill. The logic of risk acts as the *medium* between freedom and security. First, the notion of risk mediates between freedom of research—intended as freedom to conduct research as well as liberty to take part in it—and protection of research participants. Second, the notion of risk mediates between the liberty of the mentally ill *not to* be subject to coercion and society and the person's need for security. The risks involved in the logic of risk might be called the *risks of risk*: that in using the logic of risk as a medium between freedom and security, we might fail to make the necessary *nuanced* distinctions between different notions of risk, which serve different purposes. That we might fail to recognise the differences between psychosis risk and risk of harm, and between psychosis prediction and prevention of harm. That in trying to minimize risk and prevent harm we may instead cause harm by means of unnecessary surveillance, coercion, or detention. I argue that in order to maximize the benefits that derive from the expansion of AI into psychiatry we ought to be aware of and minimize the risks of risk.

6. Conclusions

This article has claimed that there is a potential for virtuous applications of ML-mediated prediction in mental health. It has suggested that we may have an obligation (a duty?) to promote the benefits that derive from using ML for prediction of psychosis and self-harm, while at the same time avoid the downsides of unnecessary coercion and surveillance. First, I have argued that ML could effectively support the practice of psychiatric assessment. There is little reason to doubt that using ML to improve accuracy in predicting psychosis transition or psychosis relapse could benefit patients and service users. However, before this happens, researchers and RECs ought to ensure that the use of ML applications in clinical research respects the rights and interests of research participants. This includes acknowledging any additional risks posed to participants by the use of ML for data analysis and outcome prediction. It also includes minimizing potential risks of research participation, which vary depending on the populations that researchers wish to recruit, their age, and their capacity to consent. Second, I have argued that the notion of *psychosis risk* and the notion of *risk of harm* are different in nature. Establishing someone's risk to transition to, or to relapse into psychosis is not the same as assessing risk of harm. The two practices are distinct, and so should remain when using ML for psychiatric assessment. Third, I have argued that reducing diagnostic uncertainty in psychiatric assessment could benefit the mentally ill. However, using psychosis prediction as a criterion to justify coercion could potentially harm them. The extent to which the use of ML-mediated prediction might benefit or harm individuals who experience psychosis will depend on: (i) which type of *risk* we are assessing, whether this is psychosis risk or risk of harm, and (ii) what we are trying to predict, whether this a psychosis transition, a psychotic relapse, self-harm and suicidality, or harm to others.

Acknowledgements

This article was developed as part of the author's work as a doctoral candidate in Bioethics and Medical Jurisprudence at the Centre for Social Ethics and Policy, The University of Manchester. The author has received a doctoral scholarship from the former School of Law, now School of Social Sciences, Department of Law, The University of Manchester to complete this work. The author thanks his supervisors, Prof Søren Holm and Mr. Neil Allen for their invaluable support in

[53] It must be noted that M. Large and C. Ryan also criticise the use of *risk* criteria in the context of suicidality and psychiatric coercion, see their commentary on the *Rabone* case, M. Large, Ryan, and Callaghan (2018).

[54] M. M. Large et al. (2008), p. 879.

[55] See *Reynolds v. The United Kingdom* (2012) 55 EHRR 35, at para 16: the applicant, David Reynolds' mother Mrs. Patricia Reynolds, "[…] considered that her son had not attempted to commit suicide but rather had wished to go home and had not realised he was on the sixth floor."

[56] *Reynolds v. The United Kingdom* (2012) 55 EHRR 35, para 61.

preparing and revising the manuscript.

Declaration of Competing Interests

The author is a recipient of a doctoral scholarship from the former School of Law, now School of Social Sciences, Department of Law, The University of Manchester. He also holds an appointment as a part-time research assistant at the Department of Psychiatry, University of Oxford. The author declares no conflict of interest.

References

Allen, N. (2013). The right to life in a suicidal state. *International Journal of Law and Psychiatry, 36*, 350–357.

Alpaydin, E. (2016). *Machine learning: The new AI.* Cambridge, Massachusetts: The MIT Press.

Appelbaum, P. S. (2005). Decisional capacity of patients with schizophrenia to consent to research: Taking stock. *Schizophrenia Bulletin, 32*, 22–25.

Barnett, I., Torous, J., Staples, P., Sandoval, L., Keshavan, M., & Onnela, J. P. (2018). Relapse prediction in schizophrenia through digital phenotyping: A pilot study. *Neuropsychopharmacology, 43*, 1660–1666.

Bartlett, P., & Sandland, R. (2014). *Mental health law. Policy and practice* (IV ed.). Oxford: Oxford University Press.

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

Brazier, M., & Cave, E. (2016). *Medicine, patients and the law* (VI ed.). Manchester: Manchester University Press.

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

Corsico, P., Griffin-Doyle, M., & Singh, I. (2018). What constitutes 'good practice' in early intervention for psychosis? Analysis of clinical guidelines. *Child and Adolescent Mental Health, 23*, 185–193.

Dadich, S. (2016). Barack Obama, neural nets, self-driving cars, and the future of the world. https://www.wired.com/2016/10/president-obama-mit-joi-ito-interview/, Accessed date: 20 July 2019.

Emsley, R., Chiliza, B., Asmal, L., & Harvey, B. H. (2013). The nature of relapse in schizophrenia. *BMC Psychiatry, 13*, 50.

Fanning, J. (2016). Continuities of risk in the era of the mental capacity act. *Medical Law Review, 24*, 415–433.

Fernandes, A. C., Dutta, R., Velupillai, S., Sanyal, J., Stewart, R., & Chandran, D. (2018). Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific Reports, 8*, 7426.

Fusar-Poli, P., Borgwardt, S., Bechdolf, A., Addington, J., Riecher-Rossler, A., Schultze-Lutter, F., ... Yung, A. (2013). The psychosis high-risk state: A comprehensive state-of-the-art review. *JAMA Psychiatry, 70*, 107–120.

Gifford, G., Crossley, N., Fusar-Poli, P., Schnack, H. G., Kahn, R. S., Koutsouleris, N., ... McGuire, P. (2017). Using neuroimaging to help predict the onset of psychosis. *NeuroImage, 145*, 209–217.

Glover-Thomas, N. (2011). The age of risk: Risk perception and determination following the Mental Health Act 2007. *Medical Law Review, 19*, 581–605.

Harris, E. C., & Barraclough, B. (1997). Suicide as an outcome for mental disorders. *British Journal of Psychiatry, 170*, 205–228.

Holland, S. K., Altaye, M., Robertson, S., Byars, A. W., Plante, E., & Szaflarski, J. P. (2014). Data on the safety of repeated MRI in healthy children. *NeuroImage: Clinical, 4*, 526–530.

Information Commissioner's Office (2017). *Big data, artificial intelligence, machine learning, and data protection*.

Jain, S. H., Powers, B. W., Hawkins, J. B., & Brownstein, J. S. (2015). The digital phenotype. *Nature Biotechnology, 33*, 462–463.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology, 2*, 230–243.

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D. L., 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.

Kempton, M. J., & McGuire, P. (2015). How can neuroimaging facilitate the diagnosis and stratification of patients with psychosis? *European Neuropsychopharmacology, 25*, 725–732.

Kopelman, L. M. (2004). Minimal risk as an international ethical standard in research. *Journal of Medicine and Philosophy, 29*, 351–378.

Koutsouleris, N., Riecher-Rossler, 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*, 471–482.

Large, M., Ryan, C. J., & Callaghan, S. (2018). Hindsight bias and the overestimation of suicide risk in expert testimony. *The Psychiatrist, 36*, 236–237.

Large, M. M., Ryan, C. J., Nielssen, O. B., & Hayes, R. A. (2008). The danger of dangerousness: Why we must remove the dangerousness criterion from our mental health acts. *Journal of Medical Ethics, 34*, 877–881.

Martinez-Martin, N., Insel, T. R., Dagum, P., Greely, H. T., & Cho, M. K. (2018). Data mining for health: staking out the ethical territory of digital phenotyping. *npj Digital Medicine, 1*, 68.

McGuire, P., Howes, O. D., Stone, J., & Fusar-Poli, P. (2008). Functional neuroimaging in schizophrenia: Diagnosis and drug discovery. *Trends in Pharmacological Sciences, 29*, 91–98.

Mechelli, A., Lin, A., Wood, S., McGorry, P., Amminger, P., Tognin, S., ... Yung, A. (2017). Using clinical information to make individualized prognostic predictions in people at ultra high risk for psychosis. *Schizophrenia Research, 184*, 32–38.

Meynen, G. (2013). A neurolaw perspective on psychiatric assessments of criminal responsibility: Decision-making, mental disorder, and the brain. *International Journal of Law and Psychiatry, 36*, 93–99.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*, 1–21.

O'Neill, C. (2016). *Weapons of Math Destruction. How big data increases inequality and threatens democracy.* Penguin Books.

Read, J., & Dillon, J. (2013). *Models of madness. Psychological, social and biological approaches to psychosis* (2nd ed.). London & New York: Routledge.

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. *npj Schizophrenia, 5*, 9.

Rose, N. (1998). Governing risky individuals: The role of psychiatry in new regimes of control. *Psychiatry, Psychology and Law, 5*, 177–195.

Ryan, C., Nielssen, O., Paton, M., & Large, M. (2010). Clinical decisions in psychiatry should not be based on risk assessment. *Australasian Psychiatry, 18*, 398–403.

Shah, S., Whittle, A., Wilfond, B., Gensler, G., & Wendler, D. (2004). How do institutional review boards apply the federal risk and benefit standards for pediatric research? *JAMA, 291*, 476–482.

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*, 1426–1448.

Spencer, B. W. J., Gergel, T., Hotopf, M., & Owen, G. S. (2018). Unwell in hospital but not incapable: Cross-sectional study on the dissociation of decision-making capacity for treatment and research in in-patients with schizophrenia and related psychoses. *British Journal of Psychiatry, 213*, 484–489.

Spranger, T. M. (2012). *International Neurolaw. A comparative analysis.* Berlin, Heidelberg: Springer.

de Stefano, A., & Ducci, G. (2008). Involuntary admission and compulsory treatment in Europe: An overview. *International Journal of Mental Health, 37*, 10–21.

Szmukler, G. (2017). *Men in white coats. Treatment under coercion.* Oxford: Oxford University Press.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*, 751–752.

Torous, J., Barnett, I., Staples, P., Sandoval, L., Onella, J. P., & Keshavan, M. (2018). Towards digital phenotyping for relapse prediction in schizophrenia. *Early Intervention in Psychiatry, 12*, 40.

Torous, J., Larsen, M. E., Depp, C., Cosco, T. D., Barnett, I., Nock, M. K., & Firth, J. (2018). Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: A review of current progress and next steps. *Current Psychiatry Reports, 20*, 51.

Turner, J. (2018). *Robot rules.* Regulating Artificial Intelligence: Palgrave Macmillan.

Valli, I., Marquand, A. F., Mechelli, A., Raffin, M., Allen, P., Seal, M. L., & McGuire, P. (2016). Identifying individuals at high risk of psychosis: Predictive utility of support vector machine using structural and functional MRI data. *Frontiers in Psychiatry, 7*, 52.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine, 15*, e1002689.

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

Xu, Z., Muller, M., Heekeren, K., Theodoridou, A., Metzler, S., Dvorsky, D., ... Rusch, N. (2016). Pathways between stigma and suicidal ideation among people at risk of psychosis. *Schizophrenia Research, 172*, 184–188.

Yung, A. R., Nelson, B., Thompson, A. D., & Wood, S. J. (2010). Should a "Risk Syndrome for Psychosis" be included in the DSMV? *Schizophrenia Research, 120*, 7–15.