Commentary

OPEN ACCESS

# AI-enabled suicide prediction tools: ethical considerations for medical leaders

Daniel D'Hotman,[1] Erwin Loh ,[2,3] Julian Savulescu[1]

[1]Oxford Uehiro Centre for Practical Ethics, Oxford University, Oxford, United Kingdom
[2]Monash Centre for Health Research and Implementation, Monash University, Melbourne, Victoria, Australia
[3]Group Chief Medical Officer, St Vincent's Health Australia Ltd, East Melbourne, Victoria, Australia

**Correspondence to**
Dr Daniel D'Hotman, Oxford Uehiro Centre for Practical Ethics, Oxford University, Oxford OX1 1PT, Oxfordshire, UK; daniel.dhotman@philosophy.ox.ac.uk

Received 27 April 2020
Revised 21 July 2020
Accepted 2 August 2020
Published Online First 23 September 2020

**To cite:** D'Hotman D, Loh E, Savulescu J. *BMJ Leader* 2021;**5**:102–107.

## INTRODUCTION

Suicide accounts for 1.5% of deaths worldwide, with over 800 000 deaths from suicide annually.[1 2] Over 80% of suicides occur in low-income and middle-income countries. While significant work is taking place to reduce the impact of suicide, there is more to be done. In many cases, people at risk of suicide do not engage with their doctor or community due to concerns about stigmatisation and forced medical treatment; worse still, people with mental illness (who form a majority of people who die from suicide) may have poor insight into their mental state, and not self-identify as being at risk. These issues are exacerbated by the fact that doctors have difficulty in identifying those at risk of suicide when they do present to medical services.

In an attempt to reduce the impact of suicide, there is increased interest in using artificial intelligence (AI), data science and other analytical techniques to improve suicide prediction and risk identification. With the proliferation of electronic medical records (EMRs) and online platforms where people share insights on their emotional state (social media), there is now a wealth of relevant health data available to researchers. When linked with other data sources, analysis of these complex sets of information (known colloquially as 'big data') can provide a snapshot of biological, social and psychological state of a person at one time. Machines can learn to detect patterns, which are indecipherable using traditional forms of biostatistics, by processing big data through layered mathematical models (AI algorithms). Correcting algorithm mistakes (training) can improve the accuracy of an AI predictive model.[3] As such, AI is well positioned to address the challenge of navigating big data for suicide prevention.

Broadly, these fall under two categories:

1. *Medical suicide prediction tools:* researchers and doctors can use AI techniques such as machine learning to determine patterns of information and behaviour that indicate suicide risk by leveraging data from EMRs, hospital records and potentially other government data sources. Most typically, these tools would be used in a hospital setting or general practitioner (GP) surgery to provide 'decision support' for doctors when determining a patient's suicide risk. Development of these tools is occurring in traditional research settings with promising results.
   i. *Example*: by applying machine learning to EHRs, Walsh *et al* (2017) achieved 80%–90% accuracy (AUC = 0.80 - 0.84) when predicting whether a suicide attempt was likely to occur within the next 2 years, and 92% accuracy in predicting whether a suicide attempt would occur within the next week.[4] It is important to note that the clinical applicability of these tools in the real world remains unproven; however, initial results are extremely promising.[5]
2. *Social suicide prediction tools:* AI can be used to leverage information from social media and browsing habits to determine suicide risk. These efforts are underway in the scientific community, but are also present in the private sector—for example, Facebook, Google and Apple use data from platforms to determine which users are at risk of suicide and deploy appropriate interventions, such as free information and counselling services. Social suicide prediction tools could theoretically be combined with medical suicide prediction tools in a medical setting (eg, by at risk patients providing access to social media accounts), but up to this point the only practical implementation of these tools has currently been by private sector companies who use algorithms to monitor suicide risk and intervene in a variety of ways. While many companies are actively producing tools to predict suicide, there is little data available on their effectiveness. Facebook, for instance, has not released data on the effectiveness of its tool or its intervention methods—which range from 'soft touch' interventions such as providing information on counselling services, to more intrusive interventions for high-risk cases, such as in the USA, where Facebook staff can call emergency services to an individual's home if there is an immediate risk to life.
   i. *Example*: While Facebook and other companies do not release data on the effectiveness of their tools, some researchers have had promising results in applying social media data to suicide prediction algorithms. A study published in *Biomedical Informatics Insights* by Coppersmith *et al* applied machine learning and natural language processing to social media data from a variety of sources (eg, Facebook, Twitter, Instagram, Reddit, Tumblr, Strava and Fitbit, among others) in order to determine suicide risk of attempted suicide.[6] By linking medical records (which were used to establish whether users actually attempted suicide, rather than to identify risk)—for which they were granted permission by test subjects—Coppersmith *et al* demonstrated that their model

was up to 10 times more accurate at *correctly predicting those 'at risk' of attempting suicide* when compared with clinician averages (4%–6% vs 40%–60%). The AUC was 0.89 to 0.93 for time periods of 1 to 6 months.[6–8]

## AI-DRIVEN SUICIDE PREDICTION INTRODUCES IMPORTANT ETHICAL CONSIDERATIONS

A significant opportunity exists to leverage advances in medical and social suicide prediction tools to improve identification of people at risk of suicide and aid existing investment in suicide prevention. AI may increase our understanding of suicide prediction and potentially save lives. However, it also introduces a number of ethical risks, which are summarised below. These risks have been identified in line with existing guidelines on the ethical use of AI, as documented by the UK Government's 'Data Ethics Framework' and 'Artificial Intelligence: Australia's Ethics Framework'. Discussion of these risks does not encompass all aspects of the ethical issues at play; in the interests of brevity, they are presented here as a means to inform medical leaders who may be considering researching or using these tools within their institutions, and to prompt further exploration and debate.

### Clear public benefit and reduced harms

A widely accepted principle of any AI intervention is that it should entail a clear public benefit. In order to establish efficacy, medical suicide prediction tools should go through extensive peer-reviewed processes in multiple settings and contexts. To some extent, this process is already under way. However, local research should be prioritised to better understand the effect of these tools in specific contexts. While there is some promising emerging evidence for the efficacy of social suicide prediction tools, Facebook, Google and other technology companies have not released data on the effectiveness of their individual efforts. The continued permissibility of these programmes should be contingent on proven efficacy (as established by independent review). Finally, both medical and social suicide prediction tools must be reliable and safe, in line with community and ethical expectations.

AI tools are unlikely to achieve 100% accuracy—as such, it is inevitable that false positives and false negatives will occur. In the case of false positives, people may be exposed to unnecessary treatment or involuntary detention, which may cause psychological harms, as well as potential stigmatisation by members of the public and/or the medical community.[9 10] Such harms can be minimised by recommending that initial interventions are less restrictive on individual rights, and improving public and medical education about mental illness and suicide. The negative effects of a false negative—that is, missing an individual at risk of suicide—are more obvious; however, weighing the cost of missed individual suicide cases must also account for cost of continuing the status quo. Indeed, the problem of false positives and false negatives also exists with human therapists. If AI performs statistically better than humans, the next question is whether a mistake made by AI is 'worse' than one made by human. On initial appraisal, it is difficult to see how mistakes by AI are substantively worse than those made by people, holding all else equal. However, this question is a worthy area of future research.

### Timing of intervention

Given variable efficacy of medical and social AI tools in determining suicide risk, there is a pertinent question as to what level of certainty should be required before intervening. Furthermore, when that threshold is reached, what type of intervention should be deployed?

There are multiple options for intervention that can be broadly classified as soft and hard touch.[9] In a medical context, a soft touch intervention could include answering a suicide questionnaire administered by a GP, while a hard touch intervention for a high-risk patient would involve involuntary admission to hospital with a formal risk assessment by a psychiatrist. For social suicide prediction tools, there are a broad range of interventions available depending on the particular platform that is used. These include: free information about counselling services (Google and Facebook), text-based counselling (Crisis Text Line), online nudges (Facebook's automatic comments on videos of potential self-harm events), online crises intervention (Facebook) and, in cases where an individual's life is deemed to be at immediate risk, intervention by emergency services (Facebook).

The least restrictive alternative is an ethical principle suggesting that an intervention should inflict the smallest rights infringement possible to achieve a specific public health aim.[11 12] The least restrictive alternative is seen to be important because it recognises and respects individual liberty and autonomy. In the context of mental health, we often think of patients as having reduced autonomy, thus justifying infringements on liberty. However, the least restrictive alternative is still useful in these instances, as respecting autonomy is particularly important if we are already justifying forms of rights infringement (such as restricting freedom of movement) on account of mental illness. In light of this principle, it is important that any intervention that is deployed following the identification of risk by a medical or social suicide prediction tool is proportionate to the degree of risk detected by the algorithm.

Determining suicide risk accurately is less meaningful if at-risk individuals are not provided with access to high-quality treatments that have proven efficacy in preventing suicide (it is also important that the individual is willing and/or able to access the intervention).[8] Many of the technology companies, including Google and Facebook, recommend crisis hotlines to those identified as at-risk, despite mixed evidence that such hotlines save lives.[13–15] Governments and medical leaders should work to connect organisations engaged in social suicide prediction (such as Facebook, Google, etc) with suicide treatment experts to ensure that appropriate and effective interventions are deployed. Reasonable questions remain as to what should be done if interventions are not effective. Of course, we will only know if interventions are not effective if technology companies are transparent in providing their results for independent scrutiny. This situation should be carefully monitored by medical experts and researchers; if interventions are ineffective, new interventions could be applied, as long as there is no evidence that they cause harm.

### Transparency and explainability

Most standards for governance of AI and data recognise the importance of transparency to ensure public trust.[16 17] Any future deployment of medical suicide prediction tools in the community should be announced after clear public consultation; benefits and risks of the technology should be explained, as well as how it is expected to affect the community. This could be done by government, doctors, technology companies or even a national independent review committee formed specifically to oversee suicide prediction tools. Particular respect should be paid to those who may be disproportionally affected by this technology—for example, vulnerable groups which experience higher rates of suicide when compared with the general population, such as those living with mental illness, LGBTIQ people and those living in rural/remote areas.

Improving the transparency of social suicide prediction tools, such as Facebook's tools, should be a matter of pressing urgency for regulators. Doing so will help to ensure public trust and galvanise support for broader development of these potentially valuable tools. In addition to the ethical reasons for sharing methodologies and results, the accuracy of predictive algorithms for detecting suicide risk would also be improved by enabling cooperation between private companies and the academic community.[18] Technology companies might argue that the AI underpinning such efforts represents proprietary knowledge. However, given these tools deal with sensitive health information, the public interest in transparency trumps commercial considerations. This is the case because it would be ethically impermissible for Facebook, or any other firm, to use suicide risk information for commercial aims in the first place. As such, there should be no commercial value to suicide risk information, or the algorithms that inform users risk. Even if companies argued that the algorithm had commercial value for other purposes, the duty of easy rescue suggests that if one can save a person's life at a reasonably low cost, then it is ethically required to do so. Given Facebook's highly profitable business model—which posted an operating profit exceeding US$18 billion in 2019[19]—it seems reasonable to suggest that the company could absorb any small impact on profitability in order to save lives.

## Privacy and security

Following media coverage of initiatives such as My Health Record in Australia and Deep Mind's data sharing with the UK's National Health Service, there have been numerous concerns raised by the public around privacy and how their medical data are used. These concerns may make citizens reluctant to share their data with governments, researchers and health providers. In order to construct ethical AI and data systems, privacy should be upheld and protected wherever possible. Data governance standards will also play an important role so that access of health data reflects ethical and legal standards.

The following arguments are relevant to both medical and social prediction tools; however, as we have outlined, medical suicide prediction tools are generally less problematic than social suicide prediction tools, as they would likely be used in medical settings where subjects have provided consent.

With regards to social suicide prediction by private firms, there is an important and obvious tension between identifying people in imminent danger from suicide and deidentifying individual data. Some may argue that social media companies should not be required to conform with medical research standards, as they are not conducting 'research'. Indeed, social suicide prediction tools developed in the private sector are currently self-regulated, and do not conform to medical research standards. Facebook, for example, has an internal ethics committee to review challenging projects; however, it is unclear what criteria are used to determine when the committee is consulted, and what threshold is required for projects to 'pass' committee review.[20] Acknowledging that the approach of Facebook and other companies to suicide prediction is innovative also acknowledges that it likely falls under what a reasonable person would consider research.[18] As such, these tools should be governed by the same legal and ethical standards as other medical research.

The level of identifiability of social suicide prediction systems is contingent on how we balance a right to individual privacy with the responsibility to protect vulnerable individuals at risk of suicide. Privacy is clearly an important goal in any liberal democracy, and one which should be protected; however, suicidality represents an example where limited infringements on privacy may be potentially justified.

Four reasons are detailed below.

1. Reduced agency: certain people at risk of suicide may have reduced agency on account of severe mental illness and/or being under 18 years of age—in such instances, we may be justified in protecting these people to ensure the preservation of their future agency/autonomy.
2. Addressing inequities: social suicide prediction tools offer a unique opportunity to identify at-risk and vulnerable people (such as adolescents, indigenous people, LGBTIQ people and those living in rural/remote areas) who may otherwise not engage with the health system, due to limited access to services, stigma or cultural differences. As a result, these tools could facilitate provision of more effective and targeted mental health interventions, which could, in turn, reduce the outsized impact of suicide and health inequities.
3. Citizens might be open to these tools: given the significant human, economic and social cost of suicide, there is a similar public interest in reducing its impact. Citizens may be open to small infringements on their privacy through the monitoring of social media data in order to save lives. This deserves further research and consideration.
4. The tools are already here: as described by Coppersmith *et al*, the 'cat is already out of the bag,' so to speak.[6] That is, the widespread use of data coupled with advances in AI and analytical processes means that these technologies are going to be developed, whether ethicists object or not. As we have seen, many companies already have working tools—although we cannot be sure of their effectiveness. As such, there is a strong case for government, academia, clinicians and mental health advocates to work together to ensure that these technologies benefit those individuals who need it most, while proactively identifying and mitigating risks in a thoughtful and transparent fashion.

In response, critics could present two specific objections that mean suicide prediction tools are not morally permissible:

1. Disclosure of sensitive data: a potentially severe and harmful risk would be the release of sensitive suicide risk data, by mistake or through malicious actors (eg, hackers), leading to a significant privacy infringement for those identified as at risk of suicide. Yet, this harm, while serious, is unlikely, and there are steps that could be taken to minimise the risk of disclosure. An identifiable system can still protect individuals from external identification by guaranteeing strong technical protections, a rigorous governance framework, and training/oversight for staff with access to data. From a technical standpoint, control of access to data could be mediated by a computer in a distributed information environment, using block chain technology such as a personal health train and/or a data lake.[21 22] Such efforts would decrease the likelihood of unintentional release of personal information. National standards could be of assistance for developers and providers of social suicide prediction tools: Australia's National Human Medical Research Council's (NHMRC) 'National Statement on Conduct in Human Research', for example, describes a number of measures that can be taken to reduce risks in marginalised groups, such as separating the analysis of data from those who hold identifiers.[23]
2. Unjustified infringements on privacy through medicalisation: as discussed, it is possible that a small number of people may

be incorrectly identified as at risk of suicide by a prediction algorithm. This may lead to unjustified rights infringements, ranging from uncomfortable questioning by doctors to forced hospitalisation for high-risk patients. Minimising these invasions of liberty and autonomy should be a moral imperative for any suicide prediction tool. This risk could be minimised by improving the accuracy of tools to minimise false positives. Interventions should also be proportionate to the degree of risk and impose the smallest rights infringement possible. Soft touch interventions, such as providing information and optional counselling of users, should come first. It should be noted again, however, that these considerations are not specific to AI; they would be the same if we only relied on human therapists.

If identification of individuals is deemed unacceptable, AI could still be used to inform suicide prevention policy at a population level—thus not compromising individual privacy. Using AI to inform suicide prediction/prevention though deidentified data represents 'low hanging fruit' for policy makers; that is, some of the benefits of AI could be realised without negatively impacting community trust and avoiding potentially difficult ethical and political battles over consent and privacy.

## Consent

The development of any medical suicide prediction tool will have to navigate questions of consent; that is, should consent be required for an algorithm to scan patient records and determine a sensitive piece of information (an individual's suicide risk)?

There is a body of ethical literature that outlines a moral justification for infringing on the rights of an individual who may pose a risk of harm to themselves or others.[24] Existing legislation in many jurisdictions allows psychiatrists to detain and treat mentally ill people without their consent, for example, if it is judged to be in their best interests.[25 26] While there is not direct moral equivalence between such action and the monitoring of patient records to determine suicide risk, it seems at least *morally plausible* that researchers and health providers could be justified in using patient data from EMRs to create risk assessment tools that improve doctor capabilities in identifying suicide risk. Clearly this process would require rigorous ethical and legal oversight.

The question of consent in social suicide prediction presents a more difficult ethical problem. In addition to the concerns outlined above, another pertinent question is whether these systems should be consent-in, or opt-out? Both options have shortcomings. A consent-in system will not be able to reach all of those deemed at risk of suicide, while an opt-out system would raise concerns around privacy and consent. Given that a primary benefit of social suicide prediction is to identify people at risk who may not engage with traditional health services, from an effectiveness standpoint, it makes sense to cast a 'wide net'—that is, for a system to be opt-out. One option is for researchers to follow the lead of Public Health Canada by developing a tool that identifies suicide trends at a population level in order to avoid the consent-in/opt-out controversy that plagued My Health Record in Australia.[27] It is worth noting that Facebook and Google do not even allow users to opt-out of their own suicide prediction tools.

Governance arrangements for medical and social suicide prediction tools will require extensive public consultation. A review of 15 studies examined public support for using social media for research purposes, finding strong support for research of social media that advances the 'public good'.[28] While similar consultation would need to take place in local contexts to determine public support for suicide prediction tools, these results suggest that there may be public support for medical suicide prediction tools—including those combined with other linked data sources, such as social media data—when used to reduce the risk of suicide, particularly in vulnerable populations.

What reasons might exist for greater public support for the use of AI/data to reduce the impact of suicide, when compared with other uses of technology in healthcare? Suicide might be the subject of particular community concern when compared with other medical issues, and there could be decreased privacy concerns for those deemed to be at risk—for example, adolescents suffering from mental illness. In other words, an issue like adolescent suicide might unite the community in such a way that there is public approval for more intrusive, opt-out social and medical prediction models that do not require consent. This idea warrants further consideration by legal experts and should be the subject of consultation with the public.

## Independent ethics oversight

Research into medical suicide prediction is governed by general principles of medical ethics—autonomy, beneficence and non-maleficence (do no harm).[24] Research conducted on medical suicide prediction tools has proceeded in line with standard national ethics guidelines in those countries, such as the NHMRC National Statement in Australia.[23] A UK-based study by DelPozo-Banos *et al* received ethics approval from the Information Governance Review Panel, an independent body consisting of government, regulatory and professional agencies that provide approval for the use of the Secure Anonymised Information Linkage (SAIL) integrated dataset, an advanced data platform specifically designed for research.[29–31] It is imperative that any future research on medical suicide prediction adheres to ethics guidelines and national laws. A national, independent research body could be established to provide oversight of suicide prediction tools, capitalising on advanced security tools (such as blockchain) to help minimise the risk of privacy breaches.[32]

Social suicide prediction tools developed in the private sector are currently self-regulated, and do not conform to medical research standards. Facebook, for example, has an internal ethics committee to review challenging projects; however, it is unclear what criteria are used to determine when the committee is consulted, and what threshold is required for projects to 'pass' committee review.[20] As mentioned, acknowledging that the approach of Facebook and other companies to suicide prediction is innovative also acknowledges that it likely falls under what a reasonable person would consider research.[18] Thus, these tools should be governed by the same legal and ethical standards as other medical research.

## Accountability

No medical or social suicide prediction tool will ever achieve perfect accuracy. What happens when things go wrong? In other words, when an algorithm fails to identify a patient at risk of suicide who subsequently dies, who should be held accountable? In addition, how should we compensate those who suffer rights infringement on behalf of false positives? The legal community is wrestling with similar questions in other areas where AI may play a role, such as self-driving cars. In the short term, developers of AI suicide prediction tools should be identifiable to regulators and the public, regardless of how legal experts answer the question of accountability. This will guarantee a level of responsibility over outcomes, and transparency will help to galvanise public support. Regular human oversight and review of social suicide prevention tools will be particularly important. Response protocols will be required for handling high-risk cases that are flagged by AI, as well as what should be done when AI risk assessments differ from clinical opinion.[9] These questions will require multidisciplinary input to address medical, technical, legal and ethical perspectives.

### Respect for human rights and equity

AI and data can enable and/or detract from human rights. Such is the case with medical and social suicide prediction tools. As discussed, there is a trade-off between individual privacy and ensuring vulnerable people are protected from killing themselves. We have discussed a number of ways that privacy can be protected through rigorous technical safeguards and governance structures. Furthermore, some interference with human rights (in this instance, privacy) may be justified if there is substantial benefit and the infringement is reasonable and proportionate.[11] With the right regulation and research, AI tools have the potential to save hundreds or even thousands of lives per year, reducing the health, economic and social impact of suicide in the community. Given that the burden of suicide disproportionally falls within disadvantaged groups, including rural/regional, low-income and LGBTIQ communities', targeting mental health interventions in these high-risk groups will help to enable autonomy and equity.

Data produced by medical and social suicide prediction tools should not be used for purposes that do not improve the health of the community. For example, the transfer of health data to third parties (such as insurers and advertisers) to inform commercial efforts is unjust, as it places an inequitable burden on those who are deemed to be at risk of suicide when compared with those who are not.

Bias along gender, racial or cultural lines is another concern to human rights. There have been a number of well publicised cases of bias in machine-learning algorithms.[33] Bias is unlikely to be eliminated entirely. As described by AI expert Professor Toby Walsh, the word machine learning is actually defined as 'inductive bias'.[34] Nonetheless, researchers and technology companies must take reasonable measures to eliminate bias where possible.

### Contestability

The ethical principle of contestability can also enhance public trust in dealing with AI systems. Contestability means providing a practical mechanism by which individuals or groups from the community can challenge the use of an AI system on an ongoing basis when it significantly impacts that individual or group. In this context, an independent organisation could be formed to provide a mechanism for ongoing public input and mediation about the role of medical and social suicide prediction tools in society. Such an organisation should be multidisciplinary, with medical, policy, and technical members, as well as representatives from patient groups.

### Slippery slope to minority report

Some will object that these tools are the thin end of the wedge in behavioural prediction. Similar technologies might be used to prevent the killing of others, for example. This might lead to widespread rights infringements. However, surely there is an even stronger justification to use such predictive tools to prevent harm to others. With the right ethical and legal oversight, these tools are unlikely to lead to a dystopian future such as that outlined by George Orwell in 1984. Nonetheless, as predictive technologies become more advanced, these questions will require careful consideration by policymakers, doctors, technologists, ethicists, and members of the public.

## CONCLUDING THOUGHTS AND IMPLICATIONS FOR MEDICAL LEADERS

As technological advances open up opportunities to predict suicide risks and potentially prevent harm through the use of AI, there is a need to ensure that there is ethical oversight on the use of these emerging algorithms to ensure that they are not misused.

Doctors in leadership roles will be pivotal to this process. Medical leaders must ensure that AI systems, which may have a clinical impact on patients, are introduced in a safe way. A robust governance process, around how new technologies and clinical practices are introduced in hospitals and mental health services, need to be established—this process should closely examine the effectiveness, cost and appropriateness of these tools, while paying respect to ethical principles outlined in this paper. Similarly, if there is a lack of evidence for a new technology, then research ethics approval may need to be considered.

Before using suicide prediction tools, appropriate training and education programmes should be put in place. This could involve credentialing of clinicians to ensure that only those trained in the use of AI technologies and those who know how to interpret the data are allowed to use such systems. Training will help to improve staff engagement on AI that identifies and treats those at risk of mental illness, as well as the use of AI in other areas of clinical practice.

Medical leaders who understand the ethical risks associated with AI will be better prepared to work with technical and ethical experts to navigate the risks and ethical challenges involved. The implementation of medical suicide prediction tools may be relatively unproblematic, as long as they produce clear public benefit, do not cause harm, patients consent to their use, and rigorous technical protections are applied. Social suicide prediction tools may also be acceptable in a medical setting if they are implemented in a fashion that clearly explains the use of private social media data to patients so as to facilitate informed consent.

On the other hand, the use of social suicide prediction tools by private companies, such as Facebook, is more problematic, particularly where these companies are not transparent about their results. Medical leaders can play an important role in advocating to policymakers for the rights of those with mental illness, with the aim of bringing about improvements in the way private firms use these tools so that they pay respect to ethical principles.

Every medical leader will have to decide how new technologies can be utilised to improve patient outcomes. Determining whether AI is appropriate for use in an institution will depend on the resources and requirements of different clinical settings. Front and centre of every medical leader's considerations should be to advocate for the interests of patients and their families by ensuring that AI tools are used in a way that is consistent with medical ethics.

**Twitter** Daniel D'Hotman @danieldhotman and Erwin Loh @erwinloh

**Contributors** DDH planned and wrote the original draft of this paper. EL and JS provided feedback on this draft and contributed to the final version of the paper.

**Funding** DD's DPhil is funded on a Rhodes Scholarship. This work was supported by the Wellcome Trust [WT203132], [WT104848].

**Competing interests** None declared.

**Patient consent for publication** Not required.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**ORCID iD**

Erwin Loh http://orcid.org/0000-0001-7157-0826

## REFERENCES

1 World Health Organization. Suicide, 2019. Available: https://www.who.int/news-room/fact-sheets/detail/suicide [Accessed 9 Jul 2020].

2 Ropper AH, Seena F, Runeson B. D. *N Engl J Med* 2020;382:266–74.

3 Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? *Am J Med* 2018;131:129–33.

4 Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci* 2017;5:457–69.

5 Loh E. Medicine and the rise of the robots: a qualitative review of recent advances of artificial intelligence in health. *BMJ Leader* 2018;2:59–63.

6 Coppersmith G, Leary R, Crutchley P, *et al*. Natural language processing of social media as screening for suicide risk. *Biomed Inform Insights* 2018;10:117822261879286.

7 Nock MK, Borges G, Bromet EJ, *et al*. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *Br J Psychiatry* 2008;192:98–105.

8 Kann L, Kinchen S, Shanklin SL, *et al*. Youth risk behavior surveillance--United States, 2013. *MMWR Suppl* 2014;63:1–168.

9 Mason M. Artificial intelligence based suicide prediction, 2019. Available: https://ssrn.com/abstract=3324874

10 Sheehan L, Dubke R, Corrigan PW. The specificity of public stigma: a comparison of suicide and depression-related stigma. *Psychiatry Res* 2017;256:40–5.

11 Childress JF, Faden RR, Gaare RD, *et al*. Public health ethics: mapping the terrain. *J Law Med Ethics* 2002;30:170–8.

12 Atkinson JM, Garner HC. Least restrictive alternative – advance statements and the new mental health legislation. *Psychiatr Bull* 2002;26:246–7.

13 Hunt T, Wilson CJ, Woodward A, *et al*. Intervention among suicidal men: future directions for telephone crisis support research. *Front Public Health* 2018;6:1.

14 Beautrais A, Fergusson D, Coggan C, *et al*. Effective strategies for suicide prevention in New Zealand: a review of the evidence. *N Z Med J* 2007;120:U2459.

15 Scott A, Guo B. For which strategies of suicide prevention is there evidence of effectiveness? 2012. Available: https://www.euro.who.int/__data/assets/pdf_file/0010/74692/E83583.pdf[Accessed 9 Jul 2020].

16 GDPR. General data protection regulation (GDPR) compliance guidelines, 2020. Available: https://gdpr.eu/ [Accessed 9 Jul 2020].

17 Department for Digital, Culture, Media and Sport. Data ethics framework, 2018. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/737137/Data_Ethics_Framework.pdf [Accessed 10 Jul 2020].

18 Barnett I, Torous J, Ethics TJ. Ethics, transparency, and public health at the intersection of innovation and Facebook's suicide prevention efforts. *Ann Intern Med* 2019;170:565.

19 Facebook. United States: Facebook, Investor relations, 2020. Available: https://investor.fb.com/investor-news/press-release-details/2020/Facebook-Reports-Fourth-Quarter-and-Full-Year-2019-Results/default.aspx [Accessed 10 Jul 2020].

20 Gomes de Andrade NN, Pawson D, Muriello D, *et al*. Ethics and artificial intelligence: suicide prevention on Facebook. *Philos Technol* 2018;31:669–84.

21 Soest V, Sun C, Ole M, *et al*. *Using the personal health train for automated and privacy-preserving analytics on vertically partitioned data*. Europe: MIE, 2018.

22 Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. *Health Aff* 2014;33:1115–22.

23 National Health and Medical Research Council. *National statement on ethical conduct in human research 2007*. Canberra, Australia, 2018.

24 Beauchamp T, Childress J. *Principles of biomedical ethics*. New York: Oxford University Press, 2001.

25 Freeman M, Pathare S. *WHO resource book on mental health, human rights and legislation*. Geneva: World Health Organization, 2005.

26 RANZCP. Mental health legislation Australia and New Zealand. Available: https://www.ranzcp.org/practice-education/guidelines-and-resources-for-practice/mental-health-legislation-australia-and-new-zealan [Accessed 16 Jul 2020].

27 Vogel L. Ai opens new frontier for suicide prevention. *CMAJ* 2018;190:E119.

28 Golder S, Ahmed S, Norman G, *et al*. Attitudes toward the ethics of research using social media: a systematic review. *J Med Internet Res* 2017;19:e195.

29 Ford DV, Jones KH, Verplancke J-P, *et al*. The SAIL databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009;9:157.

30 Lyons RA, Jones KH, John G, *et al*. The Sail databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 2009;9:3.

31 DelPozo-Banos M, John A, Petkov N, *et al*. Using neural networks with routine health records to identify suicide risk: feasibility study. *JMIR Ment Health* 2018;5:e10144.

32 Porsdam Mann S, Savulescu J, Sahakian BJ. Facilitating the ethical use of health data for the benefit of society: electronic health records, consent and the duty of easy rescue. *Philos Trans A Math Phys Eng Sci* 2016;374:20160130.

33 Hao K. This is how al bias really happens—and why It's so hard to fix, 2019. Available: <https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/> [Accessed 13 July 2020].

34 D'Hotman D. *Telephone interview with Professor Toby Walsh*. Canberra, Australia, 2019.