The ethics of probabilistic suicide risk prediction models: a response to Cockburn and Large

As the data scientist and ethicist Fabian Beigang (2022) notes, algorithmic decision-making involves two ethically distinct stages: a prediction, and a decision taken (a least partly) based on that prediction. Predictions aim at accuracy, while decisions concern the allocation of scarce resources according to appropriate distributive principles. A recent feature article by Cockburn and Large (2025) argues that probabilistic suicide risk prediction models—such as OxSATS (Fazel et al. 2023)—should not be used in clinical care. However, their critiques conflate prediction with decision-making and are better directed at problems of resource scarcity and unreflective allocation practices than at the predictive function of OxSATS itself.

i. Conflating predictions and decisions

Cockburn and Large use vignettes to illustrate how OxSATS might affect clinical care. One vignette concerns the prioritisation of older male patients over younger female patients, based on the higher correlation between suicide risk, age, and male gender. ‘’If clinicians and services wanted to introduce such tools’, they argue, ‘their introduction of such biases would necessitate candid, thoughtful discussions among professionals, patients and healthcare funders.’

But if suicide risk genuinely correlates with age and gender, incorporating these variables into a predictive model does not itself introduce bias. To suggest otherwise conflates prediction, which is evaluated in terms of accuracy, with allocation, which is evaluated in terms of distributive justice. The relevant ethical question is whether it is permissible to prioritise some individual patients over others based on factors such as group membership, even when their predicted risk is lower.

There may be principled reasons to do so. A relational egalitarian (Anderson 1999) might argue that historical and ongoing failures to adequately address women’s mental health justify additional prioritisation as a corrective measure to inequalities that reflect hierarchies of moral value. A utilitarian (Eggleston 2012) might contend that preventing suicide among younger patients produces greater overall benefit by preserving more future life-years. These arguments are contestable, but they concern allocation decisions rather than the legitimacy of risk prediction. Importantly, such distributive questions would arise regardless of whether OxSATS is used.

Indeed, OxSATS may increase transparency by making explicit the role of age and gender in suicide risk prediction, in contrast to unaided clinical judgement where such factors may influence decisions implicitly and without scrutiny. One advantage of algorithmic tools is that they can expose morally salient considerations that would otherwise remain

opaque within individual clinicians' reasoning, thereby facilitating ethical reflection on allocative decision-making.

ii. Resource scarcity as the underlying problem

Another vignette describes a young patient with a low suicide risk score who may be deprioritised for mental health referral despite significant distress. Cockburn and Large suggest this outcome would constitute a distributive injustice attributable to the use of OxSATS. However, under conditions of severe resource scarcity, the fact that a patient who could benefit from care does not receive it is not, by itself, evidence of injustice. From a distributive justice perspective, injustice arises when someone with a stronger claim to a scarce resource is denied it in favour of someone with a weaker claim. This situation is analogous to transplant ethics: although all patients on a waiting list would benefit from a transplant, scarcity necessitates prioritisation. Ethical evaluation therefore concerns the criteria used to allocate resources, not the mere existence of unmet need. Vignette-based attention to disadvantaged individuals is insufficient to demonstrate that OxSATS would produce unjust outcomes.

Cockburn and Large may nevertheless be correct that the broader context of overstretched mental health services constitutes a systemic injustice, insofar as insufficient societal resources are directed towards mental health care. In such non-ideal conditions, allocation decisions must still be made, but they are constrained by unjust background circumstances. OxSATS does not create this scarcity. The same allocation challenges would arise whether decisions were made by algorithms, clinicians, or a combination of both.

iii. The risk of unreflective allocation

Cockburn and Large further criticise the potential for predictive tools to be treated as determinative of allocation decisions. One vignette concerns a patient with emergent schizophrenia who might be denied appropriate treatment if a low OxSATS score were used as a screening threshold for further assessment. However, this would constitute an inappropriate use of the tool. A judgement that a patient is at low risk of suicide does not entail that they do not require mental health support, particularly given the complex and contested relationship between mental illness and suicide risk (cite?).

The possibility of misuse does not itself justify abandoning a predictive tool. Rather, it highlights the importance of clearly specifying appropriate uses and limitations. Cockburn and Large offer inconsistent accounts of OxSATS's intended role. At one point they ask readers to ‘‘keep an open mind that good clinicians can and will outperform these tools', suggesting an adversarial relationship between clinical and algorithmic judgement, whereas the OxSATS is intended as a consultative tool for clinicians, exactly because, as they put it 'the role of the clinician is much broader than risk prediction'. Of OxSATS, they note that ‘‘It is designed to be used alongside clinicians as a clinical adjunct, not as a replacement. However,

the authors suggest that it could be used to allocate treatment to those most at risk.’ But using such a tool to inform allocation decisions is compatible with clinicians retaining ultimate responsibility, and there is evidence from other areas of healthcare that predictive tools can augment clinical judgement when properly integrated (Thurtle et al. 2019; Wijnberge et al. 2020) .

Cockburn and Large are clearly concerned that algorithmic outputs will be routinely deferred to, leading to automation bias, clinician deskilling, or complacency. These risks are genuine and warrant careful examination of how OxSATS should be embedded in clinical workflows so that its predictions can be used productively in resource allocation decision-making to improve patient outcomes. However, assessing these risks requires empirical investigation of actual practice rather than presuming misuse in advance. The possibility that a tool might be used poorly is not sufficient reason to conclude that it should not be used at all.

Bibliography

Anderson, Elizabeth S. 1999. ‘What Is the Point of Equality?’ *Ethics* 109 (2): 287–337.

Beigang, Fabian. 2022. ‘On the Advantages of Distinguishing Between Predictive and Allocative Fairness in Algorithmic Decision-Making’. *Minds and Machines* 32 (4): 655–82. https://doi.org/10.1007/s11023-022-09615-9.

Cockburn, Alastair, and Matthew Large. 2025. ‘A Clinician’s Guide to Probabilistic Suicide Risk Prediction Tools: Cautions and Pitfalls’. *The British Journal of Psychiatry*, November 5, 1–5. https://doi.org/10.1192/bjp.2025.10458.

Eggleston, B. 2012. ‘Utilitarianism’. In *Encyclopedia of Applied Ethics*. Elsevier. https://doi.org/10.1016/B978-0-12-373932-2.00220-9.

Fazel, Seena, Maria D. L. A. Vazquez-Montes, Yasmina Molero, et al. 2023. ‘Risk of Death by Suicide Following Self-Harm Presentations to Healthcare: Development and Validation of a Multivariable Clinical Prediction Rule (OxSATS)’. *BMJ Mental Health* 26 (1). https://doi.org/10.1136/bmjment-2023-300673.

Thurtle, David R., Valerie Jenkins, Paul D. Pharoah, and Vincent J. Gnanapragasam. 2019. ‘Understanding of Prognosis in Non-Metastatic Prostate Cancer: A Randomised Comparative Study of Clinician Estimates Measured against the PREDICT Prostate Prognostic Model’. *British Journal of Cancer* 121 (8): 715–18. https://doi.org/10.1038/s41416-019-0569-4.

Wijnberge, Marije, Bart F. Geerts, Liselotte Hol, et al. 2020. ‘Effect of a Machine Learning–Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial’. *JAMA* 323 (11): 1052–60. https://doi.org/10.1001/jama.2020.0592.