



Bioethics

ISSN: 2989-5448 (Online)

Publisher: Tomorrow's Research Publications (TRP)

An International Peer-Reviewed Open Access Journal

Policy Analysis

Ethical Dimensions of Predictive Algorithms in Psychiatric Risk Assessment

Dimitris Eberhardt^{1*}, Fredrik Toth² and Etienne Berg³¹Department of Philosophy, Baltic Institute of Medical Humanities, Sweden²Department of Bioethics, Baltic Institute of Medical Humanities, Sweden³Center for Medical Ethics and Law, Shannon University Centre for Bioethics, Ireland

ABSTRACT

Predictive algorithms trained on health records and passively collected behavioral data are increasingly used to estimate a patient's risk of suicide, violence, or psychiatric crisis, often outperforming unstructured clinical judgment on retrospective measures.

This policy analysis reviews four recurring concerns in the recent literature: the gap between validation-set accuracy and real clinical utility, demographic bias inherited from historically skewed psychiatric data, the ambiguous privacy status of digital phenotyping, and the use of risk scores in involuntary treatment decisions.

Existing mental health legislation and consent practices were largely written without algorithmic inputs in mind, leaving clinicians and tribunals with little guidance on how much weight a machine-generated score should carry.

We argue that predictive accuracy is necessary but not sufficient, and that disclosure, fairness auditing, and a genuine mechanism for contesting a score need to be built into deployment from the outset.

We close with recommendations for narrowing the gap between what these tools can do and what responsible clinical and legal use of them requires.

Keywords: predictive analytics; psychiatric risk assessment; algorithmic fairness; digital phenotyping; involuntary treatment

Citation: Eberhardt D, Toth F and Berg E (2024) Ethical Dimensions of Predictive Algorithms in Psychiatric Risk Assessment. *Bioethics*. 1(1):01-05.

Corresponding Author: Dimitris Eberhardt

E-mail Address: d.eberhardt@bimh.se

Article History: Received 25 November 2023 | Accepted 03 March 2024 | Published 10 June 2024

Copyright: © 2024 Dimitris Eberhardt, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (<https://creativecommons.org/licenses/by/4.0/>).

INTRODUCTION

Psychiatry has wanted a reliable way to see risk coming for as long as it has existed as a discipline.

Clinical judgment about who is likely to attempt suicide, become violent, or relapse into crisis has always been imperfect, and for decades the honest response was to acknowledge that limitation and act

cautiously around it. Machine learning models trained on electronic health records, insurance claims, and increasingly on smartphone sensor streams now promise something different: a quantified, individualized probability of a bad outcome, delivered before it happens. Systematic reviews of these models report genuinely improved discrimination over unstructured clinical judgment on retrospective data [1], and models built directly from routine electronic health record data have shown they can flag elevated crisis risk days before an event that clinicians did not otherwise anticipate [2]. That capability is real, but its arrival has outpaced the ethical and regulatory scaffolding that would make its clinical use defensible.

The concern animating this review is not that predictive psychiatry is inherently illegitimate, but that the specific properties of psychiatric risk prediction, contested criterion validity, a checkered history of coercion in the specialty, data sources that were never collected with prediction in mind, and outcomes (suicide, violence, involuntary commitment) that carry unusually high stakes for the people being scored, make it a harder case for algorithmic deployment than most other areas of clinical prediction [3]. A recent scoping review of the ethical and social literature on psychiatric risk prediction organized the concerns into recurring clusters: consent and disclosure, fairness and bias, clinical utility versus stigmatization, and the difficulty of meaningfully contesting an algorithmic score once it has entered a chart [4]. This review follows a similar structure, examining what the recent literature says about validity and consent, fairness, privacy in the specific context of passively collected behavioral data, and the use of algorithmic risk scores in decisions that curtail liberty.

VALIDITY, CONSENT, AND THE LIMITS OF WHAT THESE MODELS ACTUALLY PREDICT

A recurring finding across the literature is that predictive accuracy at the population level does not translate cleanly into a clinically actionable warning for a given patient. Systematic reviews of machine learning approaches to suicide prediction in psychiatric populations report models with impressive area-under-the-curve statistics in validation samples, but far weaker positive predictive value at the low base rates that characterize real clinical populations, meaning that most patients flagged as high-risk will not go on to attempt suicide, while some patients who do were never flagged [5]. Earlier reviews of the field's methodological maturity made a similar point more bluntly: much of the published work in machine learning for suicide risk prediction has been evaluated on data and in settings that do not resemble how the tools would actually be deployed, which leaves a gap between publishable performance and defensible clinical use [6].

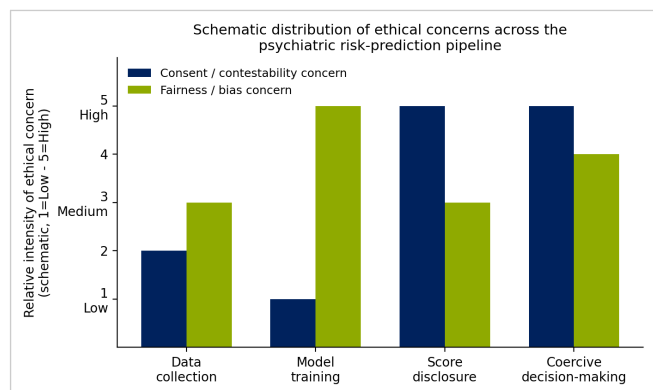
That gap has direct consequences for consent.

Patients are rarely told that a risk score is being generated from their record, let alone what data feed it or how it will be acted upon, and a widely cited stakeholder-informed framework for suicide risk models built from electronic health records concluded that transparency about the existence and basis of a score is a precondition for ethical deployment, not an optional add-on to be handled later [7]. Ethical analyses aimed specifically at clinicians treating children and adolescents have raised a related worry: a young patient's future is unusually malleable, and an algorithmically generated risk label attached early in a chart risks becoming a self-fulfilling institutional fact, shaping how subsequent clinicians and services treat that patient regardless of how the underlying risk evolves [8]. Perspective pieces on personalized predictive psychiatry converge on a similar caution: the promise of individualized risk stratification is real, but only when it is paired with a like-for-like acknowledgment of uncertainty communicated to the patient rather than delivered as a settled fact [3]. Table 1 lays out where in the prediction-to-action pipeline these ethical concerns concentrate, from data collection through to the clinical decision the score is meant to inform, and Figure 1 sketches, schematically, how the intensity of each concern tends to shift across that same pipeline.

Table 1. Where ethical concerns concentrate along the psychiatric risk-prediction pipeline, from data collection through to the clinical decision the score is meant to inform.

Pipeline stage	Primary ethical concern	What the literature reports	Key sources
Data collection	Consent for passive/behavioral data	Patients often unaware continuous sensor or record data is being repurposed for risk scoring	npjdm2018 phenotyping; jmir2022 language
Model training	Fairness and representativeness	Historical bias in psychiatric records and narrow validation samples propagate into model output	npjmhr2026 aggression; molpsych2022clinical
Score disclosure	Transparency and contestability	Scores rarely disclosed to patients in an actionable form; basis often not narratable	archsui2022stakeholder; lssp2026child
Coercive decision-making	Legitimacy of algorithmic input to liberty-restricting decisions	Legislation largely silent on algorithmic evidence; override mechanisms weak or informal	diagnostics2022algorithmic; plos2024judgment

Figure 1. Schematic comparison of consent/contestability concern and fairness/bias concern across four stages of the psychiatric risk-prediction pipeline. Illustrative only, not derived from a single empirical study or quantitative meta-analysis.



FAIRNESS, BIAS, AND THE DATA THESE MODELS LEARN FROM

Predictive models inherit the biases embedded in the records they are trained on, and psychiatric records carry a particularly fraught history on this front, including well-documented disparities in diagnosis, involuntary commitment, and use of restraint across race and socioeconomic status. A first-of-its-kind fairness analysis of models predicting inpatient aggression found that models calibrated on aggregate data performed unevenly across demographic subgroups in ways that tracked existing disparities in how aggression is documented and acted upon in acute psychiatric wards, rather than reflecting any underlying difference in true risk [9]. A parallel analysis of models predicting depression risk from smartphone-sensed behavioral data reported comparable subgroup disparities, and argued that behavioral-sensor data, often treated as more "objective" than clinician-recorded diagnoses precisely because it is passively collected, is not thereby exempt from encoding the same demographic and socioeconomic patterning found in conventional records [10].

Work specifically on psychosis-risk prediction has pushed this further by asking whether "fairness," as usually operationalized in the machine learning literature (equalized error rates across groups), is even the right target for a domain where the clinical categories themselves, such as "clinical high risk" for psychosis, are contested and unevenly applied across populations to begin with; the analysis found that several standard fairness metrics gave contradictory verdicts on the same models depending on which was chosen, undermining the assumption that fairness is a single well-defined property a model either has or lacks [11]. Reporting on this broader pattern, one recent research summary described what several fairness audits are now converging on: risk-prediction tools trained on historically biased psychiatric data can reproduce and even sharpen that bias at scale, doing so under a veneer of algorithmic objectivity that makes

the resulting disparity harder for clinicians to notice and challenge than an individual clinician's own biased judgment would be [12]. A broader review of predictive analytics literature in mental illness reached a related methodological conclusion, that most published models are validated on demographically narrow samples relative to the populations where they are eventually deployed, which limits how much a strong validation-set performance figure can be trusted to hold up elsewhere [1].

DIGITAL PHENOTYPING AND THE BOUNDARIES OF PSYCHIATRIC PRIVACY

A second major fault line concerns the shift from records generated within a clinical encounter to continuous behavioral data collected passively through smartphones and wearables, so-called digital phenotyping. An early and still widely cited ethical analysis of this shift argued that digital phenotyping data occupies an uncomfortable middle ground between clinical data, subject to familiar consent and confidentiality norms, and ordinary consumer data, subject to essentially none, and that regulators and institutional review boards were unprepared for data streams that blur the two [13]. That ambiguity has not been fully resolved; a recent review of ethical frameworks specific to mental healthcare found continued disagreement over whether passive behavioral monitoring requires the same standard of informed consent as an active clinical assessment, given that the monitored behavior (phone use patterns, movement, sleep, typing dynamics) is generated continuously and often without the patient consciously producing it as clinical information [14].

There is also a more subjective harm reported in the qualitative literature: patients monitored through digital phenotyping platforms have described a felt loss of interpretive authority over their own mental state, since a passive sensor stream can be used to override or contradict a patient's own first-person account of how they are doing. Work on this theme has framed the resulting dynamic as a form of epistemic injustice, in which the algorithmically-derived reading of a patient's data is treated by clinicians as more credible than the patient's own testimony, inverting a relationship that clinical psychiatry has spent decades trying to correct in the opposite direction [15]. Ethical guidance on language-analysis tools used for psychiatric diagnosis and prediction, drawing on speech and text samples that patients may not realize are being analyzed for diagnostic signal, has raised a structurally similar concern about consent for data uses that are not obvious from the point of collection [16].

ALGORITHMIC RISK SCORES IN COERCIVE DECISIONS

The sharpest ethical stakes arise when a risk score feeds directly into a decision that restricts a patient's liberty, most notably involuntary hospitalization. A

detailed ethical and legal analysis of algorithmic assessments used in decisions about voluntary, assisted, or involuntary psychiatric treatment concluded that existing mental health legislation in most jurisdictions was not written with algorithmic inputs in mind, leaving unclear how much weight a clinician or tribunal may legitimately give to a machine-generated score when the underlying reasoning is not fully inspectable [17]. This matters because the standard justification for compulsory treatment, that a clinician has exercised structured professional judgment and can explain the basis for that judgment to a tribunal or the patient, does not transfer cleanly to a statistical model whose internal weighting of hundreds of variables cannot be narrated in the way a clinician's reasoning can. A systematic review of how clinicians actually use structured professional judgment tools in secure admission decisions found that even well-validated structured instruments are frequently overridden or reinterpreted by clinical teams in practice, which raises the question of what a similar review would find for less transparent statistical models where override requires contesting a number rather than a documented rationale [18].

Table 1 summarizes how these concerns distribute unevenly across the stages of implementation, and Figure 1 presents the same comparison schematically, showing why interventions concentrated only at model-building (better training data, better calibration) leave the deployment-stage concerns, consent, contestability, and coercive use, largely untouched.

TOWARD ACCOUNTABLE IMPLEMENTATION

The literature reviewed here converges on a few recurring recommendations even where it disagrees on emphasis. First, predictive accuracy reported on held-out validation data should not be treated as a proxy for clinical utility until it has been tested prospectively in the setting where it will actually be used, at the base rates that setting actually has [5]. Second, fairness auditing needs to happen before deployment and periodically afterward, not as a one-time check, since demographic drift in the deployment population and changes in clinical documentation practice can both silently shift a model's error distribution across groups [9]. Third, patients should be told, in terms they can act on, that a risk score exists, roughly what feeds it, and what decision it is meant to inform, a standard that current practice in suicide risk prediction and digital phenotyping alike generally falls short of [7] [14]. Finally, where a score is used to support a coercive decision, some mechanism for the patient or their advocate to contest the score's basis, not merely its numerical output, is close to a precondition for treating the resulting decision as legitimate rather than merely efficient [17]. Commentary from bioethics centers examining AI-based suicide prediction specifically has argued that

interpretability requirements should scale with the severity of the downstream decision, meaning a screening tool used to prompt a conversation warrants a lower transparency bar than a score cited in a commitment hearing [19].

None of this amounts to an argument against predictive psychiatry as such. The reviewed evidence that these tools can catch signal clinicians miss is credible and, in aggregate, larger literature reviews of the broader clinical prediction field in psychiatry over the past two decades show real methodological progress alongside the persistent problems [1]. The argument is narrower: that the ethical infrastructure around consent, fairness auditing, and contestability has to be built into deployment from the outset, because retrofitting it after a tool is already shaping hospitalization decisions is far harder than building it in from the start [20].

CONCLUSION

Predictive algorithms in psychiatry sit at an unusually difficult intersection of genuine clinical promise and genuine ethical hazard, and the recent literature surveyed here suggests the two cannot be neatly separated by simply building better models. Improving discrimination and calibration addresses only the technical half of the problem; the harder half, disclosure to patients, fairness across groups whose psychiatric records already carry historical bias, the privacy status of passively collected behavioral data, and the legitimacy of using an uninspectable score to justify a coercive decision, requires governance choices that model developers alone cannot make. The recurring theme across validity, fairness, privacy, and coercion is the same: predictive accuracy is necessary but not sufficient, and the missing piece each time is a mechanism by which the person being scored can know the score exists, understand roughly what produced it, and contest it when it is wrong.

DECLARATIONS

Acknowledgements

The authors are grateful to colleagues at their respective institutions for discussions that helped sharpen the scope and framing of this review.

Funding

This review received no external funding.

Conflict of Interest

The authors declare no conflicts of interest relevant to this work.

Author Contributions

Dimitris Eberhardt conceived the review, coordinated the literature synthesis, and drafted the initial manuscript. Fredrik Toth performed the analysis of the fairness and digital phenotyping literature and critically revised the manuscript for intellectual content. Etienne

Berg contributed the sections on coercive decision-making and accountable implementation and approved the final version of the manuscript. All authors read and approved the final manuscript.

Ethics Statement

As a literature review synthesizing previously published, publicly available research and policy analysis, this article did not involve new data collection from human participants and did not require institutional ethics committee approval.

Data Availability

No new data were generated or analyzed for this review. All sources discussed are cited in the reference list and are publicly available through their respective publishers.

REFERENCES

1. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges (2022) *Molecular Psychiatry*.
2. Machine learning model to predict mental health crises from electronic health records (2022) *Nature Medicine*.
3. Towards personalised predictive psychiatry in clinical practice: an ethical perspective (2024) *The British Journal of Psychiatry*.
4. Ethical and social issues in prediction of risk of severe mental illness: a scoping review and thematic analysis (2025) *BMC Psychiatry*, PMID:40389909.
5. Machine learning and the prediction of suicide in psychiatric populations: a systematic review (2024) *Translational Psychiatry*.
6. Machine learning in suicide science: applications and ethics (2019) *Behavioral Sciences & the Law*, PMID:30609102.
7. A stakeholder-informed ethical framework to guide implementation of suicide risk prediction models derived from electronic health records (2022) *Archives of Suicide Research*, PMID:35446244.
8. Ethical considerations in the clinical application of prediction tools for severe mental disorders: perspectives from child and adolescent psychiatrists (2026) *Life Sciences, Society and Policy*, DOI:[10.1080/11287462.2026.2648374](https://doi.org/10.1080/11287462.2026.2648374).
9. Fairness analysis of machine learning predictions of aggression in acute psychiatric care (2026) *npj Mental Health Research*.
10. Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data (2024) *npj Mental Health Research*.
11. Algorithmic fairness in precision psychiatry: analysis of prediction models in individuals at clinical high risk for psychosis (2026) *The British Journal of Psychiatry*.
12. AI risk prediction tools in psychiatry can reinforce systemic bias (2026) *CAMH Research News*.
13. Data mining for health: staking out the ethical territory of digital phenotyping (2018) *npj Digital Medicine*.
14. Ethical dimensions of digital phenotyping within the context of mental healthcare (2024) *Journal of Technology in Behavioral Science*.
15. First-person disavowals of digital phenotyping and epistemic injustice in psychiatry (2023) *Medicine, Health Care and Philosophy*.
16. Ethical implications of the use of language analysis technologies for the diagnosis and prediction of psychiatric disorders (2022) *JMIR Mental Health*.
17. Algorithmic assessments in deciding on voluntary, assisted or involuntary psychiatric treatment (2022) *Diagnostics*, DOI:[10.3390/diagnostics12081806](https://doi.org/10.3390/diagnostics12081806).
18. Clinicians' use of the structured professional judgement approach for adult secure psychiatric service admission assessments: a systematic review (2024) *PLOS One*.
19. Ethical considerations for using AI to predict suicide risk (2023) *The Hastings Center Bioethics Forum*.
20. Machine learning applications in suicide risk prediction: data sources, modeling techniques, ethical considerations (2025) *European Psychiatry*.