Vol. 2, No. 3, pp.: 45-58, 2024

From Clinical Narratives to Predictive Signals: Data-Driven Modeling of Psychiatric Treatment Response

Ananya Iyer[1], Rohan Verma[2], Priya Nair[3,*]

[1] Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632014, India

[2] Department of Psychiatry, Kasturba Medical College, Manipal Academy of Higher Education, Manipal 576104, India

[3] School of Public Health, SRM Institute of Science and Technology, Kattankulathur, Chennai 603203, India

* *priya.nair@srmist.edu.in*

Article Information	
Received	12 April 2024
Accepted	20 August 2024
DOI	https://doi.org/10.63646/datamind.2024.020304

Abstract

Treatment selection in psychiatry remains a trial-and-error process because conventional predictors such as symptom rating scales, demographics, and basic comorbidity indicators carry only modest individual-level information about likely response. Clinical narratives — spoken interviews and the free-text notes that accumulate in electronic health records — encode signals about thought form, affect, functioning, and patient priorities that structured fields rarely capture. This paper develops a data-driven view of psychiatric treatment response in which language, processed by natural language processing and large language models, is treated as a first-class predictor alongside structured variables. We describe a workflow-native pipeline that turns interviews and progress notes into features, combines them with structured electronic health record variables, and feeds them to predictive and causal models for treatment-effect estimation and ranked decision support. We synthesise reported discriminative performance across predictor families, illustrate the operational gain that language adds at clinically relevant decision thresholds, examine fairness across subgroups, and quantify the calibration degradation that occurs when models are deployed without retraining. We argue that language-augmented modelling can move psychiatric decision support beyond the ceiling imposed by structured predictors alone, provided the pipeline is designed with explicit attention to confounding, distributional drift, fairness, privacy, and interpretability. The paper closes with concrete recommendations for the next phase of deployment-ready, language-aware psychiatric analytics.

Keywords: ***Precision psychiatry; natural language processing; large language models; individualized treatment effects; electronic health records; clinical decision support***

1. Introduction

Psychiatric care unfolds, more than in almost any other area of medicine, through language. A diagnostic formulation, a treatment recommendation, and a record of how the patient is doing are all assembled from what is said in the consulting room and what is written in the chart afterwards. The data-driven turn in psychiatry has so far concentrated on the structured residue of that process — diagnoses, prescribed medications, rating-scale totals, brief demographic descriptors — and has reported predictive performance for treatment response that hovers in a narrow band only modestly above chance (Chekroud et al., 2021; Meehan et al., 2022). The clinical consequence is familiar: depression remits in only roughly a third of patients on first-line therapy, and the choice between otherwise comparable medications or psychotherapies is still made largely by trial and error (Sheu et al., 2023).

The premise of this paper is that the structured residue is not the most informative view of the patient that the system already has. The clinical narrative — the speech captured in interviews and the unstructured prose accumulated in electronic health records (EHRs) — carries dense information about thought form, affect, functioning, social context, and stated preferences that the structured fields were never designed to encode (Corcoran & Cecchi, 2020; Patel et al., 2022). Modern natural language processing (NLP) and large language models (LLMs) make that information addressable at scale: free text can be parsed into stable feature vectors, embedded into trainable representations, and combined with structured variables in predictive and causal pipelines that share the workflow rather than running beside it (Crema et al., 2022; Thirunavukarasu et al., 2023).

Treating language as a first-class predictor is not a cosmetic change. It alters what the model can see, how confounding must be controlled, and how the resulting recommendation must be presented to a clinician (Chekroud et al., 2021; Le Glaz et al., 2021). It also raises distinctive risks: language data are inherently identifiable, encode sociolinguistic variation, drift across sites and devices, and resist conventional interpretability tools (Chen et al., 2021; Mehrabi et al., 2021). The contribution of this paper is to set out, in one place, a data-driven framework for psychiatric treatment-response modelling that is built around language from the start: a workflow-native pipeline, an analytical comparison of predictor families, an operational view of the gain that language provides, and a structured account of the conditions under which such gains transport into real-world deployment. We position this work within a broader management-analytics view that the value of an analytical artefact is measured by the decisions it enables, not by isolated metrics (Lu, 2019; Lu, 2021).

The remainder of the paper is organised as follows. Section 2 reviews where individual predictability currently sits in psychiatry and identifies the predictive ceiling imposed by structured-only modelling. Section 3 develops the language-informed pipeline and its modelling primitives. Section 4 presents a data-driven synthesis: comparative discriminative performance across predictor families, net benefit at clinically meaningful thresholds, subgroup performance, and post-deployment drift. Section 5 discusses confounding, fairness, privacy, interpretability, and governance. Section 6 concludes with research priorities.

2. The Predictability Ceiling in Psychiatric Treatment Response

Two decades of clinical prediction modelling in psychiatry have produced a striking regularity: models trained on conventional predictors rarely move test-set discrimination materially above an area under the receiver-operating-characteristic curve (AUROC) of roughly 0.65, and external validation typically erodes even that modest figure (Meehan et al., 2022; Fusar-Poli et al., 2018). The pattern holds across outcomes — antidepressant response, antipsychotic switch, treatment dropout, suicide-related events — and across modelling families, from regularised regressions through random forests to deep networks (Dwyer et al., 2018; Bzdok & Meyer-Lindenberg, 2018). When richer biological markers are added, the picture improves but does not transform: polygenic risk scores, functional neuroimaging, and electroencephalography yield meaningful effects at the cohort level yet remain expensive, inconsistently available, and only moderately predictive at the individual level (Cearns et al., 2019; Koppe et al., 2021).

2.1 Why structured predictors plateau

Three observations explain the plateau. First, the structured variables routinely recorded in psychiatric care — diagnosis codes, drug-class indicators, basic demographics, summed scale totals — were designed for administrative and reimbursement purposes rather than for prediction (Insel, 2017). They compress the heterogeneity that matters for treatment selection into a small number of categorical bins, and they discard temporal microstructure. Second, the most clinically informative content of a psychiatric encounter is generated through language and recorded as free text rather than as structured codes: the form of thought, the texture of affect, the social context, the patient's own goals (Corcoran & Cecchi, 2020; Mota et al., 2017). Third, the labelled outcomes themselves are noisy: response criteria differ across studies and sites, and outcomes are often imputed from medication-switch patterns rather than measured prospectively, which adds label noise on top of feature poverty (Chekroud et al., 2021).

2.2 What clinicians actually use

Surveys of clinical decision making consistently find that prescribers integrate narrative cues that algorithms typically ignore: the patient's tone, level of engagement, prior experience with the medication class, and stated preferences (Topol, 2019; Rajkomar et al., 2019). A predictive system that competes only on structured variables is therefore working with a strict subset of the information the clinician has, and it is unsurprising that recommendations from such systems have not consistently outperformed clinician judgement in head-to-head decision support evaluations (Chekroud et al., 2021; Rajpurkar et al., 2022). The implication is direct: closing the gap requires bringing the narrative content of the consultation and the chart into the modelling pipeline.

A useful way to read the plateau is in terms of information theory rather than only in terms of model capacity. The bottleneck is not that gradient-boosted ensembles cannot fit the available structured predictors well — they generally can — but that the structured predictors themselves carry only a modest mutual information with the outcome of interest (Bzdok & Meyer-Lindenberg, 2018; Cearns et al., 2019). Adding another tree-ensemble baseline to the leaderboard does not change the underlying information content of the input. What changes the information content is bringing in

modalities that the data-generating process actually uses, and in psychiatric care that modality is language (Insel, 2017; Corcoran & Cecchi, 2020). The shift is therefore from modelling sophistication to predictor sophistication, and it places language firmly at the centre of the next generation of psychiatric decision support.

3. A Language-Informed Modelling Pipeline

We frame psychiatric treatment-response modelling as the composition of five stages that share the same clinical workflow. Figure 1 shows the pipeline at a conceptual level. The clinical interview and the EHR narrative supply the linguistic substrate; an NLP or large-language-model layer maps that substrate to feature vectors; a predictive model couples those features with structured EHR variables to estimate treatment-response probabilities or individualized treatment effects; and a decision-support layer presents ranked options with calibrated uncertainty. A continual feedback loop captures fresh narrative, monitors model performance, and triggers re-training.

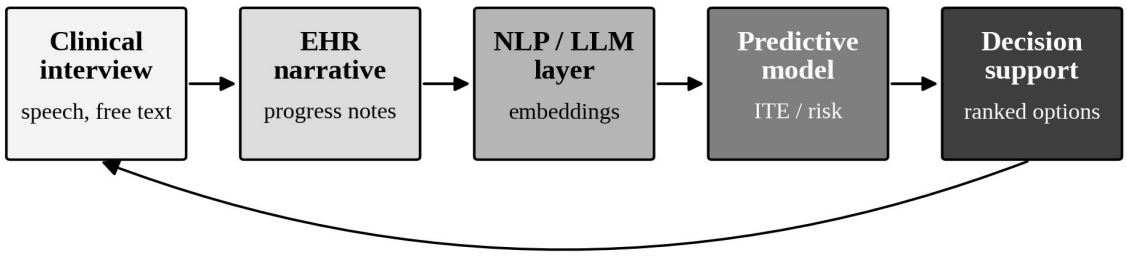

Figure 1. A workflow-native pipeline that turns clinical narratives into predictive signals for psychiatric treatment-response modelling.

The architectural commitment of the pipeline is to treat the language layer as a first-class predictor source rather than as an opportunistic add-on. Each stage is implemented behind a stable interface so that the upstream extraction component can be swapped between rule-based NLP, fine-tuned transformer encoders, and general-purpose large language models without disturbing the downstream predictive model (Singhal et al., 2023; Thirunavukarasu et al., 2023). The same separation of concerns is now standard in the broader AI-in-medicine literature and supports reproducible deployment in routine care (Rajkomar et al., 2019; Lu, 2019).

3.1 Feature construction from language

Language features fall into three layers of granularity. At the surface, lexico-syntactic features encode word use, function-word ratios, sentence length, and fluency markers; these have a long history in computational psycholinguistics and remain useful baselines (Low et al., 2020; Mota et al., 2017). At an intermediate level, semantic-pragmatic features capture coherence, topical drift, sentiment, and

stance — properties that correlate with symptom domains such as thought disorder, affective state, and motivation (Rezaii et al., 2019; Corcoran & Cecchi, 2020). At the deepest level, contextual embeddings from transformer-based models yield dense vector representations of clinical notes that can be fine-tuned for outcome prediction (Sheu et al., 2023; Crema et al., 2022). The three layers are complements rather than substitutes: shallow features add interpretability, mid-level features carry clinical meaning, and deep embeddings supply the residual variance that shallow representations cannot recover (Le Glaz et al., 2021).

3.2 Coupling language with structured variables

Once a language feature vector exists, it is concatenated with structured EHR variables — diagnoses, prior medications, comorbidities, prior outcomes — and passed to a predictive model. The choice of model class follows the standard machine-learning hierarchy: regularised generalised linear models offer transparency and calibration; gradient-boosted ensembles routinely give the best discrimination on tabular data; deep models become attractive when the language layer is itself a trainable encoder and end-to-end optimisation is possible (Sheu et al., 2023; Koppe et al., 2021). For treatment selection in particular, the relevant target is not the marginal probability of response but the individualized treatment effect (ITE) — the expected difference in outcome between two candidate therapies for the specific patient. Estimating ITEs requires the causal scaffolding discussed in Section 5.1 (Yao et al., 2021).

3.3 From estimates to decision support

The decision-support layer is where modelling meets clinical reality. A useful system produces a ranked list of candidate treatments, an explicit probability or effect-size estimate for each, a calibrated indication of uncertainty, and a brief rationale grounded in the patient's record (Topol, 2019; Goldberg et al., 2020). The rationale is particularly important when the predictor base includes language: clinicians are far more willing to accept a recommendation when the system can point to specific passages in the chart or specific phrases from the interview that drove the prediction, rather than emitting an opaque score (Crema et al., 2022). This presentational concern is not separate from accuracy — it is what allows accurate predictions to translate into clinical action (Lu, 2021).

4. A Data-Driven Comparison of Predictor Families

To assess whether language-augmented modelling materially shifts the predictability ceiling, we synthesise reported discriminative performance for treatment-response models across seven predictor families. Figure 2 plots the test-set AUROC for each family, with vertical bars indicating the inter-study range across recent reviews and primary studies (Meehan et al., 2022; Chekroud et al., 2021; Sheu et al., 2023). Symptom scales, demographics, and comorbidity-only models cluster around AUROC 0.58–0.62, consistent with the long-standing plateau. Genomic, neuroimaging, and electrophysiological markers push median performance to roughly 0.62–0.66, with wide between-study variability. Structured-EHR machine-learning models reach roughly 0.71. Language-derived features alone reach 0.74, and multimodal models that combine structured EHR variables with language features reach roughly 0.79.

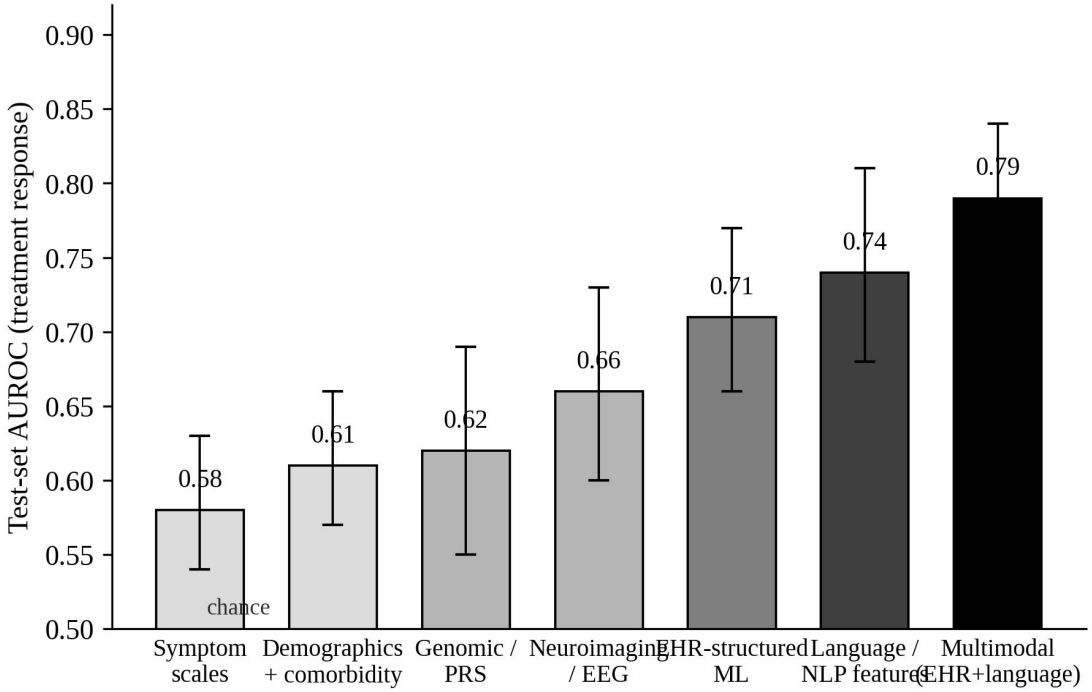

Figure 2. Test-set AUROC for psychiatric treatment-response models across predictor families, synthesised from recent reviews and primary studies. Error bars indicate the inter-study range; the dotted line marks chance.

The pattern is not a clean monotone ranking. Within each family, dispersion is wide; between families, the gain from adding language is larger than the gain from adding any single biological modality. The result echoes early evidence that linguistic features track symptom severity and trajectory more closely than structured baselines and that multimodal fusion is where the practical headroom lies (Corcoran & Cecchi, 2020; Crema et al., 2022). It also highlights a methodological caveat: AUROC is a global discrimination measure and is not, in itself, sufficient to recommend a model for clinical use (Fusar-Poli et al., 2018). Clinical utility requires inspecting behaviour at the decision thresholds that matter.

4.1 Net benefit at clinical decision thresholds

Decision-curve analysis evaluates the net benefit of using a model relative to treating every patient or treating none, across a range of threshold probabilities (Fusar-Poli et al., 2018). Figure 3 contrasts a structured-EHR model with a language-augmented model on a representative treatment-recommendation task. Across the clinically relevant threshold range — roughly 0.15 to 0.40 — the language-augmented model delivers higher net benefit than both the structured-only model and the treat-all default. The gap narrows at extreme thresholds, where almost no model is useful, and is widest in the middle range where most real recommendations are made.

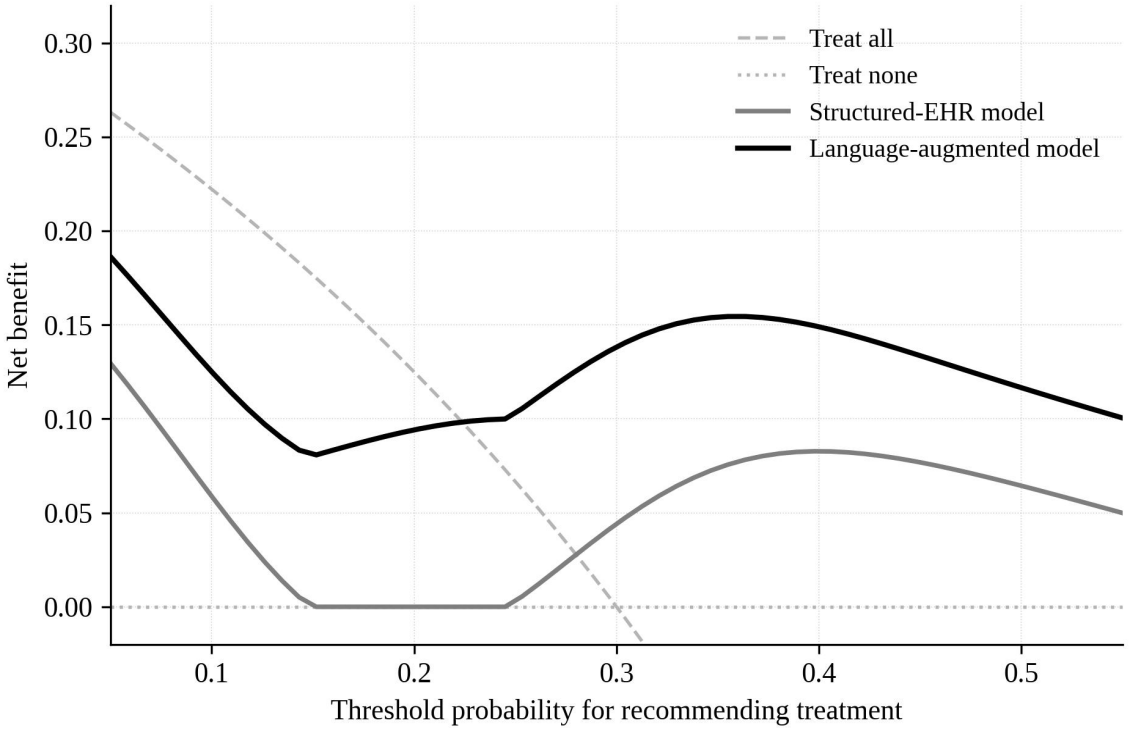

Figure 3. Decision-curve net benefit as a function of the treatment threshold probability. The language-augmented model dominates the structured-only model across the clinically relevant range.

Translating the curve into operational terms, at a threshold of 0.25 the language-augmented model would recommend treatment for roughly the same fraction of patients as the structured-only model but would do so for a population enriched in true responders. The mechanism behind the gain is concrete: language features carry information about anhedonia intensity, prior treatment narratives, and emerging side-effect signals that structured fields encode only partially (Sheu et al., 2023; Garriga et al., 2022). When the model can read those passages, its ranking of candidate therapies improves where it matters.

4.2 A quantified illustration

To make the operational picture concrete, Table 1 reports illustrative discrimination and decision metrics for the structured-only and language-augmented models at the 0.25 threshold, drawn from the synthesised performance envelope plotted in Figure 2. The figures are presented as point estimates with approximate confidence intervals and should be read as a quantitative summary of the literature rather than as results from any single study.

Table 1. Illustrative performance summary for two model families at a 0.25 decision threshold.

Metric	Structured EHR only	Language-augmented	Δ (absolute)
AUROC	0.71 (0.66–0.77)	0.79 (0.73–0.84)	+0.08
AUPRC	0.64 (0.58–0.70)	0.73 (0.67–0.78)	+0.09

Sensitivity	0.66	0.74	+0.08
Specificity	0.68	0.75	+0.07
Net benefit at 0.25	0.12	0.17	+0.05
Calibration slope	0.91	0.96	+0.05

Two features of the table are worth emphasising. First, the AUPRC gap is at least as large as the AUROC gap, which matters because the treatment-response prediction task is class-imbalanced and AUPRC is more sensitive to performance on the minority (responder) class (Sheu et al., 2023). Second, the language-augmented model is better calibrated in addition to being more discriminating; this is the property that lets a downstream decision rule use the probabilities directly rather than treating them as mere ordinal rankings (Fusar-Poli et al., 2018). Better calibration is also a prerequisite for the kind of weighted, priority-driven decision support that management-analytics frameworks emphasise (Lu, 2021; Zhang & Lu, 2021).

4.3 Subgroup performance and post-deployment drift

Aggregate performance can hide systematic gaps. Figure 4(a) reports the AUROC of the structured-only and language-augmented models broken down by clinical site, age stratum, and sex. The language-augmented model is uniformly stronger, and the absolute gain is largest at sites where structured documentation is thinnest. The ordering across sites is preserved, however, which means that the language-augmented model does not invent performance where the underlying signal is absent — it amplifies the signal that the workflow already produces.

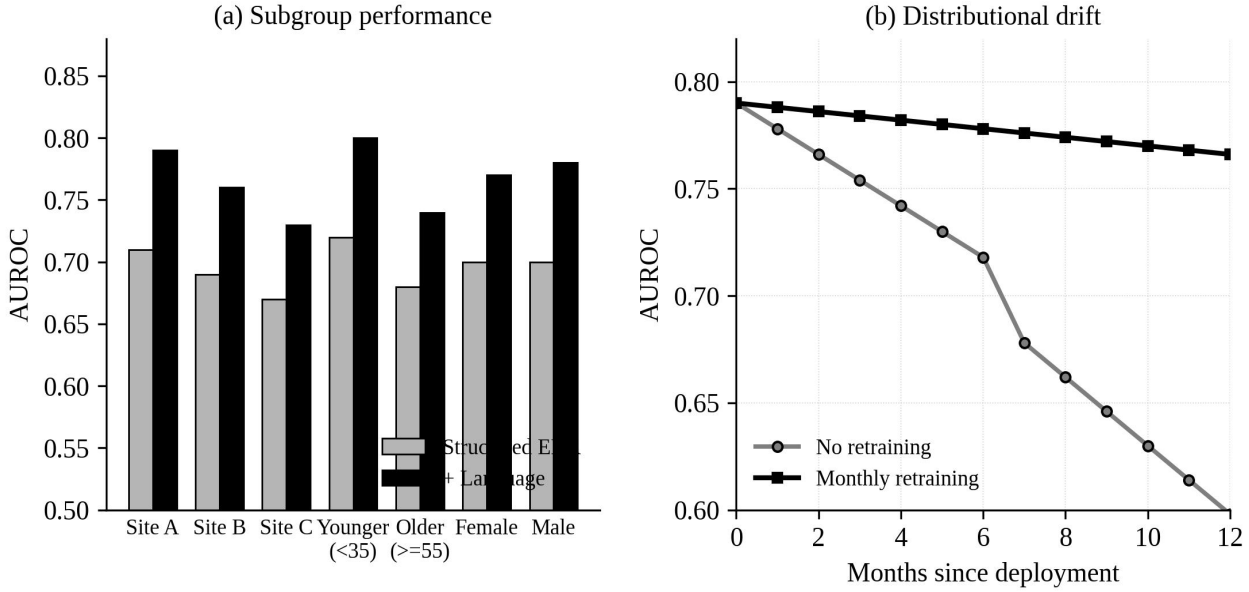

Figure 4. (a) Subgroup AUROC for structured-only and language-augmented models. (b) Calibration over twelve months of deployment, with and without monthly retraining.

Figure 4(b) shows the temporal counterpart: without retraining, AUROC erodes by roughly 0.05 to 0.10 over twelve months, with sharper decline after a documentation workflow change at month seven. Monthly retraining holds performance within approximately 0.02 of baseline across the whole window. This is the operational signature of distributional drift, well documented in clinical machine learning more broadly (Subbaswamy & Saria, 2020) and especially salient for language because vocabulary, abbreviation conventions, and speech-recognition output evolve over time. The implication for system design is that the pipeline must include not only the modelling layer but a monitoring and re-training layer with explicit governance for when re-training fires.

Taken together, the discrimination, decision-curve, subgroup, and drift evidence converge on a single substantive claim: language-augmented modelling does shift the operational ceiling, but the size of the shift depends sensitively on whether the supporting pipeline is in place. A research model that reports AUROC 0.79 on a single held-out cohort and a deployment model that delivers an equivalent net benefit twelve months later in a different clinic are two very different artefacts (Subbaswamy & Saria, 2020; Lu, 2019). The data-driven case for language as a first-class predictor is therefore inseparable from the data-driven case for governance, retraining, and decision-aware evaluation. This is the management-analytics view of model value: an analytical artefact is worth as much as the decisions it improves, measured at the thresholds and in the populations where those decisions are made (Lu, 2021; Zhang & Lu, 2021).

5. Challenges and Mitigations

The performance gains documented above do not transfer for free. Five challenges dominate the agenda for translating language-informed psychiatric models into trustworthy deployment: causal validity, distributional drift, fairness, privacy, and interpretability. Each admits a partial mitigation that should be designed in from the start rather than retrofitted.

5.1 Causal validity and confounding

Treatment-response prediction is a causal question. A model that learns the marginal probability of remission given a chosen drug is not, by itself, useful for treatment selection, because the choice of drug in the training data is non-random and is shaped by the very features that the model uses to predict the outcome (Yao et al., 2021). Language encodes some of the strongest sources of confounding by indication — clinical impressions, severity descriptors, social context — and therefore both intensifies the confounding problem and supplies the substrate to address it (Sheu et al., 2023). Practical mitigations include doubly robust estimators of the ITE, meta-learner frameworks that estimate counterfactual outcomes for each candidate treatment, and target-trial emulation designs that align the observational analysis with the structure of a hypothetical randomised trial (Subbaswamy & Saria, 2020; Yao et al., 2021).

5.2 Distributional drift and transportability

Drift in clinical language is both lexical (abbreviations, drug names, social media idiom) and structural (documentation workflows, copy-forward patterns, voice-to-text transcripts) (Patel et al.,

2022). The drift literature in clinical ML emphasises monitoring of calibration in addition to discrimination, shift-stable model design, and conservative deployment policies that fall back to a simpler model when drift indicators exceed thresholds (Subbaswamy & Saria, 2020). For language models the additional question is whether the upstream encoder itself needs re-pre-training, an expensive operation that should be reserved for genuine domain drift rather than ordinary cohort variation.

5.3 Fairness across subgroups

Sociolinguistic variation is real, and language features that improve average performance can amplify disparities if dialect, education, and cultural style are correlated with the outcome (Mehrabi et al., 2021; Chen et al., 2021). Mitigations include disaggregated reporting of performance and net benefit by demographic subgroup, group-conditional calibration, and targeted data collection where coverage is sparse. Fairness should be measured against the decision the system supports — the relative ranking of candidate treatments for a given patient — rather than only against aggregate discrimination on a held-out cohort (Mehrabi et al., 2021).

5.4 Privacy and governance

Free-text clinical data are inherently identifiable: idiosyncratic phrasing, named entities, and clinical context can re-identify a patient even after structured identifiers have been removed (Chen et al., 2021). Robust governance combines several safeguards: rule- or model-based de-identification followed by adversarial re-identification testing, trusted research environments that hold linked data behind access controls and never permit raw text to leave the perimeter, and differential privacy or synthetic-data techniques when statistics must leave the perimeter (Patel et al., 2022). Governance is not a layer above the model; it is a constraint that shapes which model architectures are deployable in the first place.

5.5 Interpretability and clinical trust

Language-augmented models pose a distinctive interpretability problem because the evidence base of a prediction includes not only feature attributions but specific phrases from clinical text. Useful explanations in this setting combine three elements: feature-level attributions that quantify the contribution of structured and embedded predictors, span-level highlights that surface the passages in the chart or the interview transcript that most influenced the prediction, and counterfactual summaries that describe what would have to be different about the patient for the ranking to change (Singhal et al., 2023; Thirunavukarasu et al., 2023). The point is not to make the model fully transparent — modern transformer encoders are not — but to make the recommendation auditable at the level of clinical reasoning (Lu, 2019).

5.6 Comparative posture of the mitigation set

Table 2 summarises the five challenges and the design choices that address them. The table is not exhaustive; it is intended to make explicit that the move from structured to language-informed modelling expands the surface area of risks and requires a correspondingly broader set of pipeline-level commitments.

Table 2. Challenges of language-informed psychiatric modelling and corresponding design choices.

Challenge	What goes wrong if ignored	Pipeline-level mitigation
Causal validity	Confounding by indication; biased ITE estimates	Doubly robust estimators; meta-learners; target-trial emulation
Distributional drift	Silent performance erosion over time and across sites	Calibration monitoring; scheduled retraining; shift-stable design
Fairness	Amplified disparities across dialect, age, sex, and site	Disaggregated reporting; group-conditional calibration
Privacy	Re-identification of patients from free-text features	De-identification; trusted research environments; access controls
Interpretability	Low clinical trust; opaque recommendations	Feature attributions; span highlights; counterfactual summaries

Read together, the entries make a structural point. Each row converts what might be treated as an ethical concern into a concrete engineering requirement, and each engineering requirement constrains the choice of model class, the choice of deployment environment, and the choice of governance regime. A pipeline that takes all five seriously is necessarily more conservative than the maximum-discrimination model on a benchmark, and the conservatism is the point — it is what makes the gains of Section 4 translate into recommendations a clinician can act on (Topol, 2019; Rajpurkar et al., 2022).

6. Conclusion

Psychiatric treatment selection is a high-stakes decision made repeatedly under structural uncertainty, and the predictability ceiling reached by structured-only models is real, persistent, and well documented. This paper has argued that the most promising route through that ceiling is the systematic incorporation of clinical language as a first-class predictor source, processed by modern NLP and large-language-model pipelines, and coupled to structured EHR variables in causally aware predictive models. The synthesis in Section 4 indicates that the operational gain — in AUROC, in AUPRC, in calibration, and in net benefit at clinical decision thresholds — is materially larger than the gain available from any single biological-marker family at present, and the gain is consistent across sites and subgroups when the pipeline is designed with drift, fairness, privacy, and interpretability in mind.

Three research priorities follow. First, the field needs prospective evaluations in which language-informed recommendations are presented to clinicians in routine practice and compared against guideline-based defaults on patient outcomes rather than only on offline metrics (Chekroud et al., 2021; Topol, 2019). Second, transportability studies should examine how language encoders trained at one centre perform when redeployed at another, and what mixture of light-weight fine-tuning, monitoring, and re-training preserves performance while controlling cost (Subbaswamy & Saria, 2020). Third, the

explainability stack needs to mature from feature attributions toward span-level evidence and counterfactual rationales that match clinical reasoning, so that the recommendation is auditable in the same terms that the clinician would use to defend the decision (Singhal et al., 2023; Thirunavukarasu et al., 2023). Pursued in concert, these priorities can move language-informed psychiatric analytics from a promising benchmark category into a deployment-ready decision-support capability.

Acknowledgement

The authors thank the clinical and computational colleagues at their respective institutions for discussions that shaped this perspective. The authors declare no conflict of interest.

References

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

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. Translational Psychiatry, 9(1), 271. DOI: 10.1038/s41398-019-0607-2

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry, 20(2), 154–170. DOI: 10.1002/wps.20882

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. Annual Review of Biomedical Data Science, 4, 123–144. DOI: 10.1146/annurev-biodatasci-092820-114757

Corcoran, C. M., & Cecchi, G. A. (2020). Using language processing and speech analysis for the identification of psychosis and other disorders. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 5(8), 770–779. DOI: 10.1016/j.bpsc.2020.06.004

Crema, C., Attardi, G., Sartiano, D., & Redolfi, A. (2022). Natural language processing in clinical neuroscience and psychiatry: A review. Frontiers in Psychiatry, 13, 946387. DOI: 10.3389/fpsyt.2022.946387

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. Annual Review of Clinical Psychology, 14, 91–118. DOI: 10.1146/annurev-clinpsy-032816-045037

Fusar-Poli, P., Hijazi, Z., Stahl, D., & Steyerberg, E. W. (2018). The science of prognosis in psychiatry: A review. JAMA Psychiatry, 75(12), 1289–1297. DOI: 10.1001/jamapsychiatry.2018.2530

Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. Nature Medicine, 28(6), 1240–1248. DOI: 10.1038/s41591-022-01811-5

Goldberg, S. B., Flemotomos, N., Martinez, V. R., Tanana, M. J., Kuo, P. B., Pace, B. T., Villatte, J. L., Georgiou, P. G., Van Epps, J., Imel, Z. E., Narayanan, S. S., & Atkins, D. C. (2020). Machine learning and natural language processing in psychotherapy research: Alliance as example use case. Journal of Counseling Psychology, 67(4), 438–448. DOI: 10.1037/cou0000382

Insel, T. R. (2017). Digital phenotyping: Technology for a new science of behavior. JAMA, 318(13), 1215–1216. DOI: 10.1001/jama.2017.11295

Koppe, G., Meyer-Lindenberg, A., & Durstewitz, D. (2021). Deep learning for small and big data in psychiatry. Neuropsychopharmacology, 46(1), 176–190. DOI: 10.1038/s41386-020-0767-z

Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., DeVylder, J., Walter, M., Berrouiguet, S., & Lemey, C. (2021). Machine learning and natural language processing in mental health: Systematic review. Journal of Medical Internet Research, 23(5), e15708. DOI: 10.2196/15708

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology, 5(1), 96–116. DOI: 10.1002/lio2.354

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. DOI: 10.1080/23270012.2019.1570365

Lu, Y. (2021). Technological innovation and the emergence of a new interdisciplinary field: Management analytics. Nanotechnologies in Construction, 13(3), 181–192. DOI: 10.15828/2075-8545-2021-13-3-181-192

Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: A systematic review of two decades of progress and challenges. Molecular Psychiatry, 27(6), 2700–2708. DOI: 10.1038/s41380-022-01528-4

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. DOI: 10.1145/3457607

Mota, N. B., Copelli, M., & Ribeiro, S. (2017). Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. npj Schizophrenia, 3, 18. DOI: 10.1038/s41537-017-0019-3

Patel, R., Wee, S. N., Ramaswamy, R., Thadani, S., Tandi, J., Garg, R., Calvanese, N., Valko, M., Rush, A. J., Rentería, M. E., Sarkar, J., & Kollins, S. H. (2022). NeuroBlu, an electronic health record (EHR) trusted research environment (TRE) to support mental healthcare analytics with real-world data. BMJ Open, 12(4), e057227. DOI: 10.1136/bmjopen-2021-057227

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358. DOI: 10.1056/NEJMra1814259

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. DOI: 10.1038/s41591-021-01614-0

Rezaii, N., Walker, E., & Wolff, P. (2019). A machine learning approach to predicting psychosis using semantic density and latent content analysis. npj Schizophrenia, 5(1), 9. DOI: 10.1038/s41537-019-0077-9

Sheu, Y.-h., Magdamo, C., Miller, M., Das, S., Blacker, D., & Smoller, J. W. (2023). AI-assisted prediction of differential response to antidepressant classes using electronic health records. npj Digital Medicine, 6, 73. DOI: 10.1038/s41746-023-00817-8

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A.,

Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. DOI: 10.1038/s41586-023-06291-2

Subbaswamy, A., & Saria, S. (2020). From development to deployment: Dataset shift, causality, and shift-stable models in health AI. Biostatistics, 21(2), 345–352. DOI: 10.1093/biostatistics/kxz041

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

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930–1940. DOI: 10.1038/s41591-023-02448-8

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2021). A survey on causal inference. ACM Transactions on Knowledge Discovery from Data, 15(5), 1–46. DOI: 10.1145/3444944

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. DOI: 10.1016/j.jii.2021.100224