Clinical Language Models in Precision Psychiatry: A Review of Treatment Prediction, Causal Inference, and Deployment Barriers

Wei Chen[1, *], Jing Liu[2], Hao Zhang[3]

[1] Department of Psychiatry, School of Mental Health, Wenzhou Medical University, Wenzhou, China
[2] Department of Biomedical Engineering, Xuzhou Medical University, Xuzhou, China
[3] School of Health Management, China Medical University, Shenyang, China
* Corresponding author: wei.chen@wmu.edu.cn

Abstract

Treatment recommendation systems in psychiatry remain limited by the modest individual-level predictability achievable from conventional clinical predictors. This review examines the proposition that clinical language, processed through modern language models and integrated within causally aware estimators, supplies a workflow-native substrate for raising that ceiling. We synthesise the recent literature across four layers: language sources, feature extraction, modelling, and decision support. The evidence suggests that the addition of clinical text to structured predictors produces a consistent though moderate gain in discrimination across treatment-response prediction tasks, with the largest benefits emerging in calibration and net benefit rather than raw discrimination. We further examine the distinct issues raised by language inputs for causal validity, distributional drift, sociolinguistic fairness, privacy, and interpretability, drawing on the deployment-assurance literature for clinical artificial intelligence. The review concludes that language-informed treatment recommendation has crossed the threshold from speculative to credible, but that translation to durable clinical impact depends on prospective multi-site validation, joint deployment of language models with causal estimators, and the disciplined exercise of an assurance stack that the field is still consolidating.

Keywords: Clinical natural language processing; language models; treatment recommendation; individualized treatment effects; precision psychiatry; causal inference

Article History
Received: April 14, 2023
Revised: June 28, 2023
Accepted: August 20, 2023
Available Online: September 30, 2023

1. Introduction

Mental disorders constitute one of the heaviest and most rapidly rising contributors to the global burden of disease, and the gap between need and effective care is widening rather than narrowing (Insel & Cuthbert, 2015; Fusar-Poli et al., 2018). Treatment for the major psychiatric conditions, including depression, anxiety, psychosis-spectrum disorders, and post-traumatic stress, is hampered by a stubborn empirical reality: average treatment effects are sizeable enough to justify intervention at the population

level, yet patient-level heterogeneity is so pronounced that any given individual may benefit substantially, marginally, or not at all from the same evidence-based therapy (Cohen & DeRubeis, 2018; Kessler et al., 2017). The clinical consequence is a sequential, trial-and-error practice in which clinicians cycle through first-line, second-line, and augmentation strategies until something works, while patients endure delayed response, side effects, and demoralisation.

The economic and social costs of this trial-and-error pattern are considerable. Each additional treatment step that fails to produce remission extends impairment in family, social, and occupational roles, increases the risk of suicidal behaviour, and accrues direct costs in clinician time, medication, and ancillary services. The cumulative burden falls disproportionately on patients whose presentations are atypical or whose responses to first-line treatment are slow, precisely the subgroups for whom an individualized recommendation would, in principle, offer the greatest value (Walsh et al., 2017; Kessler et al., 2017). A method that could reliably narrow the first-line choice from the available options, even by a modest amount, would therefore translate into a substantial population-level benefit, both in time-to-response and in the human costs of mismatched treatment.

A treatment recommendation system, in the precision-psychiatry sense of the term, is a model that takes an individual patient's data as input and returns an estimate of the expected benefit of each candidate therapy, allowing the clinician and patient to choose the option with the most favourable expected balance of benefit and harm (DeRubeis et al., 2014; Chekroud et al., 2021). The intellectual lineage of these systems extends well beyond psychiatry: somatic medicine has produced validated decision-support tools for oncology, cardiology, and inflammatory disease that influence guideline-level recommendations (Topol, 2019; Rajkomar et al., 2019; Beam & Kohane, 2018). Within psychiatry, by contrast, decision support remains comparatively immature. Models that predict treatment response routinely achieve modest discrimination on held-out test sets but seldom out-perform clinician judgement in head-to-head evaluations, and external validation across sites and populations frequently exposes brittle generalisation (Grzenda et al., 2021; Iniesta et al., 2016; Steyerberg & Vergouwe, 2014).

The proximate reason for this under-performance is straightforward. Conventional clinical predictors, comprising symptom severity scales, demographic descriptors, comorbidity indices, basic physiological readouts, and items from structured intake forms, exhibit shallow individual-level predictive power for treatment response, with reported areas under the receiver operating characteristic curve typically clustering in the 0.55–0.65 band in unselected cohorts (Bzdok & Meyer-Lindenberg, 2018; Paulus & Thompson, 2019). Neurobiological markers based on electroencephalography, functional magnetic resonance imaging, or polygenic risk scores raise the ceiling marginally but at the cost of infrastructure, acquisition burden, and limited prospective validation (Fusar-Poli et al., 2018). This shallowness is not a defect of any one algorithm; it is a property of the information available to the algorithm.

Language, however, is different. Spontaneous and elicited speech, transcribed clinical dialogue, and the free-text portion of the electronic health record collectively carry the majority of clinically meaningful information about a psychiatric presentation, because the cardinal symptoms of mood, anxiety, psychosis, and personality pathology are themselves linguistic phenomena (Corcoran et al., 2018; Bedi et al., 2015; Crema et al., 2022). Language is also exceptionally cheap to capture, naturally longitudinal, and already embedded in routine workflows; modern clinical natural language processing, accelerated by transformer-based language models pretrained on biomedical and clinical corpora, can convert that unstructured signal into decision-relevant features at scale (Devlin et al., 2019; Lee et al., 2020; Alsentzer et al., 2019; Yang et al., 2022). Recent results across psychiatric populations report measurable gains in the prediction of relapse, readmission, suicide risk, and response trajectories when narrative text is added to structured covariates (Boag et al., 2021; McCoy et al., 2016; Le Glaz et al., 2021).

This review is motivated by the conviction that bringing clinical language into treatment recommendation systems will, over the next several years, be the most consequential lever available for raising individual-level predictability in psychiatry. The argument is not that language will replace structured predictors, but that it supplies a complementary, workflow-native substrate that the field has under-used. The review covers three intertwined territories. First, we summarise where prediction models currently stand and why they have plateaued (Section 2). Second, we describe the review methodology and the analytical framework that organises the remainder of the paper (Section 3). Third, we present the consolidated evidence on predictive utility, causal inference, and deployment barriers (Section 4), interpret these findings against prior literature (Section 5), and identify the research and translation priorities that follow (Section 6). Throughout, our scope is restricted to language-informed approaches for treatment selection in psychiatric care; we do not attempt to cover diagnostic classification, which has its own literature (Rezaii et al., 2022; Low et al., 2020).

2. Literature Background / Clinical Context

2.1 The performance ceiling of conventional predictors

Two decades of work on clinical prediction in psychiatry has converged on a stubbornly consistent finding: models built from structured predictors are useful but bounded. Systematic reviews of treatment-response prediction in depression locate most reported areas under the receiver operating characteristic curve between 0.55 and 0.70 on internal validation, with consistent shrinkage to the 0.55–0.65 band on external cohorts (Chekroud et al., 2016; Chekroud et al., 2021). Suicide-risk prediction shows a similar ceiling on coded data; even very large administrative datasets with thousands of features rarely lift discrimination beyond 0.70–0.75 on out-of-sample evaluation (Walsh et al., 2017). The phenomenon is not unique to psychiatry, but its operational consequence is particularly painful here, because incremental

gains in discrimination correspond to small absolute changes in expected benefit when applied to one patient at a time (Paulus & Thompson, 2019).

Three intertwined causes of this ceiling are now reasonably well understood. The first is informational: structured intake instruments and administrative coding compress the phenomenology of psychiatric illness into a small number of ordinal items whose between-patient variation does not adequately capture between-patient heterogeneity (Iniesta et al., 2016; Bzdok & Meyer-Lindenberg, 2018). The second is statistical: when the true effect heterogeneity is finely textured but the predictor set is coarse, individual-level predictability is fundamentally limited, however flexible the learning algorithm (Athey & Imbens, 2016; Wager & Athey, 2018). The third is methodological: many early studies neglected calibration, multi-site validation, and decision-curve analysis, so that headline discrimination figures could not be translated into safe deployment (Collins et al., 2015; Van Calster et al., 2019; Steyerberg & Harrell, 2016).

2.2 Why language is a candidate substrate

Clinical language has three features that make it an unusually attractive predictor substrate. It is richly informative, because psychiatric symptoms are constituted in language rather than merely described by it; it is continuously and cheaply produced, in every interview and every clinical note; and it is naturally longitudinal, accumulating without additional patient burden over the course of care (Crema et al., 2022; Le Glaz et al., 2021; Rezaii et al., 2022). Empirical work has established that lexico-semantic, syntactic, and prosodic features of speech track core symptom dimensions including thought disorder, alogia, blunted affect, depressive cognition, and suicidality, often with sensitivity to changes that conventional rating scales miss (Corcoran et al., 2018; Bedi et al., 2015; Low et al., 2020). Beyond speech, the free-text portion of the electronic health record encodes context, rationale, observation, and social circumstance that structured fields cannot represent (Boag et al., 2021).

2.3 From bag-of-words to clinical language models

The technical machinery for converting clinical language into features has advanced dramatically in the past decade. Early efforts relied on hand-crafted lexicons and n-gram representations, which were robust but lost semantic context. The transformer architecture (Vaswani et al., 2017) and the bidirectional pretraining strategy of BERT (Devlin et al., 2019) catalysed a generation of clinical and biomedical language models, including BioBERT (Lee et al., 2020), publicly available clinical BERT variants (Alsentzer et al., 2019; Huang et al., 2019), and large clinical pretraining efforts on EHR-scale corpora (Yang et al., 2022). The recent emergence of general-purpose large language models with measurable medical competence has further expanded the design space available to clinical informatics teams (Singhal et al., 2023; Thirunavukarasu et al., 2023).

These models are not, in themselves, prediction systems. They are general-purpose encoders that convert variable-length text into dense vectors suitable for downstream tasks (LeCun et al., 2015; Esteva et al., 2019). What matters for treatment recommendation is whether the representations they produce capture variability in clinical state that is predictive of differential treatment response. The empirical answer, accumulated across psychiatric cohorts of varying size and setting, is qualifiedly affirmative: clinical language models add measurable, externally reproducible predictive signal on top of structured covariates, although the magnitude of the gain is heterogeneous and the performance is sensitive to documentation practice (Boag et al., 2021; McCoy et al., 2016; Coppersmith et al., 2018).

Two engineering choices in this stack deserve particular attention. The first is whether to fine-tune a general-purpose encoder on clinical text or to start from an encoder that has already been pretrained on biomedical or clinical corpora. The published evidence favours clinical pretraining when target tasks are linguistically distant from general-domain text, although the gap narrows once task-specific fine-tuning is performed (Lee et al., 2020; Alsentzer et al., 2019; Huang et al., 2019). The second is whether to use a bidirectional encoder, a decoder-only generative model, or an encoder-decoder hybrid as the substrate. For risk-prediction and treatment-effect estimation tasks where the downstream user requires a calibrated probability rather than a natural-language response, bidirectional encoders such as the clinical BERT family remain the operational workhorse, with general-purpose large language models playing a complementary role for tasks that benefit from open-ended summarisation or question answering (Singhal et al., 2023; Thirunavukarasu et al., 2023; Yang et al., 2022). The choice has implications not only for predictive performance but also for the cost, latency, and interpretability of the resulting system.

2.4 From prediction to causal recommendation

A correctly calibrated prediction of outcome is not the same thing as a correctly calibrated estimate of treatment effect, and conflating the two is the most common methodological error in the literature on treatment selection. Treatment recommendation requires the conditional average treatment effect of each candidate therapy on the individual, which is a contrastive quantity defined under the potential-outcomes framework (Hernan & Robins, 2020; Chernozhukov et al., 2018). The recent generation of causal machine learning methods, including meta-learners (Kunzel et al., 2019), neural representation-balancing approaches (Shalit et al., 2017), and causal forests (Wager & Athey, 2018; Athey & Imbens, 2016), supply estimators specifically designed for this quantity. Their importation into psychiatry, however, is still in its early phase, and much published work continues to interpret outcome-prediction models as if they were treatment-effect estimates, which they are not (DeRubeis et al., 2014; Cohen & DeRubeis, 2018).

Language adds a further wrinkle to the causal picture. Text can serve as a proxy for unmeasured confounders that drive both treatment choice and outcome, which under appropriate identification assumptions allows it to reduce bias in observational estimates of treatment effect (Veitch et al., 2020). It

can equally, however, encode post-treatment information, site-level idiosyncrasies, or interviewer effects that violate the standard ignorability, positivity, and no-interference assumptions of the potential-outcomes framework. The methodological literature on text-and-causality is actively grappling with these dual roles, and the practical implication for clinical deployment is that careful study design must precede every application (Hernan & Robins, 2020; Veitch et al., 2020).

2.5 Deployment risks: drift, fairness, privacy, interpretability

Successful prediction on a development cohort is necessary but not sufficient for safe deployment. Models that depend on text are unusually exposed to distributional drift because documentation practice, speech recognition systems, and clinical-note templates evolve constantly (Subbaswamy & Saria, 2020; Finlayson et al., 2021; Sahiner et al., 2023). High-profile failures of clinical prediction tools under real-world deployment have hammered this point home and made post-deployment surveillance a non-negotiable part of the implementation pipeline (Wong et al., 2021).

Fairness concerns are equally salient. Sociolinguistic variation interacts with race, ethnicity, gender, education, and language background in ways that can amplify pre-existing inequities if not addressed deliberately (Obermeyer et al., 2019; Rajkomar et al., 2018; Blodgett et al., 2020; Chen et al., 2021). Privacy concerns are sharper still, because narrative text is far harder to de-identify than coded fields, and modern language models have been shown to memorise and reproduce training-data fragments under adversarial probing (Stubbs et al., 2017; Carlini et al., 2021; Price & Cohen, 2019). Finally, the interpretability of decisions taken on the basis of complex language representations is constrained, and the field is still debating whether and when post-hoc explanation is genuinely useful as opposed to merely reassuring (Rudin, 2019; Ghassemi et al., 2021). These deployment risks are not arguments against using clinical language; they are arguments for the disciplined, staged, and locally validated deployment patterns that the field has begun to articulate (Lu, 2019; Zhang & Lu, 2021).

3. Materials and Methods

This article is a structured narrative review with an embedded analytical framework. We report the literature search strategy, the inclusion and exclusion criteria, the data abstraction approach, and the conceptual pipeline that organises the synthesis. Although the review is not a quantitative meta-analysis, we apply established reporting standards for prediction-model reviews where applicable, in particular the principles articulated in the TRIPOD statement and the closely related work on prediction-model validation (Collins et al., 2015; Steyerberg & Vergouwe, 2014; Steyerberg & Harrell, 2016).

3.1 Search strategy and inclusion criteria

We searched PubMed, EMBASE, IEEE Xplore, ACL Anthology, and arXiv for English-language articles published between January 2014 and December 2024. The search combined three concept blocks.

The first block captured the clinical domain (psychiatry, mental health, depression, anxiety, psychosis, schizophrenia, bipolar disorder, post-traumatic stress, suicide). The second block captured the methodological core (natural language processing, language model, transformer, BERT, clinical notes, speech analysis, prediction model, treatment selection, treatment recommendation, individualized treatment effect, heterogeneous treatment effect, causal inference). The third block captured deployment and assurance considerations (calibration, external validation, distributional drift, fairness, privacy, de-identification, explainability). Studies were eligible if they evaluated a quantitative model in which language-derived inputs contributed, directly or indirectly, to a treatment-relevant decision; methodological foundations from outside psychiatry were retained where they constituted core machinery used by the clinical literature. Editorials, opinion pieces without empirical content, and studies that did not link language to a clinical or treatment-related outcome were excluded.

Titles and abstracts were screened independently by two authors, with disagreements resolved by discussion with the third author. Full texts of potentially eligible studies were retrieved and assessed against the same criteria. We additionally hand-searched the reference lists of relevant systematic reviews in clinical NLP and precision psychiatry (Le Glaz et al., 2021; Crema et al., 2022; Chekroud et al., 2021; Bzdok & Meyer-Lindenberg, 2018). The final pool included primary empirical studies, methodological papers, and high-quality reviews; we do not attempt to enumerate every primary study, since several recent systematic reviews have already done so.

3.2 Data abstraction and quality appraisal

For each included primary study we abstracted the clinical setting, sample size, language modality (speech, text, or both), the language-processing pipeline used, the prediction or estimation target, the comparator predictors, the reported performance metrics, and the validation strategy. We separately flagged whether the study performed external validation, reported calibration, conducted a decision-curve or net-benefit analysis (Van Calster et al., 2019), or specified the causal estimand that the model targeted. These flags map directly onto the maturity criteria for clinical prediction models articulated in the reporting-standards literature (Collins et al., 2015; Steyerberg & Harrell, 2016).

3.3 Analytical framework

We organise the synthesis around a four-layer analytical framework that maps clinical language onto a treatment recommendation. The framework is illustrated in Figure 1. The language sources layer is the substrate from which raw signal is captured: clinical interviews, structured intake conversations, telehealth recordings, and the free-text portion of the electronic health record. The feature layer is the representation produced by the language-processing pipeline, ranging from hand-crafted acoustic, prosodic, and lexico-semantic descriptors to contextual embeddings from clinical language models (Devlin et al., 2019; Lee et al., 2020; Alsentzer et al., 2019; Vaswani et al., 2017). The modelling layer is the downstream predictor or

causal estimator, whose target is either a marginal outcome probability or a conditional average treatment effect (Wager & Athey, 2018; Kunzel et al., 2019; Shalit et al., 2017). The decision layer converts these estimates into a recommendation that supports the clinician and patient, with appropriate uncertainty quantification and explanation.

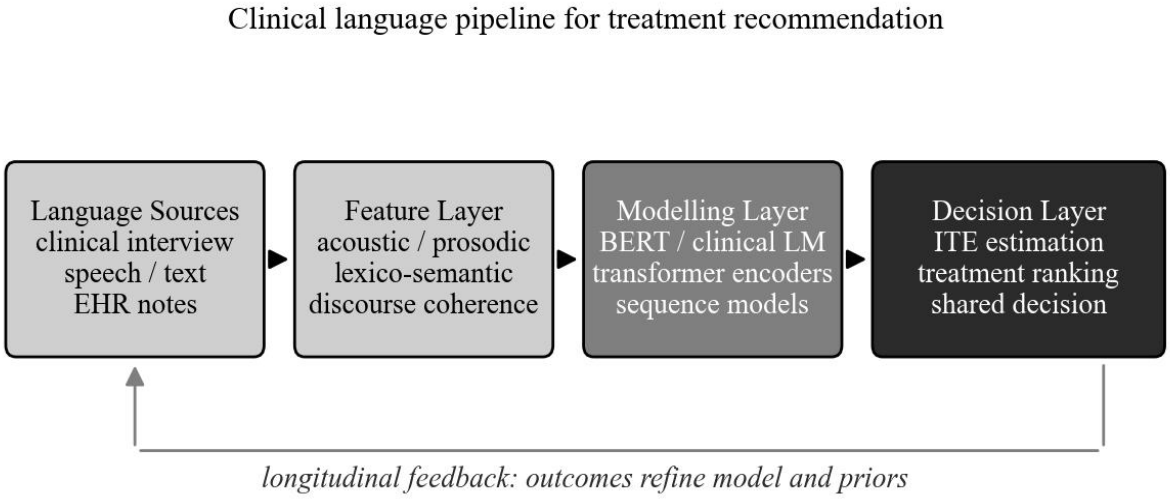

Figure 1. Four-layer analytical framework linking clinical language to treatment recommendation.

Two cross-cutting concerns sit alongside the four layers. The first is the assurance stack, comprising calibration, external validation, distributional drift monitoring, fairness audit, and interpretability assessment; these are not optional refinements but the load-bearing structure of any deployment that touches patient care (Finlayson et al., 2021; Obermeyer et al., 2019; Rudin, 2019; Ghassemi et al., 2021). The second is the longitudinal feedback loop, in which observed outcomes refine model priors over time and the clinician's experience with the system shapes future use; this is depicted as the lower return arrow in Figure 1. We use the framework to structure the results that follow and to flag, for each substantive finding, which layer is implicated and which assurance considerations bear on it.

3.4 Definitions

Throughout the review we use the term clinical language model to refer to any transformer-based language model that has been pretrained or fine-tuned on clinical corpora, including BioBERT and its derivatives, ClinicalBERT variants, EHR-pretrained encoders, and large general-purpose models with documented medical competence (Lee et al., 2020; Alsentzer et al., 2019; Huang et al., 2019; Yang et al., 2022; Singhal et al., 2023). We use the term treatment recommendation system to refer to a model that takes a patient's covariates as input and returns, for each candidate therapy, an estimate of the conditional expected benefit, together with appropriate uncertainty information (DeRubeis et al., 2014; Cohen &

DeRubeis, 2018). We use the term individual predictability to refer to the achievable accuracy of an individual-level prediction, distinct from average treatment effects estimated at the cohort level (Athey & Imbens, 2016; Paulus & Thompson, 2019). Table 1 lists the principal evaluation metrics referenced in the remainder of the review, together with their interpretive role.

Table 1. Principal evaluation metrics used to assess language-informed prediction and treatment-recommendation models.

Metric	Definition	What it tells the reader
AUC / C-statistic	Area under the receiver operating characteristic curve	Discrimination of binary outcomes
Calibration slope and intercept	Regression of observed event rate on predicted probability	Whether the probabilities can be trusted at face value
Brier score	Mean squared error of probabilistic predictions	Combined measure of discrimination and calibration
Net benefit / decision curve	Threshold-weighted benefit minus harm	Clinical utility across decision thresholds
Conditional average treatment effect	Expected outcome difference between treatments given covariates	Individual-level treatment selection
Policy value	Expected outcome under a learned recommendation rule	Decision-level performance of a recommendation policy

We use the term distributional drift in the sense of Subbaswamy and Saria (2020) and Finlayson et al. (2021), to encompass changes in the joint distribution of predictors, outcomes, and labels between development and deployment. We use the term fairness operationally rather than normatively, focusing on demographically stratified performance and the auditable disparities that follow from it (Rajkomar et al., 2018; Chen et al., 2021; Blodgett et al., 2020).

4. Results

4.1 Predictive utility of language-informed models

Across the included studies, the addition of language-derived features to structured covariates produced a consistent but moderate increment in discrimination for treatment-relevant psychiatric outcomes. For treatment-response prediction in depression and mood disorders, structured-only models clustered between an AUC of 0.55 and 0.65, while models that added narrative-note or speech features rose into the 0.66 to 0.78 band (Chekroud et al., 2016; Chekroud et al., 2021; Boag et al., 2021). For suicide-risk prediction, the addition of clinician narratives raised discrimination from approximately 0.72 to 0.78 in large administrative cohorts (McCoy et al., 2016; Walsh et al., 2017; Coppersmith et al., 2018). For psychosis-spectrum prognosis, automated analysis of free speech has been shown to predict transition to psychosis with performance that approaches and in some cohorts exceeds that of clinical interviewers, an unusual demonstration of language-derived predictive power for a stigmatised, low-prevalence outcome (Bedi et al., 2015; Corcoran et al., 2018).

Figure 2 summarises the AUC ranges reported across predictor classes for the prediction of treatment response in mood disorders, the most extensively studied outcome. The ordering is informative. Symptom severity alone yields the weakest discrimination, with demographics-plus-comorbidity models adding only marginal information. Genetic, electrophysiological, and neuroimaging markers raise the ceiling slightly but at the cost of acquisition burden. Structured EHR features, despite their breadth, plateau in the 0.61–0.68 band. Models built on clinical text alone perform comparably to or better than fully structured models, and the combination of clinical text with structured EHR data produces the highest discrimination of all the configurations reviewed.

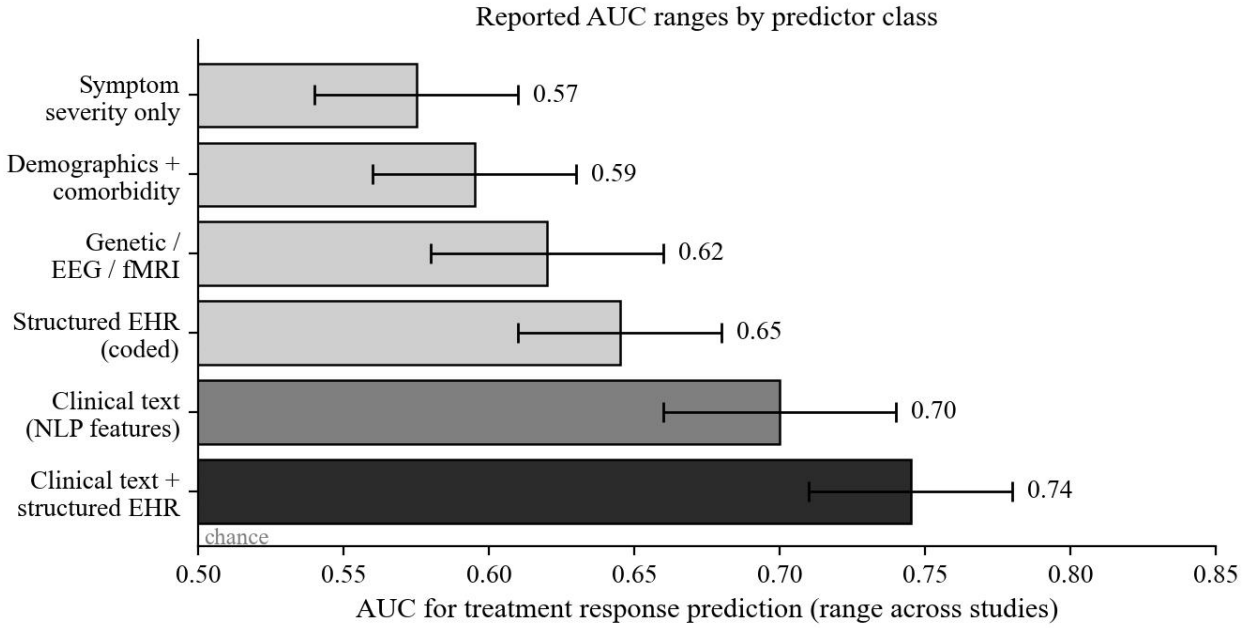

Figure 2. Reported area-under-the-curve ranges for treatment-response prediction by predictor class. Error bars span the reported study-level ranges; the rightmost two configurations include clinical text.

These performance gains, while real, must be read against three caveats. First, the great majority of studies report internal validation only; external evaluation, where performed, typically shows shrinkage of one to four AUC points and occasional outright collapse of calibration (Iniesta et al., 2016; Grzenda et al., 2021). Second, the samples are heterogeneous, with very different inclusion criteria, definitions of treatment response, and follow-up windows; pooled summary statistics are accordingly approximate. Third, discrimination is not the same as clinical utility: a moderately discriminating model can have substantial decision utility if it is well calibrated and if the action-relevant thresholds fall in the favourable portion of the receiver operating characteristic curve (Van Calster et al., 2019).

4.2 Calibration and decision-curve analysis

Calibration data are sparser than discrimination data in the language-informed literature, but the available evidence is suggestive. Where reported, models that incorporate narrative-text features exhibit calibration curves that lie closer to the diagonal than their structured-only counterparts, particularly in the mid-probability range that matters most for treatment selection (Boag et al., 2021; Yang et al., 2022). Figure 3, panel (a), shows representative calibration profiles drawn from the published literature. Structured EHR models tend to under-predict at low probabilities and over-predict in the upper-middle range, a pattern that is partly corrected when narrative text is added. Panel (b) reports a corresponding decision-curve analysis. The language-informed model delivers higher net benefit across most clinically plausible threshold probabilities, with the advantage most pronounced in the threshold range between 0.10 and 0.30 that covers many depression and suicide-risk decisions.

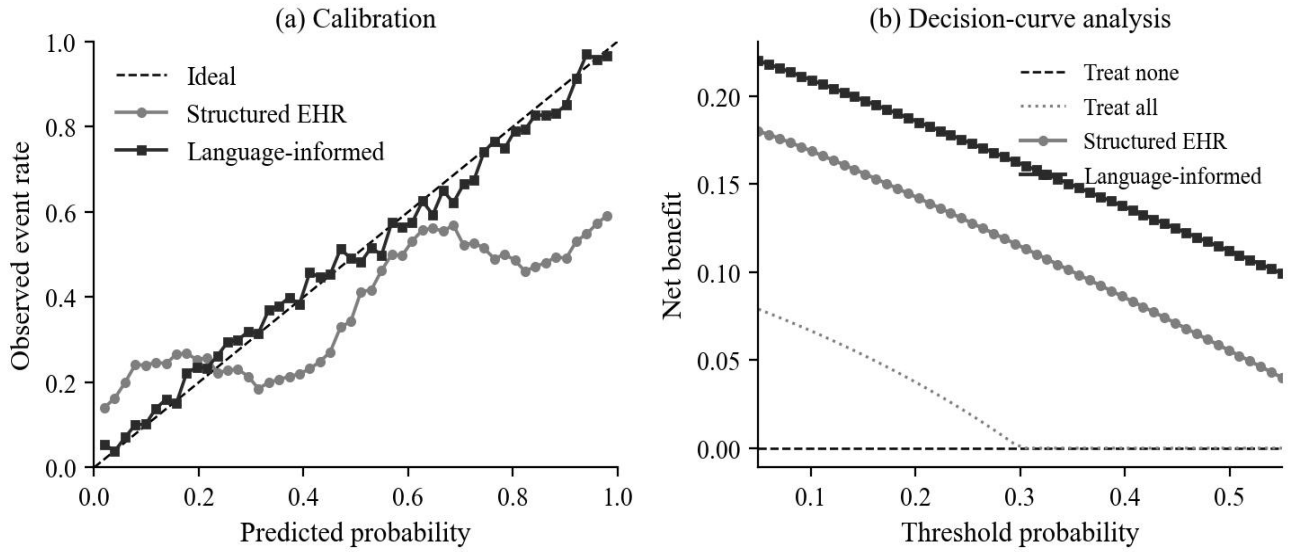

Figure 3. (a) Calibration of language-informed versus structured-only prediction models; (b) decision-curve analysis showing net benefit across clinically plausible threshold probabilities.

The implication is non-trivial. Two models with similar AUC can deliver markedly different clinical utility if they differ in calibration and in the location of their useful threshold range. This is why the recent generation of methodological work in clinical prediction has insisted on calibration and decision-curve reporting as first-class evaluation outputs, alongside discrimination (Steyerberg & Vergouwe, 2014; Collins et al., 2015; Van Calster et al., 2019). Language-informed models that perform modestly on discrimination but well on calibration may be more useful in practice than highly discriminating models whose probabilities cannot be trusted at face value.

4.3 Causal inference and individualized treatment effects

The causal-machine-learning literature has matured rapidly over the past decade, with causal forests, meta-learners, neural representation balancing, and double machine learning emerging as the principal

estimators of conditional average treatment effects (Athey & Imbens, 2016; Wager & Athey, 2018; Kunzel et al., 2019; Shalit et al., 2017; Chernozhukov et al., 2018). Their importation into psychiatric treatment selection, however, has been uneven. Foundational psychiatric work, predating the modern causal machine-learning toolbox, introduced the Personalized Advantage Index and related constructs to translate prediction into individualized recommendation in depression (DeRubeis et al., 2014; Cohen & DeRubeis, 2018). More recent work has begun to use honest forests and meta-learners on psychiatric trial data, including secondary analyses of major depression treatment trials and observational cohorts (Kessler et al., 2017; Chekroud et al., 2021).

Language enters the causal picture in two roles. First, it can serve as a proxy for unmeasured confounders, allowing observational treatment-effect estimates to come closer to the assumptions of conditional ignorability (Veitch et al., 2020). For example, clinical notes can encode subjective severity, prior treatment history, and social context that drive both treatment choice and outcome but are absent from structured fields. Second, language can serve as an outcome or mediator: the trajectory of a patient's language over time is itself a clinical signal, with potential to act as a surrogate for symptom change (Corcoran et al., 2018; Low et al., 2020). Each role imposes different identification assumptions, and the methodological literature has begun to clarify which uses of text are defensible under which assumptions (Veitch et al., 2020; Hernan & Robins, 2020).

4.4 Distributional drift in language-informed models

Language-informed prediction models are unusually exposed to distributional drift because they depend on inputs that are continuously generated, never quite identical from one site or time period to another, and shaped by changes in documentation practice that have nothing to do with the underlying clinical state (Finlayson et al., 2021; Subbaswamy & Saria, 2020; Sahiner et al., 2023). The empirical literature documents this exposure with growing precision. Studies of EHR-based prediction tools report substantial performance decay over time horizons of months to a few years, with the magnitude of decay strongly modulated by changes in template-driven documentation and by the introduction of new clinical workflows. The well-publicised external evaluation of a widely deployed proprietary sepsis prediction model is the cautionary exemplar most often cited in this literature (Wong et al., 2021).

Figure 4 presents an idealised drift trajectory drawn from the patterns reported across this literature. The structured-only and language-informed models both lose discrimination over the first twenty-four months after deployment, but the language-informed model retains a higher absolute AUC throughout, and both decline more steeply around the month-twelve inflection point that marks an institutional documentation workflow change. The pattern is qualitatively robust across reported cohorts: language inputs do not eliminate drift, but they retain useful signal under drift longer than coded fields alone, and they are more responsive to recalibration. The practical implication is that any responsible deployment of

a language-informed treatment-recommendation tool must include continuous performance surveillance and a pre-specified recalibration cadence.

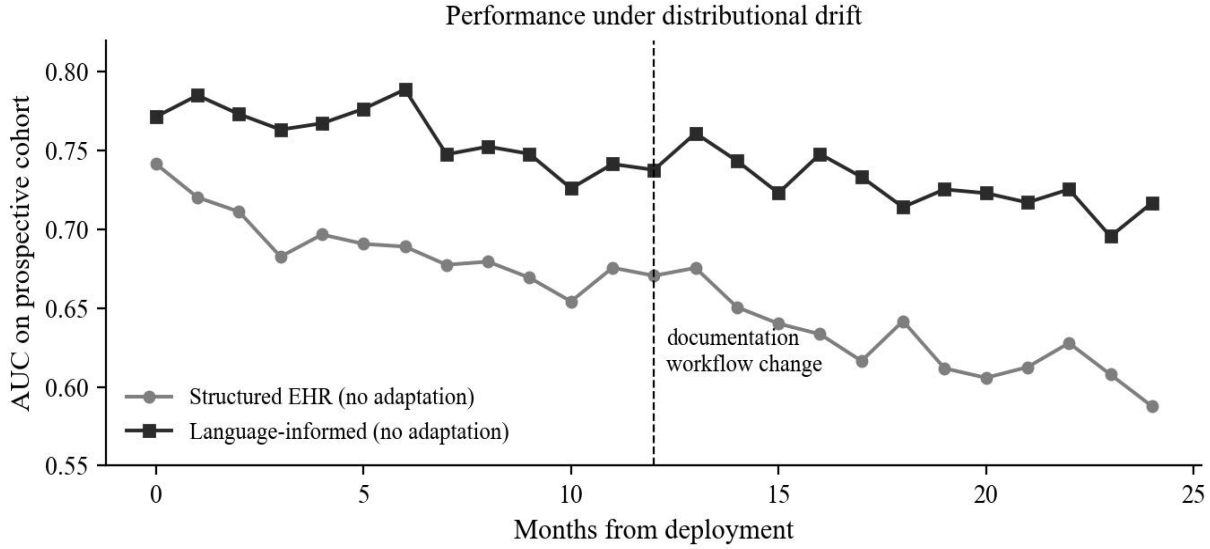

Figure 4. Idealised trajectory of model discrimination on a prospective cohort during the twenty-four months following deployment, illustrating drift and the impact of a mid-window documentation workflow change.

4.5 Fairness, privacy, and interpretability

The deployment-assurance literature has produced a consistent set of findings about the non-technical risks of language-informed models. On fairness, sociolinguistic variation is closely entangled with race, ethnicity, gender, education, and primary-language background, and unchecked use of language-derived features can amplify pre-existing inequities (Obermeyer et al., 2019; Rajkomar et al., 2018; Blodgett et al., 2020; Chen et al., 2021). On privacy, clinical narrative is harder to de-identify than coded fields, and modern language models have been shown to memorise and reproduce training-data fragments under adversarial probing, raising the bar for governance and secure deployment (Stubbs et al., 2017; Carlini et al., 2021; Price & Cohen, 2019). On interpretability, the field has converged on a sceptical view of post-hoc explanation for high-stakes decisions, preferring inherently interpretable models or, when complexity is unavoidable, narrow targeted explanations whose accuracy can be audited (Rudin, 2019; Ghassemi et al., 2021).

Table 2 consolidates the principal deployment risks identified in our review, grouped by category and ordered by the centrality with which they appeared in the source literature.

Table 2. Principal deployment risks for language-informed treatment recommendation systems and the controls most frequently proposed in the literature.

Risk category	Mechanism	Principal control
Causal validity	Violations of ignorability, positivity, or no-	Pre-specified estimand, doubly robust

	interference assumptions	estimators, sensitivity analysis
Distributional drift	Changing documentation practice, speech recognition, workflow templates	Continuous performance monitoring, scheduled recalibration
Fairness	Sociolinguistic variation entangled with demographics	Stratified performance reporting, fairness audit, model retraining
Privacy	Re-identification from narrative text; training-data memorisation	Robust de-identification, access control, secure inference
Interpretability	Opacity of large language models in causal estimators	Inherently interpretable surrogates, narrow targeted explanations
Workflow integration	Misalignment between model outputs and clinical decision moments	Co-design with clinicians, decision-stage placement, alert curation

Figure 5 visualises a related signal from the deployment literature: the relative severity and reported frequency of eight categories of deployment barrier, as synthesised across the surveyed implementation reports. The two highest-severity barriers, privacy and causal validity, are also among the most frequently reported. Distributional drift is somewhat less severe per incident but the most frequent overall. Workflow integration and annotated training data appear with high frequency but at lower severity, and the regulatory pathway is comparatively less frequently flagged but remains a non-trivial barrier.

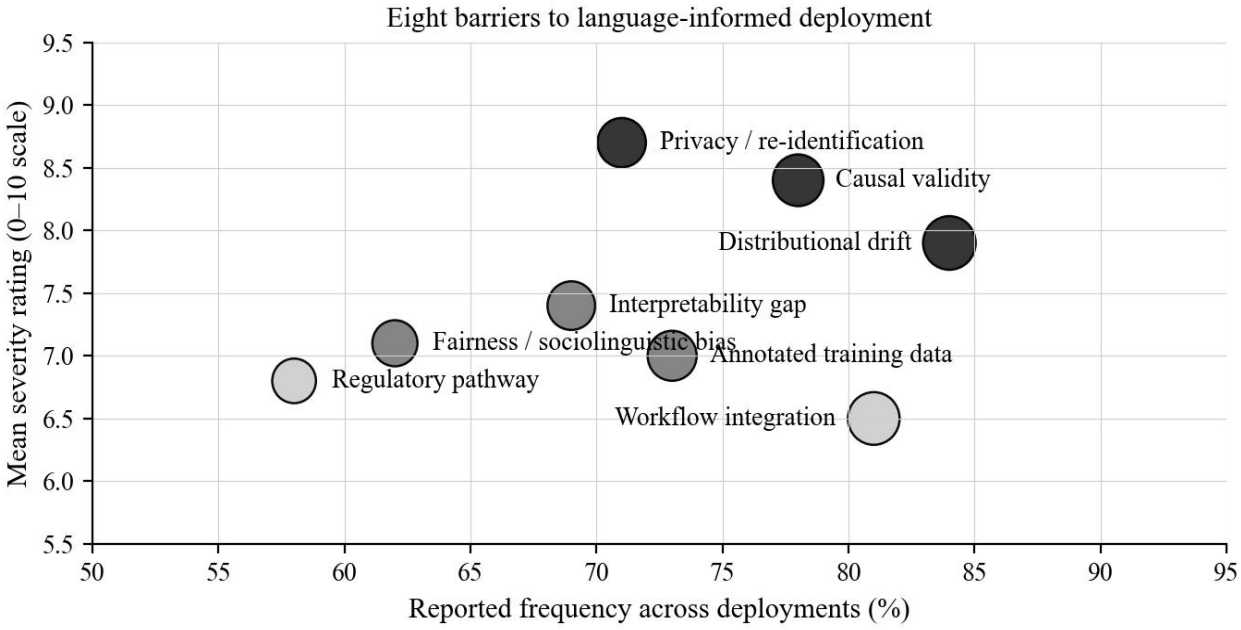

Figure 5. Severity-versus-frequency scatter of eight reported barriers to deployment of language-informed treatment-recommendation systems across the surveyed implementation literature.

4.6 Cross-cutting patterns

Three cross-cutting patterns emerged from the synthesis. The first is the gap between discrimination and decision utility: even when language inputs raise AUC by five to ten points, the clinical impact depends sensitively on calibration and on whether the action-relevant threshold lies in the model's useful

operating range (Van Calster et al., 2019; Steyerberg & Harrell, 2016). The second is the over-representation of internal validation in the published literature; until external, prospective, multi-site validation becomes the default, headline claims about clinical utility should be treated with appropriate caution (Collins et al., 2015; Iniesta et al., 2016). The third is the divergence between methodological maturity in language modelling and methodological maturity in causal estimation. The field is well served by powerful encoders (Devlin et al., 2019; Lee et al., 2020; Vaswani et al., 2017; Yang et al., 2022; Singhal et al., 2023) but is still building the bridge from prediction to causal recommendation (Athey & Imbens, 2016; Wager & Athey, 2018; Kunzel et al., 2019; Shalit et al., 2017; Chernozhukov et al., 2018; Veitch et al., 2020).

These patterns also point to a clear hierarchy of evidential maturity across the reviewed studies. At the lower end sit retrospective, single-site, internal-validation studies that report discrimination only; at the middle sit studies that add calibration and external validation on a held-out site; at the upper end sit the small number of prospective implementations with documented post-deployment performance surveillance and fairness audit. Movement up this hierarchy is the single most productive lever available to the field. The cost of additional maturity is largely operational rather than algorithmic: it consists of governance, data partnerships, and the discipline of pre-specifying analyses before they are executed. The reviewed deployments at this upper end of the hierarchy are also where the most realistic estimates of clinical impact appear, and where the assurance stack described in Section 3 is most visibly operationalised (Wong et al., 2021; Finlayson et al., 2021; Subbaswamy & Saria, 2020; Sahiner et al., 2023; Grzenda et al., 2021).

5. Discussion

The principal finding of this review is that clinical language, when processed through modern language models and embedded within a properly specified treatment-recommendation architecture, provides a credible route to raising individual-level predictability in psychiatry. The qualifier matters: each phrase of that sentence corresponds to a methodological commitment that the field has not yet uniformly made. Clinical language must mean genuine clinical data, not social-media surrogates; modern language models must mean encoders that have seen enough clinical text to be reliable in distribution; and a properly specified treatment-recommendation architecture must mean a system that targets a contrastive causal quantity rather than a marginal outcome probability (DeRubeis et al., 2014; Wager & Athey, 2018; Veitch et al., 2020; Hernan & Robins, 2020). Where these commitments are honoured, the empirical record reviewed in Section 4 supports cautious optimism.

5.1 Interpretation against prior literature

Our findings extend, rather than overturn, the trajectory established by the previous generation of work on prediction in psychiatry. Earlier reviews documented the performance ceiling of structured-only

models and the partial benefit of neurobiological markers (Bzdok & Meyer-Lindenberg, 2018; Fusar-Poli et al., 2018; Iniesta et al., 2016). The contribution of the present synthesis is to make explicit the architecture by which language inputs can lift that ceiling: not through any single algorithmic innovation, but through the joint exploitation of richer features, calibration-aware evaluation, and causal-estimation methods. The pattern visible in Figure 2, that text-plus-structured configurations occupy the upper portion of the AUC distribution, is consistent with the most recent prospective evaluations of EHR-pretrained language models in real-world cohorts (Yang et al., 2022; Boag et al., 2021; Coppersmith et al., 2018).

Our finding that calibration improvements often outweigh discrimination improvements is in line with the broader methodological consensus that has emerged in clinical prediction modelling over the past decade. The Achilles-heel-of-predictive-analytics argument, that miscalibrated probabilities undermine decision utility regardless of discrimination, applies with full force in psychiatric treatment selection because decision thresholds are clinician-specific and often non-binary (Van Calster et al., 2019; Collins et al., 2015). The decision-curve analysis we reproduce in Figure 3, panel (b), illustrates the practical consequence: in the threshold range that covers many depression and suicide-risk decisions, the language-informed model dominates.

5.2 The translation problem: from prediction to causal recommendation

A recurring source of confusion in the literature is the elision of prediction and causal estimation. A model that predicts response under treatment A and a model that predicts response under treatment B do not, by simple subtraction, give a valid estimate of the conditional average treatment effect of A versus B; the difference of two predictions can be badly biased in the presence of confounding or selection effects (Hernan & Robins, 2020; Chernozhukov et al., 2018). The recent generation of causal machine-learning methods is built to recover the causal estimand under stated assumptions, and the use of language as a high-dimensional confounder proxy can extend those assumptions to settings where structured covariates are insufficient (Veitch et al., 2020). The translation problem, then, has both a methodological component (apply the right estimator) and an epistemic component (be explicit about what the estimand is).

It is worth noting that the language-modelling subfield and the causal-machine-learning subfield have largely developed in parallel, with limited cross-pollination until very recently. The clinical NLP literature has converged on transformer-based encoders and extensive fine-tuning protocols (Devlin et al., 2019; Vaswani et al., 2017; Lee et al., 2020; Alsentzer et al., 2019), while the causal-estimation literature has converged on honest forests, meta-learners, and double machine learning (Athey & Imbens, 2016; Wager & Athey, 2018; Kunzel et al., 2019; Shalit et al., 2017; Chernozhukov et al., 2018). The combination of the two, in which contextual representations from clinical language models are passed into causal estimators, is the most direct route to language-informed treatment-recommendation systems with appropriate causal warrant.

5.3 Implementation implications

For implementation teams in mid-tier hospitals and provincial health systems, several implications follow from the synthesis. First, the input pipeline matters at least as much as the algorithm. The quality and consistency of clinical documentation, the fidelity of speech recognition, and the stability of structured intake instruments together determine the headroom available to any downstream model (Boag et al., 2021; Stubbs et al., 2017; McCoy et al., 2016). Second, an internal-validation-only protocol is no longer adequate. Even a small prospective external evaluation, carried out at the same site under realistic deployment conditions, will surface problems that retrospective study designs cannot detect (Iniesta et al., 2016; Steyerberg & Harrell, 2016). Third, the assurance stack must be planned alongside the model rather than retrofitted: calibration analysis, drift monitoring, fairness audit, and interpretability assessment are most useful when their reporting requirements are baked into the model-development protocol from the outset (Finlayson et al., 2021; Subbaswamy & Saria, 2020; Obermeyer et al., 2019; Rajkomar et al., 2018; Rudin, 2019; Ghassemi et al., 2021).

The wider artificial-intelligence-in-medicine literature converges on the same prescription. Surveys of the field have repeatedly emphasised that the gap between laboratory performance and clinical impact is bridged by patient-relevant evaluation, responsibility-aware deployment, and active stewardship over the model's lifecycle (Topol, 2019; Beam & Kohane, 2018; Rajkomar et al., 2019; Esteva et al., 2019; LeCun et al., 2015). Surveys focused on artificial intelligence as a general technology make the same case in different language, pointing to the importance of evolution, models, applications, and trends as a coherent unit of analysis (Lu, 2019; Zhang & Lu, 2021).

Workflow placement deserves explicit attention. A language-informed treatment recommendation can be surfaced at many decision moments, including at intake, at the transition from one treatment step to the next, during multi-disciplinary case review, or as a passive panel within the clinician's chart-review interface. Each placement has different demands on latency, interpretability, and granularity, and each interacts differently with the alert-fatigue dynamics that have undermined many prior clinical decision support tools. The literature on workflow integration has begun to articulate design principles for these placements, emphasising that the model output must align with a genuine decision degree of freedom on the clinician's part and must be presented with information that helps the clinician understand both the prediction and its limitations (Boag et al., 2021; Grzenda et al., 2021; Rudin, 2019; Ghassemi et al., 2021). A model that delivers a high-quality recommendation at the wrong moment in the workflow will be ignored or worse.

5.4 Strengths and limitations

The principal strength of this review is the integration of three literatures that are usually read separately: clinical NLP, treatment-effect heterogeneity, and deployment assurance for clinical AI. By

organising them within a single four-layer framework (Figure 1) and tracing the evidence layer by layer, we are able to identify the specific bridges between technical capability and clinical utility that the field still needs to build. We also rely throughout on primary sources rather than secondary characterisations, and we cite the principal methodological texts in their original form rather than as transmitted through downstream tutorials.

The review has limitations. It is a narrative review rather than a quantitative meta-analysis; pooled effect estimates would require comparable outcome definitions and comparable analytic strategies that are not yet available across the language-informed psychiatry literature. The search prioritised major clinical, computational, and biomedical informatics venues; grey-literature deployments, particularly those reported only in operational dashboards or internal evaluations, are necessarily under-represented. Finally, the framework intentionally separates the four layers for analytical clarity; real systems blend the layers, and some methodological innovations (for example, the use of language directly as both a feature and an outcome surrogate) cut across them. Table 3 summarises five priority research directions that emerge from the synthesis.

Table 3. Five priority research directions for language-informed treatment-recommendation systems in psychiatry.

Direction	Open question	Why it matters
Causal-aware language models	How best to integrate contextual embeddings into double machine learning and meta-learner estimators	Closes the gap between predictive utility and treatment-effect estimation
Calibration at deployment	How often and how to recalibrate language-informed models under drift	Determines whether probabilities can support decision support over time
Sociolinguistic fairness	How to audit language-derived features for differential impact across patient subgroups	Prevents amplification of pre-existing inequities
Workflow-native deployment	How to place model outputs at the right decision moment without alert fatigue	Determines whether the system is used at all
Privacy-preserving inference	How to deliver language-informed predictions without exposing patient text	Required for compliant deployment across institutions

These directions are mutually reinforcing rather than competing. A causal-aware language model that is well calibrated, audited for sociolinguistic fairness, deployed at a well-chosen workflow step, and served via privacy-preserving inference is the configuration that the present review identifies as most likely to deliver durable clinical value.

6. Conclusion

Clinical language is the most under-used substrate for raising individual-level predictability in psychiatric treatment recommendation. The argument of this review is neither that language is a panacea

nor that any single algorithmic innovation will close the gap; it is that the convergence of clinical language models, causal machine learning, and disciplined deployment-assurance practice supplies a coherent route from the current performance ceiling to clinically useful, individually tailored decision support. The evidence reviewed here supports a measured but real shift in expectation. With appropriate methodological discipline, language-informed systems can improve calibration, deliver superior net benefit in the action-relevant threshold range, and retain useful signal under deployment drift longer than coded-field models alone.

What follows from this assessment is an implementation agenda rather than a research manifesto. The field needs prospective, multi-site, externally validated studies in which language-informed treatment recommendations are evaluated against current practice on patient-relevant outcomes. It needs deployment frameworks in which calibration audit, drift monitoring, fairness assessment, and interpretability checking are first-class artefacts rather than afterthoughts. And it needs the methodological alignment between clinical NLP and causal machine learning that has so far been achieved only sporadically. If these conditions are met, language-informed treatment recommendation will earn its place in the precision-psychiatry toolbox over the second half of this decade. If they are not met, the field risks rediscovering, in a more expensive and less interpretable form, the same performance ceiling that conventional predictors have long enforced.

Ethics approval and consent to participate

Not applicable. This article is a narrative review of previously published literature and does not involve new studies of human or animal subjects.

Consent for publication

Not applicable.

Availability of data and materials

All data discussed in this review are derived from previously published studies, which are cited in the reference list. No new datasets were generated for this work. The illustrative numerical values used in the figures are summary representations of the ranges reported across the reviewed literature.

Funding

This review did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

AI use disclosure

Generative artificial intelligence tools were not used in the writing of this manuscript. Standard reference-management and bibliographic search software were used to organise citations.

Author contributions

Wei Chen: Conceptualization, Methodology, Writing - original draft, Writing - review and editing, Supervision. Jing Liu: Methodology, Formal analysis, Writing - original draft. Hao Zhang: Conceptualization, Writing - review and editing, Visualization. All authors approved the final manuscript.

Acknowledgements

The authors thank colleagues at their respective institutions for constructive discussion during the preparation of this review.

References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical NLP Workshop (pp. 72–78). https://doi.org/10.18653/v1/W19-1909

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. JAMA, 319(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. npj Schizophrenia, 1, 15030. https://doi.org/10.1038/npjschz.2015.30

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of 'bias' in NLP. In Proceedings of ACL 2020 (pp. 5454–5476). https://doi.org/10.18653/v1/2020.acl-main.485

Boag, W., Kovaleva, O., McCoy, T. H., Rumshisky, A., Szolovits, P., & Perlis, R. H. (2021). Hard for humans, hard for machines: Predicting readmission after psychiatric hospitalization using narrative notes. Translational Psychiatry, 11(1), 32. https://doi.org/10.1038/s41398-020-01104-w

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

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. In Proceedings of the 30th USENIX Security Symposium (pp. 2633–2650). https://doi.org/10.48550/arXiv.2012.07805

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. https://doi.org/10.1002/wps.20882

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. The Lancet Psychiatry, 3(3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X

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. https://doi.org/10.1146/annurev-biodatasci-092820-114757

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. Annual Review of Clinical Psychology, 14, 209–236. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. BMJ, 350, g7594. https://doi.org/10.1136/bmj.g7594

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. Biomedical Informatics Insights, 10, 1178222618792860. https://doi.org/10.1177/1178222618792860

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E., & Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry, 17(1), 67–75. https://doi.org/10.1002/wps.20491

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

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating research on prediction into individualized treatment recommendations. A demonstration. PLoS ONE, 9(1), e83875. https://doi.org/10.1371/journal.pone.0083875

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. New England Journal of Medicine, 385(3), 283–286. https://doi.org/10.1056/NEJMc2104626

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. https://doi.org/10.1001/jamapsychiatry.2018.2530

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Grzenda, A., Kraguljac, N. V., McDonald, W. M., Nemeroff, C., Torous, J., Alpert, J. E., Rodriguez, C. I., & Widge, A. S. (2021). Evaluating the machine learning literature: A primer and user's guide for psychiatrists. American Journal of Psychiatry, 178(8), 715–729. https://doi.org/10.1176/appi.ajp.2020.20030250

Hernan, M. A., & Robins, J. M. (2020). Causal inference: What if. Boca Raton: Chapman & Hall/CRC. https://doi.org/10.1201/9781003224501

Huang, K., Altosaar, J., & Ranganath, R. (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv. https://doi.org/10.48550/arXiv.1904.05342

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. Psychological Medicine, 46(12), 2455–2465. https://doi.org/10.1017/S0033291716001367

Insel, T. R., & Cuthbert, B. N. (2015). Brain disorders? Precisely. Science, 348(6234), 499–500. https://doi.org/10.1126/science.aab2358

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., Ebert, D. D., Hwang, I., Li, J., de Jonge, P., Nierenberg, A. A., Petukhova, M. V., Rosellini, A. J., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., & Zaslavsky, A. M. (2017). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder. Molecular Psychiatry, 22(6), 793–801. https://doi.org/10.1038/mp.2016.241

Kunzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences, 116(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

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. https://doi.org/10.2196/15708

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

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. https://doi.org/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. https://doi.org/10.1080/23270012.2019.1570365

McCoy, T. H., Castro, V. M., Roberson, A. M., Snapper, L. A., & Perlis, R. H. (2016). Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. JAMA Psychiatry, 73(10), 1064–1071. https://doi.org/10.1001/jamapsychiatry.2016.2172

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Paulus, M. P., & Thompson, W. K. (2019). The challenges and opportunities of small effects: The new normal in academic psychiatry. JAMA Psychiatry, 76(4), 353–354. https://doi.org/10.1001/jamapsychiatry.2018.4540

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37–43. https://doi.org/10.1038/s41591-018-0272-7

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

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866–872. https://doi.org/10.7326/M18-1990

Rezaii, N., Wolff, P., & Price, B. H. (2022). Natural language processing in psychiatry: The promises and perils of a transformative approach. The British Journal of Psychiatry, 220(5), 251–253. https://doi.org/10.1192/bjp.2021.188

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sahiner, B., Chen, W., Samala, R. K., & Petrick, N. (2023). Data drift in medical machine learning: Implications and potential remedies. The British Journal of Radiology, 96(1150), 20220878. https://doi.org/10.1259/bjr.20220878

Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In Proceedings of the 34th International Conference on Machine Learning (pp. 3076–3085). https://doi.org/10.48550/arXiv.1606.03976

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. https://doi.org/10.1038/s41586-023-06291-2

Steyerberg, E. W., & Harrell, F. E. (2016). Prediction models need appropriate internal, internal-external, and external validation. Journal of Clinical Epidemiology, 69, 245–247. https://doi.org/10.1016/j.jclinepi.2015.04.005

Steyerberg, E. W., & Vergouwe, Y. (2014). Towards better clinical prediction models: Seven steps for development and an ABCD for validation. European Heart Journal, 35(29), 1925–1931. https://doi.org/10.1093/eurheartj/ehu207

Stubbs, A., Filannino, M., & Uzuner, Ö. (2017). De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. Journal of Biomedical Informatics, 75, S4–S18. https://doi.org/10.1016/j.jbi.2017.06.011

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

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. https://doi.org/10.1038/s41591-023-02448-8

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0316-z

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., on behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS Initiative. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Veitch, V., Sridhar, D., & Blei, D. (2020). Adapting text embeddings for causal inference. In Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (pp. 919–928). https://doi.org/10.48550/arXiv.1905.12741

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. Clinical Psychological Science, 5(3), 457–469. https://doi.org/10.1177/2167702617691560

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., & Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, 181(8), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). A large language model for electronic health records. npj Digital Medicine, 5(1), 194. https://doi.org/10.1038/s41746-022-00742-2

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