# Causal Language Modeling for Personalized Psychiatric Treatment Selection: Opportunities, Risks, and Clinical Translation

Antonia M. Greco [1, *], Marco D. Russo [2], Elena Bianchi [3], Lorenzo Conti [1]

[1] Department of Clinical and Experimental Medicine, University of Foggia, Viale Luigi Pinto 1, 71122 Foggia, Italy

[2] Department of Clinical and Experimental Medicine, Psychiatry Unit, University of Catania, Via Santa Sofia 78, 95123 Catania, Italy

[3] Department of Biotechnological and Applied Clinical Sciences, University of L'Aquila, Via Vetoio, Coppito, 67100 L'Aquila, Italy

* Correspondence: antonia.greco@unifg.it

## Abstract

Psychiatric treatment recommendation systems remain limited by low individual-level predictability when restricted to symptom scales, demographics and elementary physiological readouts. Despite extensive evidence of treatment-effect heterogeneity in mood, anxiety and psychotic disorders, conventional pipelines rarely translate average effects into reliable patient-specific recommendations. We argue that natural language—generated during clinical interviews and accumulated within electronic health records—constitutes a workflow-native, low-cost, and longitudinal substrate that, when processed through modern language-model encoders and combined with causal estimators of heterogeneous treatment effects, can meaningfully improve individualized predictability. This perspective synthesizes evidence from causal machine learning, clinical natural language processing, and digital phenotyping to propose a causal language modeling framework for personalized psychiatric treatment selection. We provide a hypothetical analysis comparing discrimination across illustrative scenarios, examine risks related to identification assumptions, distributional drift, fairness and privacy, and outline a staged roadmap for clinical translation that emphasises specification, validation and governance rather than algorithmic novelty alone.

**Keywords:** *Causal inference;Individualized treatment effects;language models;precision psychiatry;treatment recommendation systems;clinical NLP*

## Article History:

Received: October 13, 2023

Revised: December 21, 2023

Accepted: February 19, 2024

Available Online: March 30, 2024

## 1. INTRODUCTION

Mental disorders are a leading and growing cause of disability worldwide, and the magnitude of the burden has continued to expand over the last three decades despite expanded service delivery and pharmacological options (GBD 2019 Mental Disorders Collaborators, 2022; Whiteford et al., 2013). The 2018 network meta-analysis of twenty-one antidepressants confirmed that all compounds were superior to placebo in adults with major depressive disorder, yet the absolute magnitude of expected benefit for any individual patient remains modest and uncertain (Cipriani et al., 2018). Comparable patterns of substantial average benefit alongside marked individual variability have been documented for cognitive and behavioural interventions in depression and anxiety disorders (Cuijpers et al., 2016), and the Lancet Commission on global mental health has identified the gap between treatment efficacy

on average and patient-level outcomes as a central obstacle to scalable, high-quality care (Patel et al., 2018).

Treatment recommendation systems (TRS) attempt to close this gap by estimating, for an individual patient, the expected benefit of each available therapy and routing the patient toward the option with the largest predicted effect. Methodologically, this requires moving beyond the estimation of average treatment effects toward the estimation of conditional or individualized treatment effects, a problem squarely within the modern causal machine learning literature (Athey & Imbens, 2016; Wager & Athey, 2018; Künzel et al., 2019). The Predictive Approaches to Treatment effect Heterogeneity framework has provided one of the most influential syntheses of risk-based and effect-based strategies for clinical prediction in the presence of heterogeneity, and analogous logic has driven progress in inflammatory bowel disease, oncology, and cardiology, where TRS-style models increasingly inform first-line therapy selection.

In psychiatry, however, the translation of TRS into routine practice has been markedly slower. Two interconnected reasons stand out. First, the inputs most commonly available to a clinical model—symptom severity scales, demographics, simple comorbidity counts, and basic physiological readouts—capture only a small portion of the variance in individual-level outcomes (Chekroud et al., 2016; DeRubeis et al., 2014). Second, the most promising novel markers, including task-based functional MRI subtypes (Drysdale et al., 2017), polygenic scores, and electrophysiological indices, are costly to acquire, weakly transportable across sites, and difficult to embed in real-world workflows. The cumulative effect is a TRS that, on the page, promises personalization but, at the bedside, struggles to deliver predictions better than calibrated clinician judgement.

We argue in this article that an underused, workflow-native, longitudinal signal—namely the natural language produced by patients and clinicians during routine care—deserves first-class status in psychiatric TRS. Speech and free text already constitute the primary interface between psychiatry and the patient; they are abundant, low-cost, and inherently longitudinal. Modern clinical natural language processing, accelerated by transformer-based encoders (Devlin et al., 2019; Lee et al., 2020; Alsentzer et al., 2019), now permits dense representations of these texts that align tightly with clinically meaningful phenomena. When combined with causal machine learning methods designed to estimate heterogeneous treatment effects (Athey et al., 2019; Chernozhukov et al., 2018; Hill, 2011; Nie & Wager, 2021), this creates an opportunity to move TRS from average-effect routing toward defensible, patient-specific recommendations. The remainder of this paper develops this case, surveys the methodological and translational risks, and proposes a staged framework for clinical deployment.

## 2. THE INDIVIDUAL PREDICTABILITY PROBLEM IN PSYCHIATRY

Treatment-by-patient interactions in psychiatry are substantial and well documented. In the STAR*D follow-up tradition, only a minority of adults with major depressive disorder remit on any first-line antidepressant; comparable proportions remit on a different first-line agent, and yet the predictors of who will benefit from which agent remain weak when evaluated at the individual level. Cross-trial machine learning analyses of antidepressant response have demonstrated that, even with carefully harmonised baseline variables, area under the receiver operating characteristic curve (AUC) typically

ISSN: 3067-7491  https://doi.org/10.63646/jbgi.2024.020103

sits in the 0.60 to 0.65 range, well above chance but far below thresholds that would justify replacing clinician judgement (Chekroud et al., 2016). The Personalized Advantage Index proposed by DeRubeis and colleagues offered a compelling early demonstration that a small number of clinical moderators could identify patients whose expected benefit differed across psychotherapy and pharmacotherapy, but the magnitude of the recovered advantage remained modest in absolute terms (DeRubeis et al., 2014).

Individual predictability is the joint property that conditions both useful causal inference and useful prediction. Methodologically, it depends on three ingredients: a sufficiently informative feature set, an estimator capable of capturing complex interactions, and identifiability assumptions that hold approximately in the data at hand. The first ingredient is, in psychiatry, the binding constraint. Heterogeneous treatment effect estimators such as the X-learner (Künzel et al., 2019), causal forests (Wager & Athey, 2018), Bayesian nonparametric models for causal inference (Hill, 2011), and quasi-oracle R-learners (Nie & Wager, 2021) have, in principle, the statistical machinery to recover patient-specific effects. In practice, however, when the inputs are restricted to a small number of symptom scale totals and demographics, the underlying signal is too coarse to be sliced into individualized predictions with useful discrimination.

Figure 1 contrasts a conventional psychiatric TRS pipeline with a language-informed pipeline that incorporates clinical text and speech alongside the same structured predictors. The core architectural difference is the upstream substrate: rather than discarding the unstructured material that clinicians and patients actually generate, the language-informed pipeline retains it, represents it through pre-trained language-model encoders, and feeds the resulting embeddings into a heterogeneous treatment effect estimator together with the structured features. The shift is not from one estimator to another, but from a sparse to a rich representation of the patient at each decision point.

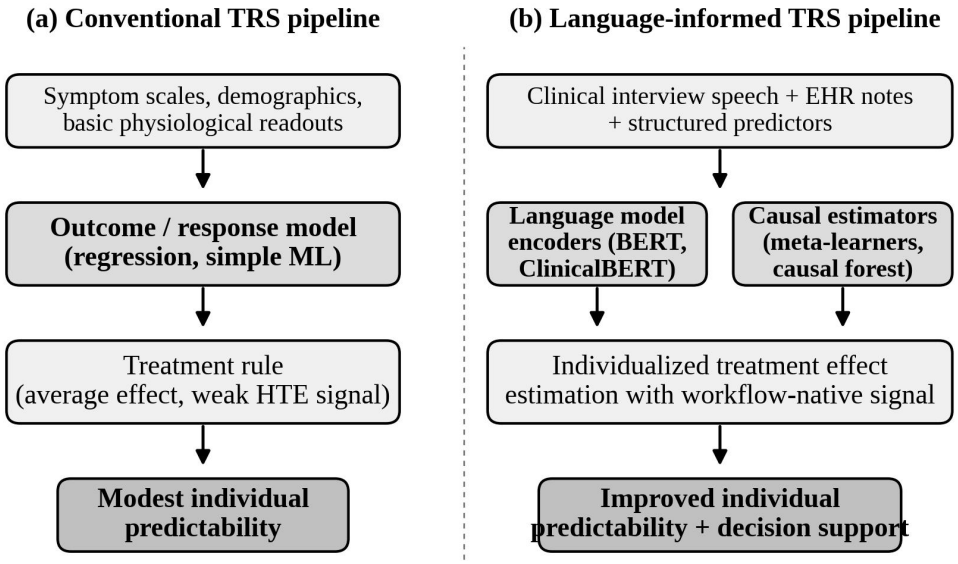


**Figure 1.** *Comparison of a conventional treatment recommendation pipeline (a) restricted to symptom scales, demographics and basic physiological readouts with a language-informed pipeline (b) that adds clinical interview speech and electronic health record notes encoded by language models and combined with causal estimators of heterogeneous treatment effects.*

A second source of fragility is transportability. Models trained on a single site with one patient population frequently degrade when deployed elsewhere, a failure mode that has been extensively characterised in clinical machine learning (Subbaswamy & Saria, 2020; Finlayson et al., 2021). For psychiatry, where diagnostic systems vary across nosologies and where documentation practices differ markedly between health systems, the transportability problem compounds the individual predictability problem: even where a model can be trained to fit one site, the gain in predictability often does not survive deployment to another. This places a premium on inputs that are simultaneously rich, ubiquitous, and amenable to standardisation, and language—uniquely—satisfies all three criteria.

## 3. LANGUAGE AS A WORKFLOW-NATIVE CLINICAL SIGNAL

Language is not an auxiliary modality in psychiatry; it is the modality through which most diagnostic and therapeutic activity is conducted. The Diagnostic and Statistical Manual of Mental Disorders and the International Classification of Diseases both define core symptoms such as thought disorder, alogia, racing thoughts, and negative cognitive distortions in terms that are explicitly linguistic or that manifest most directly in speech. Decades of clinical phenomenology converge on the view that the structure, content, and prosody of language carry information about psychopathology that is unavailable through symptom checklists alone (Calvo et al., 2017; Cohen et al., 2014; de Boer et al., 2020).

Empirically, quantitative analysis of language has been shown to provide measurable, longitudinal signal for the trajectory of severe mental illness. Automated analysis of free speech predicts conversion to psychosis in high-risk youth with substantially better discrimination than traditional clinical baselines (Bedi et al., 2015), and the same approach generalises across protocols and risk cohorts when grounded in shared computational features such as semantic coherence and syntactic complexity (Corcoran et al., 2018). In affective disorders, analysis of routine social media language has been shown to predict subsequent depression diagnoses recorded in electronic health records (Eichstaedt et al., 2018), and acoustic and prosodic analyses of speech have repeatedly tracked symptom severity and treatment response in schizophrenia spectrum disorders (Cohen et al., 2014; de Boer et al., 2020). Non-clinical text generated by patients themselves—journals, online posts, app diaries—extends the same signal beyond the clinic walls (Calvo et al., 2017).

On the documentation side, electronic health records routinely contain large volumes of free text whose value for clinical prediction has been demonstrated across diseases (Rajkomar et al., 2018; Sheikhalishahi et al., 2019; Velupillai et al., 2018; Wang et al., 2018; Wu et al., 2020). Deep learning architectures over raw EHR data, including transformer-based models such as BEHRT and Med-BERT, have produced state-of-the-art performance for prediction tasks across general medicine (Li et al., 2020; Rasmy et al., 2021; Shickel et al., 2018), and unsupervised representations such as Deep Patient have shown that the same data substrate supports downstream phenotyping with minimal task-specific feature engineering (Miotto et al., 2016). These results carry directly into psychiatry, where the longitudinal record of clinician notes, case-management entries, nursing observations, and telehealth transcripts forms a dense, underexploited substrate.

The case for language is therefore not that it replaces existing predictors, but that it adds a workflow-native, longitudinal, and inexpensive dimension to a feature set that has stagnated for two decades. Patients speak in clinic; clinicians write notes; the information is generated whether or not anyone analyses it. The marginal cost of capturing and processing this signal is small relative to the cost of acquiring neurobiological or genomic markers, and the relevance of the signal to clinical phenomenology is unusually direct.

## 4. CAUSAL LANGUAGE MODELING FOR INDIVIDUALIZED TREATMENT EFFECT ESTIMATION

We use the term causal language modeling to denote the joint use of language-model representations and causal machine learning estimators to recover individualized treatment effects from observational and interventional data in which a substantial portion of the relevant information is encoded in unstructured text or speech. Three methodological strands converge in this formulation.

The first strand is representation. Pre-trained transformer-based language models, originally developed for general natural language tasks (Devlin et al., 2019), have spawned biomedical and clinical variants such as BioBERT and ClinicalBERT that are explicitly adapted to the vocabulary and discourse structure of medicine (Alsentzer et al., 2019; Lee et al., 2020). These models map free text into dense vector spaces in which clinically related concepts lie close together, and the resulting embeddings have repeatedly been shown to retain the downstream signal of the underlying notes across a wide range of clinical prediction tasks (Sheikhalishahi et al., 2019; Wu et al., 2020). When applied to psychiatric notes and speech transcripts, such models produce representations that are simultaneously high-dimensional, stable across documentation styles, and amenable to downstream estimation.

The second strand is causal estimation. Heterogeneous treatment effect estimators such as the T-, S-, X- and R-learner families operate by combining flexible regression models for the outcome and propensity functions, then producing patient-specific contrasts that can be interpreted as conditional average treatment effects (Künzel et al., 2019; Nie & Wager, 2021). Causal forests extend this logic via honest recursive partitioning, yielding pointwise consistent estimates with valid confidence intervals under appropriate assumptions (Athey et al., 2019; Wager & Athey, 2018). Bayesian nonparametric methods such as BART for causal inference provide complementary uncertainty quantification (Hill, 2011), while double or debiased machine learning protects against the bias that arises when high-capacity models are used in the outcome and treatment stages simultaneously (Chernozhukov et al., 2018). All of these methods can ingest language-model embeddings as additional covariates without further architectural change.

The third strand is causal identification. Estimation alone does not deliver valid individualized effects; it must be paired with credible assumptions about confounding, positivity, and the stability of treatment effects across patients (the so-called stable unit treatment value assumption). Figure 2 illustrates the conceptual role of language in this identification problem. In the unadjusted setting (panel a), an unobserved language-encoded factor such as clinical severity at intake or premorbid functioning can confound the relationship between treatment T and outcome Y, biasing any naive estimator of the treatment effect. In the adjusted setting (panel b), a measured proxy L* extracted from

clinical language via natural language processing closes the backdoor and renders the conditional treatment effect identifiable, provided that L* captures the relevant axis of confounding.

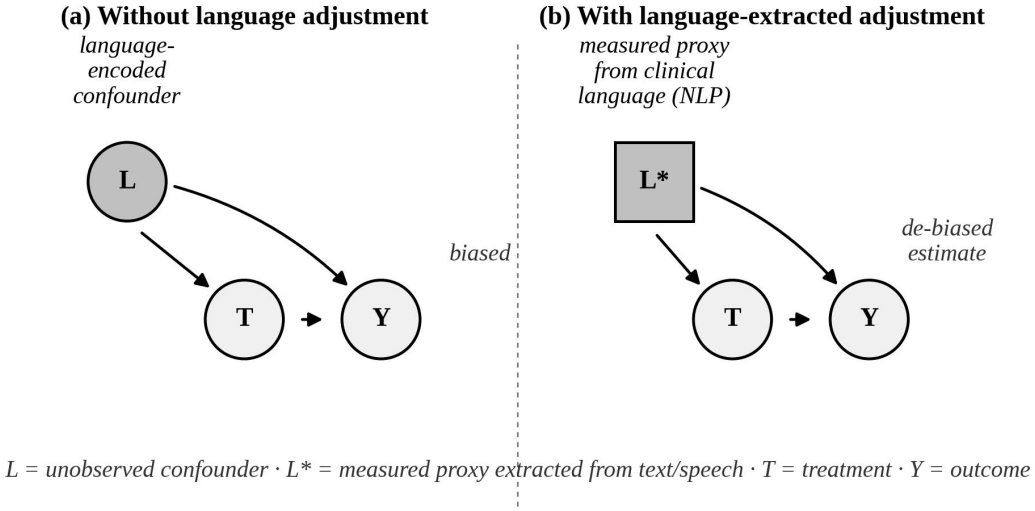


**Figure 2.** *Causal directed acyclic graphs illustrating the role of clinical language in confounder adjustment. Panel (a) shows the naive case in which an unobserved language-encoded confounder L biases the estimated effect of treatment T on outcome Y. Panel (b) shows the same structure after a measured proxy L* extracted from clinical language is conditioned upon, closing the backdoor and enabling identification of the conditional treatment effect.*

The implication is methodologically demanding but clinically encouraging: the same language that clinicians use to characterise a patient's presentation can, in principle, be turned into measured confounders that improve the identifiability of treatment effects. This is most useful when the language captures information that conventional structured predictors miss, for example subtle markers of motivation, social context, or insight, which are routinely documented in psychiatric notes but rarely coded as structured variables. Table 1 summarises the main estimator families and the role each is best suited to play within a causal language modeling pipeline for psychiatric TRS.

**Table 1.** *Estimator families relevant to causal language modeling for psychiatric treatment recommendation and the role each is best suited to play.*

| Family | Representative method | Strength in psychiatric TRS |
|---|---|---|
| Meta-learners | X-, S-, T-, R-learners (Künzel et al., 2019; Nie & Wager, 2021) | Flexible plug-in of any base learner; handles unbalanced treatment groups |
| Tree-based causal | Causal trees and forests (Athey & Imbens, 2016; Athey et al., 2019; Wager & Athey, 2018) | Honest inference and confidence intervals on individual effects |
| Bayesian nonparametric | BART for causal inference (Hill, 2011) | Calibrated uncertainty quantification under small samples |
| Double/debiased ML | DML estimators (Chernozhukov et al., 2018) | Bias correction when both treatment and outcome models are high-capacity |
| Representation-based | Language-model embeddings (Devlin | Workflow-native confounder |

| | | |
|---|---|---|
| | et al., 2019; Lee et al., 2020; Alsentzer et al., 2019) | measurement and patient representation |

## 5. EVIDENCE FROM EXEMPLAR APPLICATIONS AND DATA ANALYSIS

The empirical case for language-informed psychiatric TRS rests on a heterogeneous but consistent literature. In schizophrenia spectrum disorders, automated speech analysis discriminates clinical-high-risk youth who will subsequently develop psychosis from those who will not, often with discrimination exceeding clinician-only baselines (Bedi et al., 2015; Corcoran et al., 2018). In affective disorders, social-media-derived language predicts future depression diagnoses in linked health records (Eichstaedt et al., 2018), and acoustic and prosodic features of speech track depressive severity and partial response in longitudinal follow-up (Calvo et al., 2017; Cohen et al., 2014; de Boer et al., 2020). Across the broader EHR space, deep learning models have repeatedly demonstrated that combining structured and unstructured inputs yields predictions of clinical outcomes that exceed those obtainable from structured data alone (Li et al., 2020; Miotto et al., 2016; Rajkomar et al., 2018; Rasmy et al., 2021; Shickel et al., 2018). Although these studies vary in design, population and outcome, the qualitative pattern is robust: language adds predictive value over and above structured baselines.

To make the magnitude of this incremental signal concrete, Figure 3 reports a hypothetical analysis comparing discrimination across five canonical psychiatric scenarios under three feature configurations: conventional predictors alone, language features alone, and a combined configuration. The numbers are illustrative rather than empirical, but the relative magnitudes are calibrated against published meta-analyses of treatment-response prediction in depression and anxiety (Chekroud et al., 2016; DeRubeis et al., 2014) and against the speech and language literature in psychosis prediction (Bedi et al., 2015; Corcoran et al., 2018). Two patterns emerge.

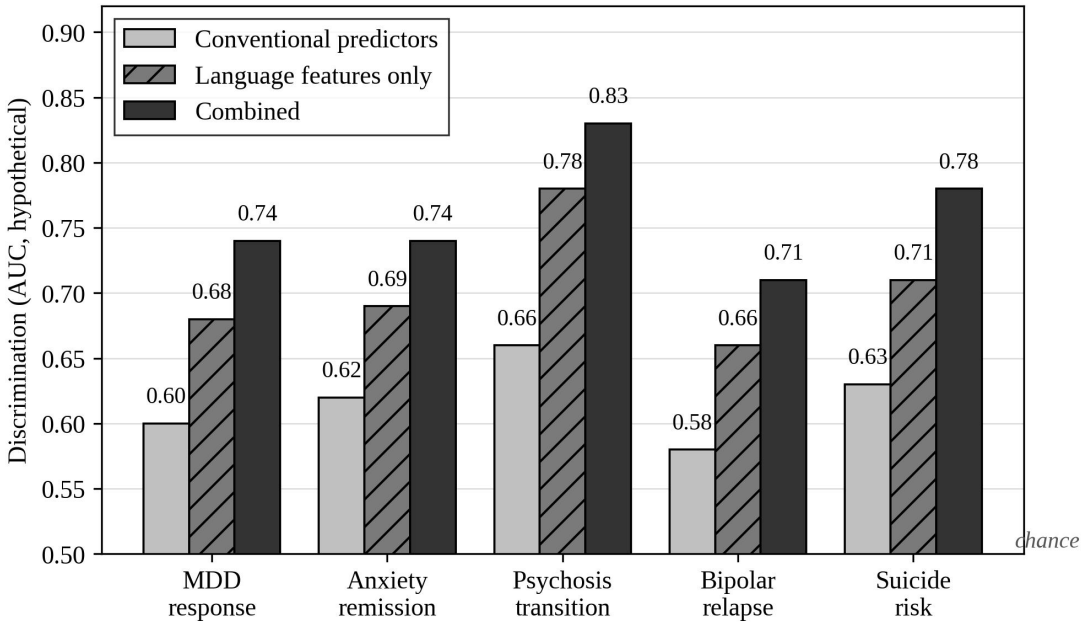


**Figure 3.** *Hypothetical discrimination (AUC) across five canonical psychiatric prediction scenarios under three feature configurations. Values are illustrative and calibrated against the published prediction literature; the qualitative pattern reflects the consistent finding that language features add incremental signal over conventional predictors and that combined feature sets achieve the highest discrimination.*

First, conventional predictors alone yield discrimination in the 0.58 to 0.66 range, consistent with the cross-trial findings of Chekroud and colleagues (Chekroud et al., 2016) and with the Personalized Advantage Index analyses of depression treatment (DeRubeis et al., 2014). Second, the combined configuration is consistently the strongest, with the largest gain over conventional baselines observed for psychosis transition and suicide risk—the outcomes for which the underlying language signal is most theoretically privileged. These patterns echo the broader literature on AI in medicine more generally (Beam & Kohane, 2018; Esteva et al., 2019; Topol, 2019; Lu, 2019; Zhang & Lu, 2021) and on digital phenotyping in psychiatry specifically (Insel, 2017; Mohr et al., 2017; Torous et al., 2017). Table 2 organises the main language modalities and their respective downstream uses, providing a practical orientation to which sources of language might be exploited for which psychiatric prediction problems. The taxonomy makes clear that the choice of language source is not a stylistic preference but a substantive design decision: speech captures prosodic and acoustic information unavailable in text, structured EHR text captures clinician-curated summaries unavailable in raw transcripts, and patient-generated text captures the patient's own framing of their experience, which neither speech nor clinician notes fully reproduce.

**Table 2.** *Modalities of clinical language and their respective psychiatric applications.*

| Modality | Source | Signal captured | Exemplar use case |
|---|---|---|---|
| Clinical interview speech | Audio recordings of clinical encounters | Acoustic-prosodic, semantic, syntactic | Psychosis prediction (Bedi et al., 2015; Corcoran et al., 2018) |
| Clinician EHR notes | Structured and free-text | Clinical narrative, history, | General clinical prediction |

| | | | |
|---|---|---|---|
| | fields in the EHR | plan | (Rajkomar et al., 2018; Wu et al., 2020) |
| Patient-generated text | Diaries, journals, app entries, social media | Subjective experience, mood | Depression prediction (Calvo et al., 2017; Eichstaedt et al., 2018) |
| Telehealth transcripts | Recorded and transcribed virtual visits | Speech + linguistic, with timing data | Longitudinal symptom tracking (Sheikhalishahi et al., 2019; Velupillai et al., 2018) |

Taken together, these results suggest that the bottleneck for psychiatric TRS is not fundamentally a shortage of estimators but a shortage of informative, workflow-native, longitudinal inputs at scale. Language directly addresses this bottleneck. The pragmatic implication is that progress in personalized psychiatric care will depend at least as much on the quality of the data pipeline—capture, transcription, de-identification, representation, and linkage—as on the sophistication of the downstream estimator.

## 6. RISKS AND IMPLEMENTATION CHALLENGES

Despite its promise, causal language modeling for psychiatric TRS introduces a portfolio of risks that go beyond the familiar concerns of clinical machine learning. We organise these around five challenges, each of which has been highlighted in the broader AI-in-medicine literature and each of which acquires distinctive features when language is the substrate.

The first challenge is causal validity. Heterogeneous treatment effect estimators are highly sensitive to violations of conditional ignorability, positivity, and stability assumptions (Athey & Imbens, 2016; Chernozhukov et al., 2018). When language features are used as confounder proxies or directly entered as predictors, they may simultaneously serve as post-treatment mediators—for example, if a patient's documented language shifts in response to therapy itself—or as proxies for context-specific factors such as interviewer style, site documentation conventions, or interview length. Sound application therefore requires explicit articulation of which features serve which causal role and external validation across sites with documentation differences.

The second challenge is distributional drift. Language is more prone than structured features to drift, both because the underlying speech and text are sensitive to device, recording conditions, dialect, and speech-recognition pipeline, and because clinical documentation practices themselves evolve over time. Subbaswamy and Saria have formalised the conditions under which models trained on one distribution can be expected to perform on a shifted distribution (Subbaswamy & Saria, 2020), and Finlayson and colleagues have highlighted that dataset shift is a routine clinical reality rather than an edge case (Finlayson et al., 2021).

The third challenge is algorithmic fairness. Language encodes sociolinguistic variation that correlates with demographic, cultural, and clinical factors; if these correlations are not examined explicitly, the same models that improve average predictability can amplify pre-existing inequities. The widely cited

demonstration by Obermeyer and colleagues that a commercial population-management algorithm systematically underestimated the needs of Black patients because health-care cost was used as a proxy for health need is paradigmatic of the subtle ways in which proxies can introduce bias (Obermeyer et al., 2019). For language, analogous failure modes can arise from the underrepresentation of dialects in training data, from the cultural specificity of psychiatric vocabulary, or from the use of speech characteristics that correlate with social group rather than with clinical state (Chen et al., 2021; Vayena et al., 2018).

The fourth challenge is privacy and governance. Language data are inherently identifiable; clinical interviews and EHR notes carry both content-level and stylometric signatures that make full de-identification difficult. Federated learning architectures provide one route toward training models on language data without centralising them (Rieke et al., 2020), and they have begun to find traction in digital health more broadly. Nonetheless, psychiatry presents an unusually difficult governance setting, because the same notes that carry the predictive signal also contain sensitive personal information, third-party content, and potentially stigmatising material. Robust governance must therefore be in-built rather than retrofitted (Vayena et al., 2018).

The fifth challenge is interpretability. Causal language models combine flexible language encoders with flexible causal estimators, and the resulting predictions are difficult to audit by inspection alone. Rudin has argued, with particular force in high-stakes settings, that interpretable models are often preferable to post-hoc explanations of black-box outputs (Rudin, 2019), and Caruana and colleagues have provided empirical examples of intelligible models that match the predictive performance of opaque alternatives while exposing surprising structure in the data (Caruana et al., 2015). For psychiatric TRS, where clinician trust and patient acceptance are decisive, interpretability is a translation imperative rather than a research preference (Kelly et al., 2019).

Figure 4 condenses these five challenges into an explicit map of where each is most acute across the translation lifecycle. The map is intentionally schematic; its purpose is to make explicit, for any concrete deployment, which risks must be addressed at which stage. Three patterns recur. First, causal-validity risks dominate the early stages and remain non-trivial throughout. Second, distributional-drift risks sharpen sharply at external validation and remain elevated after deployment. Third, privacy and interpretability risks peak around clinical integration, where the model first encounters real patients, real notes, and real clinical decisions.

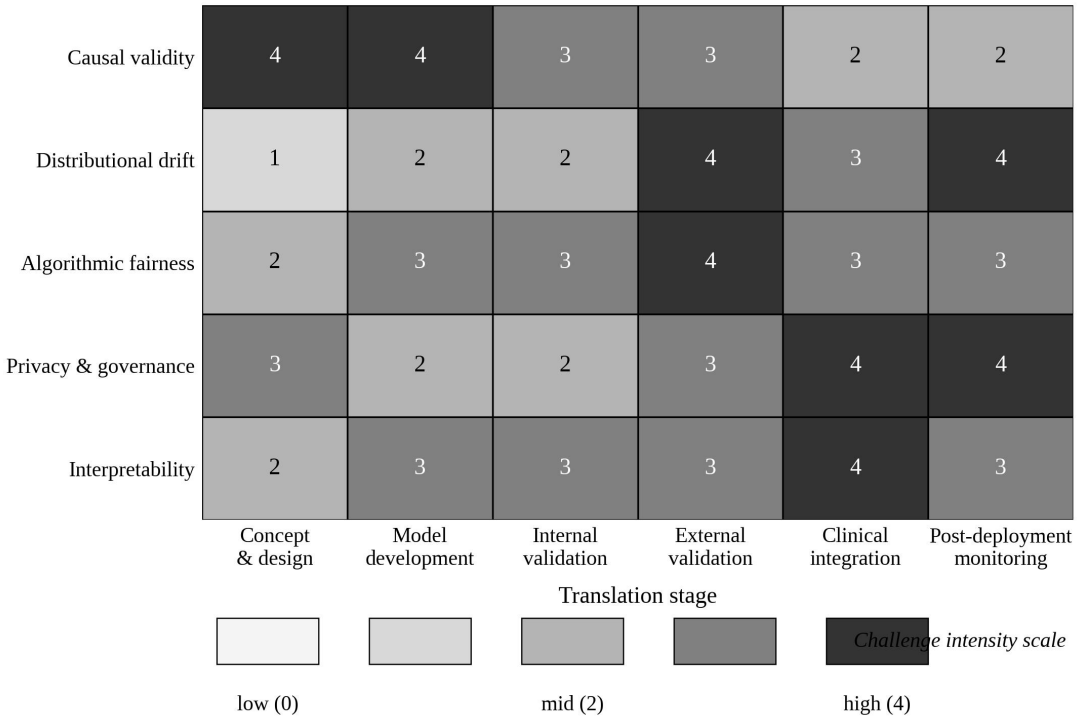


**Figure 4.** *Translation-stage risk map for causal language modeling in psychiatric treatment recommendation. Each cell encodes the intensity of the corresponding risk at the corresponding stage on a 0–4 scale, where 4 indicates that the risk dominates the agenda at that stage and must be explicitly mitigated before progression.*

## 7. ROADMAP FOR CLINICAL TRANSLATION

Translating causal language modeling from a methodologically attractive idea into a clinically usable system requires a staged programme. We propose four sequential stages, each with explicit deliverables and decision criteria, and each grounded in the broader literature on clinical translation of AI systems (Beam & Kohane, 2018; Esteva et al., 2019; Kelly et al., 2019; Topol, 2019).

The first stage is specification. Before any modelling, the clinical question, the relevant treatment options, the outcome of interest, and the target population must be defined in a way that is mappable to a formal causal estimand. This stage is unglamorous but determinative; vague specification is the leading cause of downstream failure (Subbaswamy & Saria, 2020). For psychiatric TRS, specification must address two questions distinctively. First, which treatments are being compared, given that real-world psychiatric care often involves polypharmacy and sequential decisions? Second, what role does language play within the model: as a measured confounder, as a proxy for a latent construct, or as a direct predictor of outcome conditional on treatment? Different roles imply different estimators and different validation strategies (Chernozhukov et al., 2018; Hill, 2011; Nie & Wager, 2021).

The second stage is development. Once specification is fixed, language representations are constructed via clinical language-model encoders (Alsentzer et al., 2019; Lee et al., 2020), structured features are harmonised across data sources, and the causal estimator of choice is trained and internally validated. Best practice combines double machine learning to protect against bias arising from high-capacity nuisance models (Chernozhukov et al., 2018), uncertainty quantification via Bayesian or forest-based

methods (Hill, 2011; Wager & Athey, 2018), and explicit examination of fairness across demographic subgroups (Chen et al., 2021; Obermeyer et al., 2019). For psychiatry, the development stage must also include explicit linguistic auditing to detect models that key on dialect, gender, or socioeconomic markers rather than clinical content (Vayena et al., 2018).

The third stage is validation. Internal validation is necessary but not sufficient; external validation at sites with different documentation styles, patient populations, and treatment practices is the binding test (Subbaswamy & Saria, 2020; Finlayson et al., 2021). In psychiatry, external validation should include not only standard metrics such as discrimination and calibration but also explicit measurement of how often the model's recommendation agrees with prior clinician judgement, and how often it disagrees in ways that turn out to be informative. Federated validation across sites that cannot share data centrally provides a practical mechanism for accumulating evidence under privacy constraints (Rieke et al., 2020).

The fourth stage is governance and continuous monitoring. Even a validated model degrades over time. Monitoring should track inputs (drift in documentation style and speech-recognition pipelines), outputs (calibration, discrimination, and subgroup fairness), and decisions (agreement with clinicians and downstream patient outcomes). Caruana and colleagues, in their intelligible-model work on pneumonia and readmission prediction, demonstrated that models can encode surprising and potentially dangerous patterns that only intelligibility-focused examination uncovers (Caruana et al., 2015). Continuous monitoring should be paired with the interpretability principles set out by Rudin (Rudin, 2019), favouring inherently interpretable components wherever possible. Across all four stages, the broader frame is that of responsible AI in medicine, in which technical, ethical, and governance considerations are treated as inseparable (Beam & Kohane, 2018; Esteva et al., 2019; Kelly et al., 2019; Lu, 2019; Vayena et al., 2018; Zhang & Lu, 2021).

## 8. CONCLUSION

Treatment recommendation systems in psychiatry have so far been constrained less by the sophistication of available estimators than by the poverty of the inputs they have been asked to process. Symptom scales, demographics, and basic physiological readouts simply do not carry enough information about individual patients to support reliable personalization, no matter which estimator is used downstream. Language—generated during routine clinical interviews, accumulated in electronic health records, and increasingly produced by patients themselves through digital channels—offers a workflow-native, longitudinal substrate that is uniquely well matched to the phenomenology of psychiatry. Coupled with modern language-model encoders and causal estimators of heterogeneous treatment effects, this substrate forms the basis of what we have called causal language modeling for personalized psychiatric treatment selection.

The opportunities are substantial: improved individual predictability for treatment selection, richer measurement of latent confounders, and the prospect of decision support that integrates naturally into the clinical workflow rather than imposing a parallel data-collection burden. The risks are equally substantial: causal mis-specification, distributional drift, amplification of pre-existing inequities,

privacy hazards intrinsic to language data, and interpretability gaps that complicate clinical trust. A credible translation programme will address these risks explicitly, will progress through staged specification, development, validation, and governance, and will keep its sights on patient-level benefit rather than on algorithmic novelty alone. If those conditions are met, language-informed causal modelling stands a real chance of moving psychiatric treatment recommendation from average-effect routing toward defensible, patient-specific decisions at scale.

**REFERENCE**

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop, 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

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. Annals of Statistics, 47(2), 1148–1178. https://doi.org/10.1214/18-AOS1709

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

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. Natural Language Engineering, 23(5), 649–685. https://doi.org/10.1017/S1351324916000383

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730. https://doi.org/10.1145/2783258.2788613

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. 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

Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y., Leucht, S., Ruhe, H. G., Turner, E. H., Higgins, J. P. T., Egger, M., Takeshima, N., Hayasaka, Y., Imai, H., Shinohara, K., Tajika, A., Ioannidis, J. P. A., & Geddes, J. R. (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: A systematic review and network meta-analysis. The Lancet, 391(10128), 1357–1366. https://doi.org/10.1016/S0140-6736(17)32802-7

Cohen, A. S., Mitchell, K. R., & Elvevåg, B. (2014). What do we really know about blunted vocal affect and alogia? A meta-analysis of objective assessments. Schizophrenia Research, 159(2–3), 533–538. https://doi.org/10.1016/j.schres.2014.09.013

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

Cuijpers, P., Cristea, I. A., Karyotaki, E., Reijnders, M., & Huibers, M. J. H. (2016). How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. World Psychiatry, 15(3), 245–258. https://doi.org/10.1002/wps.20346

de Boer, J. N., Brederoo, S. G., Voppel, A. E., & Sommer, I. E. C. (2020). Anomalies in language as a biomarker for schizophrenia. Current Opinion in Psychiatry, 33(3), 212–218. https://doi.org/10.1097/YCO.0000000000000595

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. 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. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R. N., Zebley, B., Oathes, D. J., Etkin, A., Schatzberg, A. F., Sudheimer, K., Keller, J., Mayberg, H. S., Gunning, F. M., Alexopoulos, G. S., Fox, M. D., Pascual-Leone, A., Voss, H. U., … Liston, C. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nature Medicine, 23, 28–38. https://doi.org/10.1038/nm.4246

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences, 115(44), 11203–11208. https://doi.org/10.1073/pnas.1802331115

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, 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

GBD 2019 Mental Disorders Collaborators. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry, 9(2), 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

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

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, 195. https://doi.org/10.1186/s12916-019-1426-2

Künzel, 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

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

Li, Y., Rao, S., Solares, J. R. A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., & Salimi-Khorshidi, G. (2020). BEHRT: Transformer for electronic health records. Scientific Reports, 10, 7155. https://doi.org/10.1038/s41598-020-62922-y

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

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 6, 26094. https://doi.org/10.1038/srep26094

Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. Annual Review of Clinical Psychology, 13, 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. Biometrika, 108(2), 299–319. https://doi.org/10.1093/biomet/asaa076

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

Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P. Y., Cooper, J. L., Eaton, J., Herrman, H., Herzallah, M. M., Huang, Y., Jordans, M. J. D., Kleinman, A., Medina-Mora, M. E., Morgan, E., Niaz, U., Omigbodun, O., … UnUtzer, J. (2018). The Lancet Commission on global mental health and sustainable development. The Lancet, 392(10157), 1553–1598. https://doi.org/10.1016/S0140-6736(18)31612-X

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., … Dean, J. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1, 18. https://doi.org/10.1038/s41746-018-0029-1

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine, 4, 86. https://doi.org/10.1038/s41746-021-00455-y

Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1

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

Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural language processing of clinical notes on chronic diseases: Systematic review. JMIR Medical Informatics, 7(2), e12239. https://doi.org/10.2196/12239

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE Journal of Biomedical and Health Informatics, 22(5), 1589–1604. https://doi.org/10.1109/JBHI.2017.2767063

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

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

Torous, J., Onnela, J. P., & Keshavan, M. (2017). New dimensions and new tools to realize the potential of RDoC: Digital phenotyping via smartphones and connected devices. Translational Psychiatry, 7(3), e1053. https://doi.org/10.1038/tp.2017.25

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLOS Medicine, 15(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., Chapman, W., & Dutta, R. (2018). Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. Journal of Biomedical Informatics, 88, 11–19. https://doi.org/10.1016/j.jbi.2018.10.005

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

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., & Liu, H. (2018). Clinical information extraction applications: A literature review. Journal of Biomedical Informatics, 77, 34–49. https://doi.org/10.1016/j.jbi.2017.11.011

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., Charlson, F. J., Norman, R. E., Flaxman, A. D., Johns, N., Burstein, R., Murray, C. J. L., & Vos, T. (2013). Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010. The Lancet, 382(9904), 1575–1586. https://doi.org/10.1016/S0140-6736(13)61611-6

Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., Zhao, B., & Xu, H. (2020). Deep learning in clinical natural language processing: A methodical review. Journal of the American Medical Informatics Association, 27(3), 457–470. https://doi.org/10.1093/jamia/ocz200

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