Task-Aligned Outcome Learning in Psychiatry: Reducing Endpoint Dilution

Eric V. Strobl

Departments of Biomedical Informatics and Psychiatry
University of Pittsburgh

Psychiatric research relies on well-defined outcomes for standardization, comparability, and replication, yet investigators often fix broad endpoints before knowing which symptom domains carry task-relevant signal. Even when psychometrically sound and clinically useful, composite measures can dilute predictive information and attenuate treatment effects when predictability or responsiveness concentrates in only a subset of symptoms—thus making studies appear negative despite meaningful change. This Perspective proposes a task-aligned, two-stage framework to learn the appropriate outcome with machine learning: constrained discovery learns a clinically interpretable outcome from a prespecified item pool; confirmatory evaluation tests replication across settings using either the fixed learned outcome when designs match closely or a relearned outcome produced by the same prespecified procedure when designs differ. The framework complements psychometrics and open-science practices, shifting focus from broad unsupervised composites to empirically supported targets, with safeguards to keep results interpretable and rigorous.

Introduction

Psychiatric trials and prediction studies often fall short despite better measurement, larger datasets, and more sophisticated models. Researchers usually blame placebo response [1], clinical heterogeneity [2], limited power [3], or weak interventions [4]. These factors matter, but they can obscure a simpler source of signal loss: investigators often commit to broad endpoints before they know which symptom domains are most predictable or most likely to change with treatment. Composite totals and diagnostic labels can be clinically useful and psychometrically defensible, yet they often pool symptom dimensions that differ in predictability, treatment responsiveness, and mechanistic relevance [5–7].

For example, a bupropion trial might prespecify the Quick Inventory of Depressive Symptomatology–Self Report (QIDS-SR) total score as its primary endpoint, even though the score includes opposite poles of key symptoms (appetite/weight gain vs loss; hypersomnia vs insomnia) [8]. If bupropion improves hypersomnia and increased appetite but worsens insomnia or appetite loss, the total score can mask offsetting symptom shifts and become less sensitive to bupropion's therapeutic profile [9]. This is not an argument to ignore adverse effects; the problem is that a single total score can mix therapeutic and adverse shifts into one number, obscuring both benefit and tolerability. More generally, broad totals can average targeted improvement with unchanged or counter-directional symptoms, diluting treatment-relevant signal before analysis begins.

This Perspective therefore critically examines the use of broad scores and its consequences across study types. Psychiatry often adopts the same broad outcomes as default targets for prediction studies, causal analyses, and clinical trials [10, 11]. Standardization supports comparability,

but poorly aligned endpoints can mask true effects and make "placebo responsiveness" a default explanation for null results [12]. I therefore propose a task-aligned machine learning framework called *outcome learning* to address the problem of endpoint dilution. In a constrained discovery stage, studies learn clinically interpretable outcomes from a prespecified item pool. In subsequent confirmatory work, researchers test replication across settings using either the fixed learned outcome when the study context closely matches the derivation setting, or a relearned outcome obtained by applying the same prespecified learning procedure and constraints when contexts, comparators, or treatments differ. This framework thus retains the aims of standardization and preregistration, but it shifts what we prespecify—from broad, largely unsupervised composites to a constrained learning procedure with empirically supported, task-relevant targets. The sections below explain why this shift may help and how it builds on existing methods, as summarized in Table 1. They also describe appropriate safeguards and illustrate outcome learning with a concrete example.

Limitations of fixed outcomes	**Prevailing response**	**Outcome learning as a solution**
Broad scoring rules fixed too early	Refine constructs or measurement	Prespecified constrained outcome discovery
One endpoint used across tasks	Increase model/predictor complexity	Task-specific targets
Symptom heterogeneity treated as noise and uncertainty	Restrict samples or subtype patients	Leverage heterogeneity to learn interpretable symptom targets

Table 1 From fixed outcomes to task-aligned outcome learning. Each row links a limitation of fixed outcome definitions to common methodological responses, and to a task-aligned outcome-learning alternative.

Limitations of Fixed Outcome Definitions in Psychiatric Research

Psychiatry relies on fixed outcomes for good reasons. Standardized scales and diagnostic definitions enable clear communication, make results easier to interpret, and support replication and meta-analysis [11, 13]. The problem is not standardization per se, but what we standardize and when. Psychiatry often standardizes both the instrument *and* its scoring rule—typically a single total score or diagnostic label. Keeping the instrument fixed is often essential, and the instrument may actually contain clear signals of change. The risk comes from fixing a broad scoring rule too early: when symptom domains differ in predictability, treatment responsiveness, or confounding, the total score can average away the very signal the instrument captures. As a result, prediction models can look weak even when they predict specific symptom domains well [5], and trials can miss meaningful benefits when an intervention changes only a subset of symptoms that the instrument still measures [6].

This problem worsens when investigators carry the same broad endpoint unchanged across tasks with different inferential goals. Prediction studies prioritize stable, learnable targets; mechanistic studies prioritize targets that track the hypothesized pathway; trials prioritize endpoints that sensitively capture treatment-specific change. A single omnibus outcome rarely optimizes all three. When investigators treat it as a universal target, they can lose power, understate clinically meaningful effects, and draw overly pessimistic conclusions about what the data can support.

These limitations do not imply that fixed outcomes are always misguided, nor do they justify endpoint shopping. The key design choice is *timing*. A more productive workflow holds the instrument fixed but uses prespecified, constrained discovery to learn a task-relevant, clinically interpretable scoring rule. Subsequent confirmatory studies then test replication across cohorts and settings—using the fixed learned outcome when designs closely match the derivation setting,

or relearning the outcome by applying the same prespecified procedure and constraints when they do not (as described later). In this way, the focus shifts from "one broad score for every task" to a prespecified outcome-learning workflow and task-aligned targets that can be validated across studies.

Prevailing Methodological Responses and Their Limitations

Psychiatry has developed several sophisticated responses to the above limitations, but many focus on other levers—constructs, measurement, models, or sampling—rather than the endpoint itself even when endpoint dilution is the main source of signal loss.

One common response focuses on construct. Researchers revise diagnostic categories [14], redefine symptom boundaries [15] or develop alternative nosologic frameworks [16], such as RDoC [17]. Related efforts improve measurement quality through new scales [18], refined item sets [8], repeated assessments [19], different informants [20] and psychometric methods such as calibration and harmonization [21]. These efforts are scientifically valuable. They can sharpen constructs, improve reliability and clarify what instruments measure. However, they often define outcomes primarily on psychometric or conceptual grounds, without explicit reference to the specific analytic task. As a result, the same broadly defined outcome is often carried across prediction, causal and treatment studies, even when its composition is poorly aligned with the signal or effect each task is trying to detect.

A second response increases predictor or model complexity while keeping the outcome fixed. Studies add more features, use more complex machine learning models or seek much larger datasets [22]. These strategies sometimes improve performance, but they also introduce trade-offs: lower interpretability [23], higher deployment burden [24] and weaker transportability across hospitals or health systems [25]. More complex models can also fail in external validation because they are fragile, as added complexity often increases fragility across settings [26]. When the main limitation lies in the outcome definition, greater model complexity may offer only modest gains.

A third response reduces heterogeneity by restricting the sample. Researchers may recruit narrowly defined subgroups or putative subtypes to create a more internally consistent cohort or enrich for treatment response [27–29]. This is especially consequential in psychiatry, where randomized trials already face persistent concerns about external validity and representativeness [30, 31]; further restriction can compound those generalizability limits. In some cases, the need to restrict or cluster patients arises partly because the outcome itself is too broad to capture the relevant signal in a more heterogeneous population. The outcome may therefore remain poorly aligned with the task, while the resulting findings generalize poorly to routine clinical populations. In treatment studies, strict restriction can also exclude patients who might benefit if the study used a more task-aligned outcome.

Taken together, the above strategies improve constructs, measurements, predictors or samples while leaving the target outcome largely unchanged. They can therefore produce only partial gains when the central problem is mismatch between the outcome and the task.

Outcome Learning as a Solution

Outcome learning takes a different approach by leveraging psychiatric symptom heterogeneity. Although heterogeneity is often framed as a measurement problem, it also creates a design opportunity: broad symptom inventories provide a structured, clinically meaningful item pool from which researchers can construct outcomes that better match a specific question. This flexibility matters because investigators often cannot know in advance which symptom domain will be most predictable from baseline data or most responsive to a treatment. Prior knowledge may support a broad hypothesis, but it rarely identifies the optimal endpoint with confidence; for example, an intervention such as oxytocin may not improve social functioning broadly, but it may meaningfully affect a narrower domain called social-emotional reciprocity [32, 33]. If a trial commits in advance only to a broad endpoint, it may miss a meaningful effect.

Outcome learning addresses this uncertainty by allowing researchers to learn outcomes under prespecified rules and then interpret them clinically after estimation. In this framework, the discovery stage does not search freely for any favorable result. Instead, it uses a constrained procedure to select, weight, or transform outcome components from a predefined instrument or item pool in a way that matches the task. The procedure can incorporate clinical constraints (for example, sparsity, non-negativity or minimum content coverage) so that the final outcome remains interpretable.

Different tasks may justify different outcomes. A prediction study may benefit from an outcome that emphasizes symptom domains that baseline variables can predict reliably [5]. A treatment study may need an outcome that captures the symptom changes most relevant to the intervention mechanism [34]. A causal analysis may require outcome construction that improves robustness to confounding [6]. In each case, the goal is to learn an outcome that better aligns with the clinical and analytic objective.

These task-aligned outcomes need not be limited to predefined subscores or individual items. In some settings, the most informative target may be an estimated latent symptom dimension derived from a specific combination of items [35]. This point is especially important in psychiatry, where many phenomena vary along continuous dimensions rather than forming clean, discrete subtypes [16, 36]. Researchers often seek latent structure, but the methods used can implicitly favor categorical groupings [37]. A factor-mixture perspective may offer a better account, allowing patients to share common latent dimensions while differing in the degree to which those dimensions are expressed. Outcome learning can build on this view by constructing clinically interpretable outcomes that precisely capture the latent dimensions most relevant to a task.

Importantly, outcome learning can also reduce pressure to escalate predictor-model complexity. When researchers define a clearer target, simpler models may recover meaningful effects and yield more interpretable and generalizable findings without the need to subtype patients [5]. This shift can thus improve robustness and implementation potential in real-world settings.

Several methodological families already support outcome learning across different aims. Some methods learn outcomes to maximize predictability from baseline covariates [5], some learn supervised low-dimensional representations that preserve interpretability while improving separation of treatment effects [9, 38], and others construct composite outcomes for causal identification [6]. These examples suggest that outcome learning is not a speculative idea but an emerging practical strategy.

Methodological and Inferential Considerations

Allowing the endpoint to be learned, rather than fixed a priori as a single total score, understandably raises concerns about validity. Readers may worry that learned outcomes will be harder to interpret, harder to compare across studies, or more vulnerable to false positive findings. These concerns do not preclude outcome learning, but they do require that outcome construction be governed by explicit safeguards as summarized in Table 2.

Importantly, task-aligned outcome learning is not a license for post hoc endpoint shopping. It is a prespecified, constrained procedure for defining the study's target outcome. Outcome learning is compatible with preregistration, although preregistration is often interpreted as requiring a single fixed endpoint score [13, 39, 40]. In discovery-oriented work, preregistration can instead specify the instrument item pool, the clinical constraints on outcome construction, and the machine learning algorithm used to derive the outcome for a given task. Confirmatory work can then evaluate replication either by fixing the discovery-derived outcome or by fixing the learning procedure. This workflow preserves transparency while clearly separating discovery from confirmation.

A key nuance is that confirmatory work does not always require transporting a single fixed scoring rule across studies. Fixing the discovery-derived score is most appropriate when the confirmatory study targets a closely matched question with the same or meaningfully equivalent comparator, population, setting, and measurement process. However, in many psychiatric applications, confirmatory evidence is often accumulated in studies that are not *exactly* matched in comparator or context (e.g., bupropion versus placebo in one trial and bupropion versus an active comparator in another). In such settings, insisting on one invariant score can reintroduce the very

Concern	Safeguard
Endpoint shopping	Preregister item pool, constraints, and learning rule; separate discovery from confirmation.
Confirmatory studies must use fixed scores	Fix the discovery-derived score for closely matched tasks; otherwise prespecify the learning procedure and constraints.
Invalid inference (false positives; overfitting)	Valid inference: permutation/bootstrapping with full refitting; prediction assessed on a held-out test set.
Endpoint proliferation / multiplicity	Learn a prespecified, small set of composite outcomes and apply multiplicity control across them.
Poor interpretability	Use constrained, inspectable composites (e.g., sparse, nonnegative, or rule-based); report chosen sets, weights, or transforms.
Subjective discounting of selected symptom content	Interpret outcomes at the item level as measured; if content priorities are desired, encode them prospectively via the item pool and constraints.
Outcome shifts with predictors/treatments	Treat context-specific learned outcomes as distinct endpoints.
Hard to compare with prior work	Report learned + conventional outcomes when feasible.

Table 2 Common concerns about outcome learning and corresponding safeguards. Outcome learning can preserve rigor and interpretability when treated as a prespecified, constrained component of the analytic pipeline and evaluated with appropriate inference and validation procedures.

problem outcome learning is meant to address: the implicit use of a universal endpoint across tasks. A principled alternative is to prespecify the outcome-learning procedure and its constraints, apply the same procedure within each study to learn the task-aligned endpoint for that study's comparator context, and then confirm replication at the level of the scientific claim—namely, that the treatment shows a reproducible pattern of differential change relative to its comparator when the endpoint is optimized for that comparator-defined task. This preserves confirmatory discipline by fixing the procedure and evaluation plan, while avoiding the stronger (and often unjustified) assumption that a single scoring rule should perform optimally across all confirmatory settings.

A common alternative to outcome learning is to analyze symptom domains or individual items directly. Item-level analyses can be valuable for mechanism and safety profiling, but they introduce two practical costs. First, they replace one endpoint with many, creating a substantial multiplicity burden: even with false discovery rate control, power can drop sharply when effects are modest and dispersed across correlated items, and family-wise error control can be prohibitively conservative [41]. A large number of tested outcomes also expands analytic degrees of freedom (items, domains, contrasts, time windows), increasing the risk of false positives unless decisions are tightly prespecified [42]. Second, item-level inference implicitly treats the available questionnaire items as the scientific targets. As mentioned previously, the clinically meaningful quantity is better viewed as a *latent* outcome that is only imperfectly measured by any single item and may not be captured identically across instruments or cohorts. Because individual items are noisy and often blend constructs, latent targets rarely map one-to-one onto a particular question. In that setting, learning a small number of constrained, clinically interpretable composites can be a pragmatic middle ground: it reduces multiplicity relative to item-by-item testing, improves signal-to-noise by aggregating consistent item-level variation, and yields outcomes that are more feasible to replicate than a long list of item-specific effects.

Valid statistical inference also requires evaluation procedures that take into account the learning of outcomes from the data. Many familiar tests implicitly assume a fixed, prespecified outcome;

applying them unchanged after outcome learning can therefore produce misleading p-values. In randomized studies, valid inference can be obtained with permutation-based procedures that rerun the *entire* outcome learning pipeline for each permuted treatment assignment [9, 34]. Uncertainty can be summarized with confidence intervals obtained by bootstrap procedures that likewise rerun the full pipeline within each resample. In prediction settings, validity similarly hinges on genuinely out-of-sample assessment: both the learned outcome and the predictor model should be evaluated on an independent test set [5]. From this perspective, concerns about "small samples" or instability are empirical rather than definitional—outcome learning does not automatically imply overfitting, and there is no universal sample-size cutoff beyond which it becomes invalid. The practical question is whether the learning procedure yields reproducible signal under appropriate evaluation, for example via permutation-based p-values in randomized studies or held-out performance in prediction. More broadly, outcome construction should be treated as an integral part of statistical inference, not as a preprocessing step that can be ignored.

Interpretation requires particular care. Investigators may be tempted to discount certain items that receive weight in a learned composite by imposing external judgments about what *should* count as part of the outcome—for example, by asserting that sleep or appetite/weight items primarily reflect "side effects" rather than "therapeutic" change (as in the bupropion example in the Introduction), or by arguing after the fact that the selected item content is "not clinically relevant." Outcome learning is not intended to support these moves. Clinical relevance is specified *up front* through the choice of instrument and item pool: the items are included precisely because they are taken to measure clinically meaningful distress or impairment. The instrument therefore defines the measurement space, and the learning procedure reweights observed item responses to optimize a prespecified task under explicit constraints. When particular items receive substantial weight, the appropriate interpretation is literal and task-specific: variation in those measured responses contributes materially to prediction or to differentiating trajectories between conditions in the study context. The learned composite does not adjudicate whether an item reflects benefit, tolerability, or some other construct beyond its measured content; it only identifies which elements of the prespecified measurement space are empirically informative for the stated objective. If investigators or stakeholders wish the endpoint to emphasize a narrower subset of content, that preference should be formalized prospectively through the item pool and constraints, rather than introduced after estimation through interpretive discounting.

Finally, learned outcomes may change when the predictor set or treatment changes. This is often framed as a threat to comparability, but it is often expected: changing predictors changes what information is available to be predicted, and changing treatments changes what symptom domains are plausibly affected. The appropriate safeguard is to make this dependence explicit—define the task, prespecify the candidate item pool and constraints, and treat predictor- or treatment-specific learned outcomes as aligned, interpretable endpoints rather than as one universal score. When feasible, reporting both learned and conventional outcomes can further support transparency and comparability with prior work.

In sum, outcome learning is most defensible when it is treated as a design choice that is specified in advance, evaluated end-to-end, and validated externally. The central trade-off is not between "fixed" and "flexible" endpoints, but between fixing a broad scoring rule before task-relevant signal is known versus prespecifying a constrained procedure that can identify a small set of empirically informative targets and then subject them to confirmatory testing. When combined with appropriate inference, multiplicity control, explicit constraints that preserve interpretability, and replication across cohorts and settings, task-aligned outcome learning can reduce endpoint dilution without sacrificing transparency or comparability. This framework therefore shifts the primary object of standardization from a single omnibus score to a reproducible pipeline for deriving, fixing, and validating clinically interpretable outcomes that match the scientific task.

A Concrete Example with Antidepressants

To illustrate outcome learning concretely, we revisit the running example with bupropion. Our goal is to test whether bupropion has a symptom profile that differs from other antidepressants,

given its distinct pharmacology as a norepinephrine–dopamine reuptake inhibitor [43]. The results reported here were previously published in [9], and we follow the best practices reported in Table 2.

To examine the hypothesis, we first analyzed Levels 2 and 2A of the STAR*D trial [44], which we treat as the constrained *discovery* stage: the item pool is fixed (QIDS-SR items) and the outcome-learning algorithm is prespecified. Patients received bupropion-SR or venlafaxine-XR (sertraline was also included in Level 2 but is omitted here for clarity; it was analyzed in [9]). Treatment response in STAR*D was commonly summarized using the total score of the 16-item QIDS-SR [45]. However, the QIDS-SR total score showed substantial overlap between bupropion-SR and venlafaxine-XR (Figure 1 (b)). Historically, clearer evidence for differential effects emerged only after large-scale meta-analytic synthesis [46]. In contrast, our aim is to detect clinically meaningful differential effects earlier, without requiring many trials to accumulate.

We applied an outcome-learning method, Supervised Varimax, to construct task-aligned symptom composites that sharpened treatment differences. Supervised Varimax is an algorithm that learns a small number of orthogonal symptom composites whose weights are chosen to maximize between-arm separation while remaining directly interpretable as item loadings [9]. In Levels 2/2A of STAR*D, the loadings of the learned outcome indicated that hypersomnia improved more with bupropion than with venlafaxine (Figure 1(a), right), whereas venlafaxine showed relative advantages across most other depressive symptoms (Figure 1(a), left). The resulting composite is interpretable because each symptom weight quantifies the extent to which that symptom favors bupropion versus venlafaxine.

Using permutation testing that reruns the full pipeline under permuted treatment assignment and controls family-wise error rate (FWER) across the learned composites, we found a significant difference in the learned outcome (difference $= -0.384$, $p_{\text{FWER}} = 0.007$, Figure 1(c)). Here, we summarize separation as the standardized difference in mean outcomes between arms at the final follow-up visit (similar to Cohen's d and negative values indicate smaller improvement for

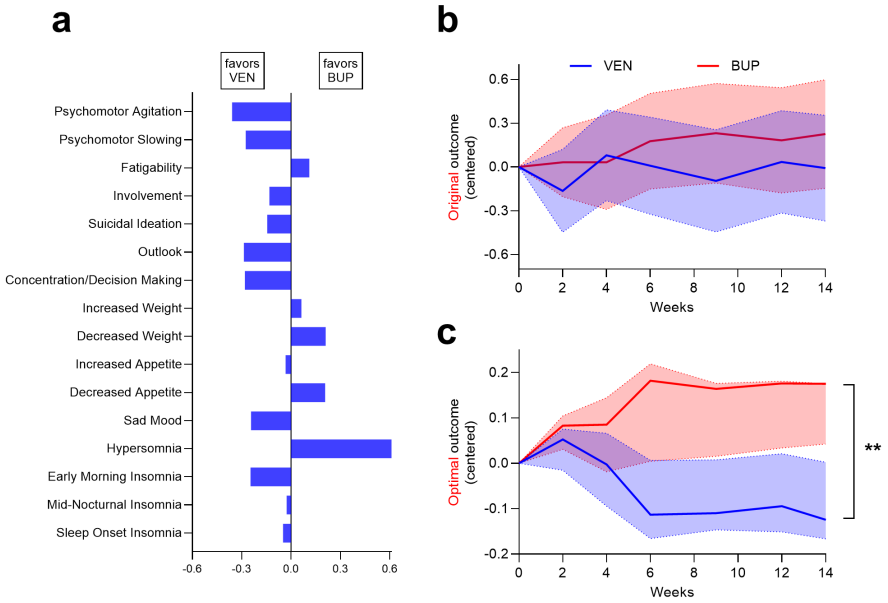

Fig. 1 STAR*D Levels 2/2A: outcome learning yields an interpretable symptom composite with stronger treatment separation. Adapted from [9]. (a) Symptom weights defining the learned composite, shown separately for symptoms favoring bupropion-SR (right) versus venlafaxine-XR (left). (b) Longitudinal trajectories using the conventional QIDS-SR total score. The outcome is mean-centered to highlight between-treatment differences over time. (c) Longitudinal trajectories using the learned composite, with uncertainty shown via bootstrap confidence intervals; overall separation is assessed using a permutation-based test with family-wise error control. * $p_{\text{FWER}} < 0.05$, ** $p_{\text{FWER}} < 0.01$.

bupropion). We also constructed the 95% confidence intervals in Figure 1(c) via the bootstrap by resampling with replacement and refitting the entire pipeline on each resample.

Recall our hypothesis: bupropion should exhibit a symptom-change profile that differs from other antidepressant strategies because of its distinct mechanism of action. We therefore evaluated this claim in CO-MED as the *confirmatory* stage. Because confirmatory trials often differ in comparators, we define replication at the level of the claim—detectable separation between a bupropion-containing regimen and at least one non-bupropion strategy in an independent trial—rather than as transporting identical symptom weights. Accordingly, CO-MED did not compare bupropion-SR with venlafaxine-XR; it compared escitalopram plus bupropion-SR with venlafaxine-XR plus mirtazapine.

As in STAR*D, the QIDS-SR total score provided little discrimination between these augmentation strategies (Figure 2(b)). Because the treatment setup differed, we followed the recommendations in Table 2 and learned a trial-specific outcome within CO-MED just like we did with STAR*D. The learned outcome suggested that escitalopram plus bupropion was superior to venlafaxine plus mirtazapine across most symptoms, with exceptions for decreased appetite, decreased weight, and insomnia, which received relatively greater weight in favor of the mirtazapine-containing strategy (Figure 2(a)). This pattern is pharmacologically plausible given mirtazapine's antihistaminergic (and related sedative/appetite-stimulating) effects [47]. The learned outcome also produced clearer longitudinal separation between the bupropion- and mirtazapine-containing strategies than the total score (difference $= 0.302$, $p_{\mathrm{FWER}} = 0.022$, Figure 2(c)). We again obtained p-values via permutation testing and constructed confidence intervals via bootstrap resampling, refitting the full pipeline on each resample as recommended in Table 2.

The point of this exercise was not that outcome learning retrospectively discovers what clinicians already recognize about bupropion. Rather, the underlying differential signal appears to be real, yet broad total scores can fail to detect it. By increasing sensitivity to task-relevant symptom patterns, outcome learning can recover clinically meaningful differences from individual trials. This is particularly important for novel treatments, where the clinically relevant symptom

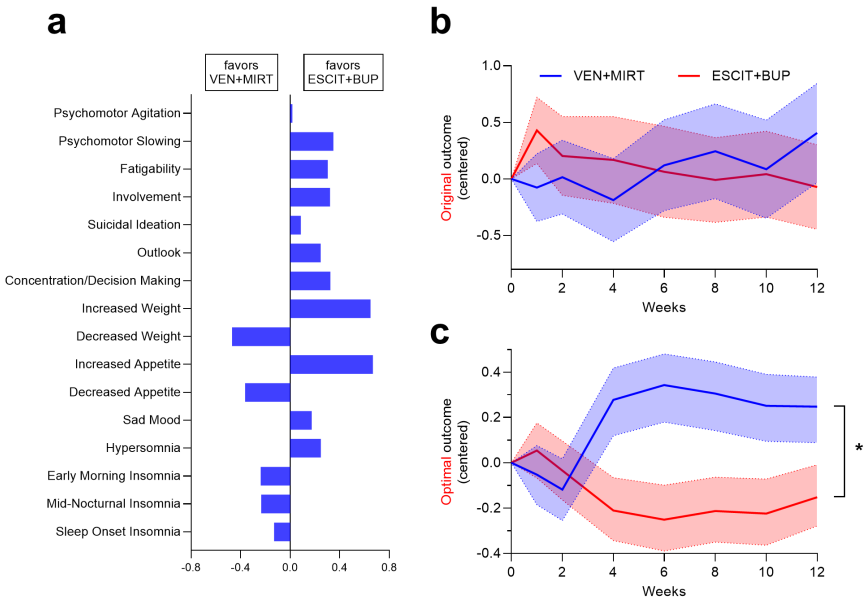

Fig. 2 CO-MED: trial-specific outcome learning under a non-identical replication design. Adapted from [9]. (a) Symptom weights for the CO-MED learned composite contrasting escitalopram plus bupropion augmentation (right) with venlafaxine-XR plus mirtazapine augmentation (left). (b) Longitudinal trajectories using the QIDS-SR total score. (c) Longitudinal trajectories using the learned composite, with uncertainty from bootstrap resampling; significance is evaluated by permutation testing applied to the full refitted pipeline.

profile may be unknown and where earlier identification of differential effects could improve trial interpretation and downstream development decisions.

Conclusion

Broad symptom inventories and standardized scales remain essential for communication and cumulative science in psychiatry, but fixing a single omnibus scoring rule too early can obscure the very signals those instruments contain. When predictability or treatment responsiveness concentrates in a subset of symptoms, total scores can average away task-relevant change, leading to underpowered prediction, null trials, and overly pessimistic interpretations. Outcome learning reframes heterogeneity from an obstacle to a design input: by prespecifying an item pool, constraints, and a machine learning algorithm, investigators can derive clinically interpretable, task-aligned targets that better match the inferential goal while retaining transparency and comparability.

This Perspective also argues that a key methodological decision is *when* to fix outcomes—not whether outcomes should be fixed at all. That timing depends on how closely the confirmatory setting matches the one in which the outcome was learned: when settings are closely aligned, a learned outcome can be transported and treated as a fixed endpoint; when comparators or contexts differ, confirmatory discipline is better achieved by fixing the learning procedure and constraints, then evaluating replication using relearned, task-aligned outcomes. With appropriate safeguards—constrained construction, end-to-end inference (e.g., permutation and bootstrap refitting), multiplicity control, and external validation—outcome learning can reduce endpoint dilution without inviting post hoc endpoint shopping. Empirically evaluating where this framework yields the largest gains is a pragmatic agenda for precision psychiatry, and may advance the field as much by improving the target of analysis as by refining predictors or models.

Finally, the above agenda defines a concrete opportunity for machine learning in psychiatry. Psychiatric studies are unusually well positioned for outcome learning because commonly used questionnaires and rating scales already contain many clinically meaningful, task-relevant items. That structure creates room to optimize outcomes in ways that are often less available in other domains. Rather than relying primarily on off-the-shelf algorithms developed for different problems, the field should continue to develop methods tailored to psychiatric measurement, with explicit constraints that preserve clinical interpretability, comparability and external validity. The central design question is therefore not only whether to learn outcomes, but how to align the learning objective and constraints with the clinical question.

References

[1] Kasper, S. & Dold, M. Factors contributing to the increasing placebo response in antidepressant trials. *World Psychiatry* **14**, 304 (2015).

[2] Sverdrup, E., Petukhova, M. & Wager, S. Estimating treatment effect heterogeneity in psychiatry: a review and tutorial with causal forests. *International Journal of Methods in Psychiatric Research* **34**, e70015 (2025).

[3] De Vries, Y. A., Schoevers, R. A., Higgins, J. P., Munafò, M. R. & Bastiaansen, J. A. Statistical power in clinical trials of interventions for mood, anxiety, and psychotic disorders. *Psychological Medicine* **53**, 4499–4506 (2023).

[4] Reiter, J. E. *et al.* Increasing psychopharmacology clinical trial success rates with digital measures and biomarkers: Future methods. *NPP—Digital Psychiatry and Neuroscience* **2**, 7 (2024).

[5] Strobl, E. Agrawal, M. *et al.* (eds) *Predicting the predictable in the psychiatric high risk.* (eds Agrawal, M. *et al.*) *Proceedings of the 10th Machine Learning for Healthcare Conference*, Vol.

298 of *Proceedings of Machine Learning Research* (PMLR, 2025). URL https://proceedings.mlr.press/v298/strobl25a.html.

[6] Strobl, E. V. Learning causally predictable outcomes from psychiatric longitudinal data. *Biocomputing 2026: Proceedings of the Pacific Symposium* 158–172 (2026).

[7] Anderson, A. E. *et al.* Measuring pathology using the panss across diagnoses: Inconsistency of the positive symptom domain across schizophrenia, schizoaffective, and bipolar disorder. *Psychiatry Research* **258**, 207–216 (2017).

[8] Rush, A. J. *et al.* The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry* **54**, 573–583 (2003).

[9] Strobl, E. V. Consistent differential effects of bupropion and mirtazapine in major depression. *Journal of Affective Disorders* **388**, 119551 (2025).

[10] Hunt, A. *et al.* Systematic review of clinical prediction models for psychosis in individuals meeting at risk mental state criteria. *Frontiers in Psychiatry* **15**, 1408738 (2024).

[11] Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Journal of British Surgery* **102**, 148–158 (2015).

[12] Hieronymus, F., Emilsson, J. F., Nilsson, S. & Eriksson, E. Consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression. *Molecular Psychiatry* **21**, 523–530 (2016).

[13] Turner, L. *et al.* Consolidated standards of reporting trials (consort) and the completeness of reporting of randomised controlled trials (rcts) published in medical journals. *The Cochrane Database of Systematic Reviews* **2012**, MR000030 (2012).

[14] American Psychiatric Association, D., American Psychiatric Association, D. *et al. Diagnostic and statistical manual of mental disorders: DSM-5* Vol. 5 (American Psychiatric Association, Washington, DC, 2013).

[15] Regier, D. A., Kuhl, E. A. & Kupfer, D. J. The dsm-5: Classification and criteria changes. *World Psychiatry* **12**, 92–98 (2013).

[16] Kotov, R. *et al.* The hierarchical taxonomy of psychopathology (hitop): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology* **126**, 454 (2017).

[17] Insel, T. *et al.* Research domain criteria (rdoc): toward a new classification framework for research on mental disorders. *American Journal of psychiatry* **167**, 748–751 (2010).

[18] Pilkonis, P. A. *et al.* Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (promis®): depression, anxiety, and anger. *Assessment* **18**, 263–283 (2011).

[19] Shiffman, S., Stone, A. A. & Hufford, M. R. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* **4**, 1–32 (2008).

[20] De Los Reyes, A. *et al.* The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin* **141**, 858 (2015).

[21] Zhao, X., Coxe, S., Sibley, M. H., Zulauf-McCurdy, C. & Pettit, J. W. Harmonizing depression measures across studies: A tutorial for data harmonization. *Prevention Science* **24**, 1569–1580

(2023).

[22] Meehan, A. J. *et al.* Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Molecular Psychiatry* **27**, 2700–2708 (2022).

[23] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019).

[24] Ahmed, M. I. *et al.* A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus* **15**, e46454 (2023).

[25] Chekroud, A. M. *et al.* Illusory generalizability of clinical prediction models. *Science* **383**, 164–167 (2024).

[26] Lasko, T. A., Strobl, E. V. & Stead, W. W. Why do probabilistic clinical models fail to transport between sites. *NPJ Digital Medicine* **7**, 53 (2024).

[27] U.S. Food and Drug Administration. Enrichment strategies for clinical trials to support determination of effectiveness of human drugs and biological products: Guidance for industry. Guidance document (2019).

[28] Drysdale, A. T. *et al.* Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine* **23**, 28–38 (2017).

[29] Posternak, M. A., Zimmerman, M., Keitner, G. I. & Miller, I. W. A reevaluation of the exclusion criteria used in antidepressant efficacy trials. *American Journal of Psychiatry* **159**, 191–200 (2002).

[30] Rothwell, P. M. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet* **365**, 82–93 (2005).

[31] Zimmerman, M., Chelminski, I. & Posternak, M. A. Generalizability of antidepressant efficacy trials: differences between depressed psychiatric outpatients who would or would not qualify for an efficacy trial. *American Journal of Psychiatry* **162**, 1370–1372 (2005).

[32] Sikich, L. *et al.* Intranasal oxytocin in children and adolescents with autism spectrum disorder. *New England Journal of Medicine* **385**, 1462–1473 (2021).

[33] Watanabe, T. *et al.* Mitigation of sociocommunicational deficits of autism through oxytocin-induced recovery of medial prefrontal activity: a randomized trial. *JAMA Psychiatry* **71** (2014).

[34] Strobl, E. V. Oxytocin enhances social-emotional reciprocity in autism. *medRxiv* 2025–07 (2025).

[35] Otto, M. E. *et al.* Item response theory in early phase clinical trials: Utilization of a reference model to analyze the montgomery-åsberg depression rating scale. *CPT: Pharmacometrics & Systems Pharmacology* **12**, 1425–1436 (2023).

[36] DeYoung, C. G. *et al.* The hierarchical taxonomy of psychopathology and the search for neurobiological substrates of mental illness: A systematic review and roadmap for future research. *Journal of psychopathology and clinical science* **133**, 697 (2024).

[37] Haslam, N., McGrath, M. J., Viechtbauer, W. & Kuppens, P. Dimensions over categories: A meta-analysis of taxometric research. *Psychological Medicine* **50**, 1418–1432 (2020).

[38] Strobl, E. V. & Kim, S. Learning outcomes that maximally differentiate psychiatric treatments. *medRxiv* 2024–12 (2024).

[39] World Health Organization. Who trial registration data set (trds), version 1.2.1 (archived). World Health Organization (ICTRP) (2026). Accessed 2026-02-27.

[40] Chan, A.-W. *et al.* Spirit 2013 statement: defining standard protocol items for clinical trials. *Annals of Internal Medicine* **158**, 200–207 (2013).

[41] Li, G. *et al.* An introduction to multiplicity issues in clinical trials: the what, why, when and how. *International Journal of Epidemiology* **46**, 746–755 (2017).

[42] Gelman, A. & Loken, E. The statistical crisis in science. *The Best Writing on Mathematics (Pitici M, ed)* **102**, 305–318 (2016).

[43] Stahl, S. M. *et al.* A review of the neuropharmacology of bupropion, a dual norepinephrine and dopamine reuptake inhibitor. *Primary Care Companion to the Journal of Clinical Psychiatry* **6**, 159 (2004).

[44] Rush, A. J. *et al.* Sequenced treatment alternatives to relieve depression (star* d): rationale and design. *Controlled Clinical Trials* **25**, 119–142 (2004).

[45] Rush, A. J. *et al.* Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a star* d report. *American Journal of Psychiatry* **163**, 1905–1917 (2006).

[46] Cipriani, A. *et al.* 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**, 1357–1366 (2018).

[47] Fawcett, J. & Barkin, R. L. Review of the results from clinical studies on the efficacy, safety and tolerability of mirtazapine for the treatment of patients with major depression. *Journal of Affective Disorders* **51**, 267–285 (1998).