Reviewed Preprint
v1 • April 13, 2026
Not revised

For correspondence:
povilas.karvelis@camh.ca

Funding: See page 20

Reviewing editor: Alex Fornito, Monash University, Australia

Evaluating biomarkers and prediction models with E2P Simulator

Povilas Karvelis[1], Daniel Felsky[1,2], Anissa Abi-Dargham[3], Guillermo Horga[4], Andreea O Diaconescu[1,5,6,7]

[1]Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health (CAMH), Toronto, Canada • [2]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada • [3]Department of Psychiatry and Behavioral Health, Renaissance School of Medicine, Stony Brook University, Stony Brook, United States • [4]Department of Psychiatry, Columbia University, New York, United States • [5]Department of Psychiatry, University of Toronto, Toronto, Canada • [6]Institute of Medical Sciences, University of Toronto, Toronto, Canada • [7]Department of Psychology, University of Toronto, Toronto, Canada

eLife Assessment

This **important** work outlines why commonly applied performance metrics in predictive modelling do not accurately reflect translational potential using the example of psychiatric care; it provides a web-based tool to contextualize effect sizes in psychiatry with respect to reliability and base rates, and to calculate the real-world utility of prediction models under different scenarios. The evidence supporting the conclusions is **convincing**, incorporating established psychometric principles that will be of use for multiple fields, along with transparent quantitative logic and example applications. The manuscript would benefit from further details about how the tool can be optimally applied and how the resulting outputs should be interpreted. The work will be of broad interest to both clinical experts and scientists in biomedicine and the life sciences.

https://doi.org/10.7554/eLife.110646.1.sa3

Abstract

Precision psychiatry aims to identify biomarkers and develop prediction models to improve diagnosis, prognosis, and treatment selection. Despite extensive research, translating findings into clinically useful tools remains challenging. Here we argue that one key contributor to this translational gap is a pervasive misalignment between routine statistical analytic practices and the criteria for clinical utility. To address this, we introduce E2P (effect-to-prediction) Simulator, an interactive web tool for estimating the real-world predictive value and clinical utility of biomarkers and prediction models by accounting for real-world outcome base rates and measurement reliability – a procedure we term *predictive utility analysis*. Similar to how power analysis helps optimize research for statistical significance, predictive utility analysis can help optimize it for practical significance. The interactive nature of E2P Simulator makes this approach accessible to non-statistician researchers while also providing publication-ready figures to aid with transparency and standardization of reporting. We demonstrate its application in three key areas: diagnostic (depression, Alzheimer's disease), treatment response (antidepressants), and risk (suicide attempts and psychosis onset) prediction, highlighting conditions under which we may expect clinically meaningful advances. While we focus on translational challenges in psychiatry, the framework and tools presented here address general statistical challenges and are broadly applicable across biomedical and behavioral sciences.

1 Introduction

The search for clinically useful psychiatric biomarkers has been a major pursuit of the past several decades, spanning structural and functional neuroimaging, genetic and multiomic approaches, cognitive and computational assays, and, more recently, digital and deep phenotyping (Singh and Rose, 2009; García-Gutéerrez et al., 2020; Jollans and Whelan, 2018; Abi-Dargham et al., 2023; Stephan and Mathys, 2014; Onnela and Rauch, 2016; Clark et al., 2025). With the emergence of precision medicine (Council et al., 2011; Collins and Varmus, 2015), these efforts have recently culminated in the vision of precision psychiatry: the idea that accounting for individual variability across different modalities can help go beyond broad diagnostic categories and trial-and-error treatments, towards more personalized mental health care (Fernandes et al., 2017; Bzdok and Meyer-Lindenberg, 2018; Rutledge et al., 2019; Patzelt et al., 2018; Huys et al., 2016; Paulus and Thompson, 2021; Karvelis et al., 2023; Van Dellen, 2024).

Despite extensive research and methodological advances, real-world clinical impact of individualized prediction models remains difficult to achieve (Salazar de Pablo et al., 2021). While there is a lot of excitement in the field, with many empirical studies reporting high predictive accuracy, a more critical look suggests that this optimism may be misplaced: overfitting (Meehan et al., 2022; Chekroud et al., 2024; Tornero-Costa et al., 2023; Bracher-Smith et al., 2021), small sample sizes, and biased validation methods are the rule rather than the exception (e.g., Karvelis et al., 2022; Schnack and Kahn, 2016; Hunt et al., 2024; Chen et al., 2025). When it comes to individual biomarkers, most of them have also failed to replicate or prove their clinical utility (Abi-Dargham et al., 2023; Cortese et al., 2023). The literature seems to be filled with weak and false positive findings (Ioannidis and Panagiotou, 2011; Collaboration, 2015; Ioannidis, 2005).

A natural question arises: what is holding back progress? While many factors are at play (Kambeitz-Ilankovic et al., 2022; Felsky et al., 2023), here we argue that one crucial bottleneck is a pervasive misalignment between routine statistical analysis practices and the requirements for clinical translation. Most common analytic approaches are optimized for identifying the existence of effects (p-values, model evidence), while effect sizes, which convey practical significance, often receive much less attention (Loth et al., 2021; Kapur et al., 2012; Wasserstein et al., 2019; Lo et al., 2015; Flora, 2020). This is further exacerbated by failing to account for measurement reliability, which attenuates effect sizes (Karvelis and Diaconescu, 2025a; Karvelis et al., 2023; Hedge et al., 2018), diminishing their translational value. When multiple predictors of small effect sizes are then combined to build prediction models, they end up underperforming (which is often masked by inflated performance metrics due to overfitting). On top of that, predictive performance is often evaluated using metrics (e.g., sensitivity, specificity) (Cortese et al., 2023) that do not account for real-world outcome base rates (e.g., prevalence, response rates) and therefore do not convey their real-world predictive value (Abi-Dargham and Horga, 2016; Carter et al., 2017; Jeni et al., 2013; Brabec et al., 2020; Foody, 2023; Guesné et al., 2024) or clinical utility (Rousson and Zumbrunn, 2011). Although these problems are documented in the literature and many are addressed by guidelines such as TRIPOD (Moons et al., 2015; Collins et al., 2024) and STARD (Bossuyt et al., 2015), a lack of accessible tools for dealing with them presents ongoing challenges for the wider research community.

In what follows, we will elaborate on how these methodological challenges collectively undermine progress towards precision psychiatry (Fig. 1). To address them, we will introduce E2P (effect-to-prediction) Simulator, *www.e2p-simulator.com* (Karvelis and Diaconescu, 2025b), an open-source web tool for evaluating effect sizes and predictive performance by accounting for real-world outcome base rates and measurement reliability to obtain predictive value and clinical utility measures – a procedure we term *predictive utility analysis*. To demonstrate the range of its applications we will consider three key areas: diagnostic, treatment response, and risk prediction, highlighting important insights along the way.

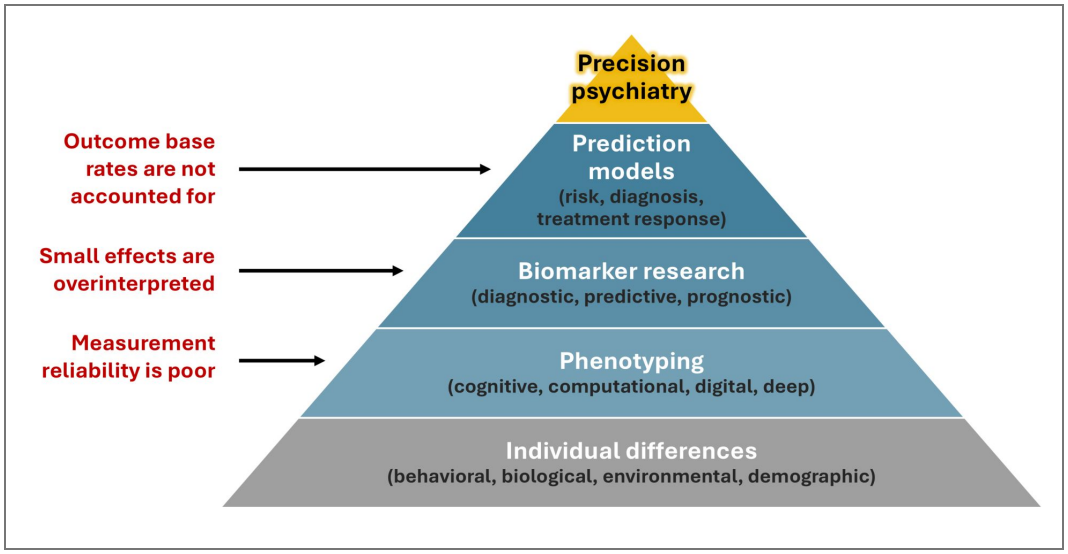


Figure 1. Precision psychiatry: from individual differences to clinical prediction.

The path to precision psychiatry begins with research on individual differences, which exist across many dimensions. The first challenge is to develop tools to measure and characterize these differences (phenotyping). Next, clinically relevant differences must be identified (biomarker research). Finally, collections of relevant markers can be used to build prediction models. However, systemic problems in research practices (highlighted in red) at multiple steps in this pathway undermine progress.

2 Statistical challenges in understanding practical significance

2.1 Statistical significance is not practical significance

Research in psychiatry and behavioral sciences is dominated by null-hypothesis significance testing (NHST) and p-values, which have well-known limitations (Kirk, 1996; Szucs and Ioannidis, 2017; Ioannidis, 2005; Wasserstein and Lazar, 2016; Wasserstein et al., 2019; Yarkoni, 2022; Meehl, 1992; Kapur et al., 2012). Most problematically, based solely on statistical significance, researchers often conclude that their findings have potential for clinical utility - but significant variables are not automatically good predictors (Wasserstein and Lazar, 2016; Lo et al., 2015; Szucs and Ioannidis, 2017; Kapur et al., 2012; Loth et al., 2021; Kühberger et al., 2015). The same limitations apply to more sophisticated measures of Bayesian model evidence (Stephan et al., 2009; Friston et al., 2016; Palminteri et al., 2017; Wilson and Collins, 2019) - both approaches simply quantify the confidence in the existence of the effects, but not their strength (Fig. 2). In the age of increasingly large datasets, this is problematic because increasingly smaller and thus more practically negligible effects become the focus of scientific inquiry (Szucs and Ioannidis, 2017; Marek et al., 2022). To be able to gauge the translational value of research findings, it is therefore important to focus on the estimation and interpretation of effect sizes (Calin-Jageman, 2018; Flora, 2020; Nakagawa and Cuthill, 2007).

2.2 Poor measurement reliability attenuates effect sizes and predictive utility

The lack of focus on effect size interpretation also leads to overlooking the importance of measurement reliability, which attenuates observed effects. This is particularly problematic in psychiatry, where many constructs - from cognitive (Karvelis et al., 2023; Enkavi et al., 2019), to neuroimaging (Elliott et al., 2020; Nikolaidis et al., 2022; Gell et al., 2024; Vidal-Piñeiro et al., 2025), to diagnostic categories themselves (Regier et al., 2013) - are measured with substantial unreliability.

For correlational analysis, the attenuation of observed effects by poor measurement reliability has been long established (Spearman, 1904):

$$r_{observed} = r_{true}\sqrt{ICC_x \cdot ICC_y}, \tag{1}$$

where ICC_x and ICC_y denote the reliabilities of the two variables expressed as Intraclass Correlation Coefficient (ICC). However, this attenuation is rarely accounted for in routine analyses in psychiatry. The attenuation effects are considered even less in the context of group differences (e.g., patients vs. controls), where the formulae describing them were not established until very recently (Karvelis and Diaconescu, 2025a); for a general case considered in this paper:

$$d_{observed} = d_{true}\sqrt{\frac{2ICC_1 \cdot ICC_2}{ICC_1 + ICC_2}\sin(\frac{2}{\pi}\kappa)}, \tag{2}$$

where the observed Cohen's d is attenuated by both predictor reliability in each of the groups (ICC_1 and ICC_2) and the reliability of the group labels themselves (measured as Cohen's κ; e.g., inter-rater agreement).

Given that poor measurement reliability makes the groups overlap more, we would expect it to have important consequences in predictive modeling. Yet, its effects have been explored only by a small number of studies (Gell et al., 2024; Nikolaidis et al., 2022; Jacobucci and Grimm, 2020;

	Research question	Methods	Metrics
	*Does the effect exist?*	Hypothesis testing, in-sample model selection	*p*-values, Bayes factors / model evidence, AIC/BIC
	*How large is the effect?*	Effect size and interval estimation	Cohen's *d*, Pearson's r, OR, confidence intervals
	*How predictive is the effect?*	Predictive modeling	Sensitivity, specificity, accuracy, ROC-AUC, R^2
	*What is the real-world predictive value and utility of the effect?*	Predictive modeling + real-world base rate	PPV, NPV, PR-AUC, Net Benefit

Figure 2. Different levels of research questions with associated methods and metrics.

Much of the research planning and clinical translation strategy is currently formulated based on the mere existence of effects. This can be misleading, as translational potential is proportional not to statistical significance or model evidence, but to effect size. These observed effects can be further expressed as discriminative ability and - by accounting for real-world base rates - as predictive value and clinical utility. Each subsequent step becomes increasingly informative for gauging translational potential, as conveyed by the emojis.

Whittle et al., 2018; Lionetti et al., 2025; Cullen et al., 2023). As we will see later, the nature of the E2P Simulator allows anyone to explore these effects without needing to code their own simulation pipelines.

2.3 The challenge of interpreting effect sizes

Even when effect sizes are reported, researchers struggle to interpret them in a meaningful way. The most common approach is to simply rely on conventional levels of small (d = .2, r = .1), medium (.5, .3), and large (.8, .5) - however, these labels are arbitrary, do not convey translational potential, and were never meant to be used so ubiquitously (Funder and Ozer, 2019; Anvari et al., 2023; Carey et al., 2023; Correll et al., 2020; Giner-Sorolla et al., 2024; Rothman and Greenland, 2018; Durlak, 2009). Another common approach is to compare observed effects with other empirical findings in one's research domain (Lovakov and Agadullina, 2021; Thompson, 2007; Durlak, 2009; Gignac and Szodorai, 2016; Bosco et al., 2015). Although this may serve as a more meaningful reference point, empirically reported effect sizes are often inflated due to publication bias (Schafer and Schwarz, 2019), and, even more importantly, they do not directly convey practical significance (Giner-Sorolla et al., 2024; Kirk, 1996; Kapur et al., 2012). Given that our ultimate goal is to build prediction models to personalize psychiatry, practical significance is best understood in terms of predictive performance (Shmueli, 2010).

2.3.1 ROC-AUC, sensitivity, specificity

Expressing effect sizes in terms of prediction metrics is actually quite straightforward. All we need is to draw a classification threshold between the two groups and determine true and false positives and negatives - all prediction metrics are then derived from these four categories (see Fig. 3). The most common metrics are sensitivity (true positive rate, or proportion of cases correctly identified) and specificity (true negative rate, or proportion of non-cases correctly identified). Moving the threshold across the whole range captures how these metrics trade off against one another, which is captured by the receiver operating characteristic (ROC) curve. Calculating the area under ROC (ROC-AUC) provides an informative summary metric which captures the probability that a predictor ranks a randomly chosen positive case (e.g., a patient with a psychiatric condition) higher than a negative case across all classification thresholds.

Even more straightforwardly, ROC-AUC can be computed directly from Cohen's d:

$$\mathrm{ROC-AUC} = \Phi\left(\frac{d}{\sqrt{2}}\right), \tag{3}$$

where Φ is the cumulative distribution function of the standard normal distribution.

2.4 Prediction metrics can be misleading if real-world outcome base rate is not accounted for

While ROC-AUC, sensitivity, and specificity are useful metrics for assessing a predictor's discriminative performance, they may create a misleading impression of translational potential because these metrics are invariant to outcome base rates (e.g., disease prevalence, treatment response rate, event rate). This is problematic because in clinical contexts many decisions occur under low or varying prevalence (Carter et al., 2017; Abi-Dargham and Horga, 2016). When prevalence is low, most predicted positives will be false - this phenomenon and the associated mistakes are known as the *false positive paradox*, *base rate fallacy*, and *base rate neglect* (Casscells et al., 1978; Kahneman and Tversky, 1973; Tversky and Kahneman, 1974). The relevant metric in such cases is the posterior probability of disease given a positive test, which is most commonly known as positive predictive value (PPV):

$$\mathrm{PPV} = \frac{\phi \cdot \mathrm{Sensitivity}}{\phi \cdot \mathrm{Sensitivity} + (1-\phi)\cdot(1-\mathrm{Specificity})}, \tag{4}$$

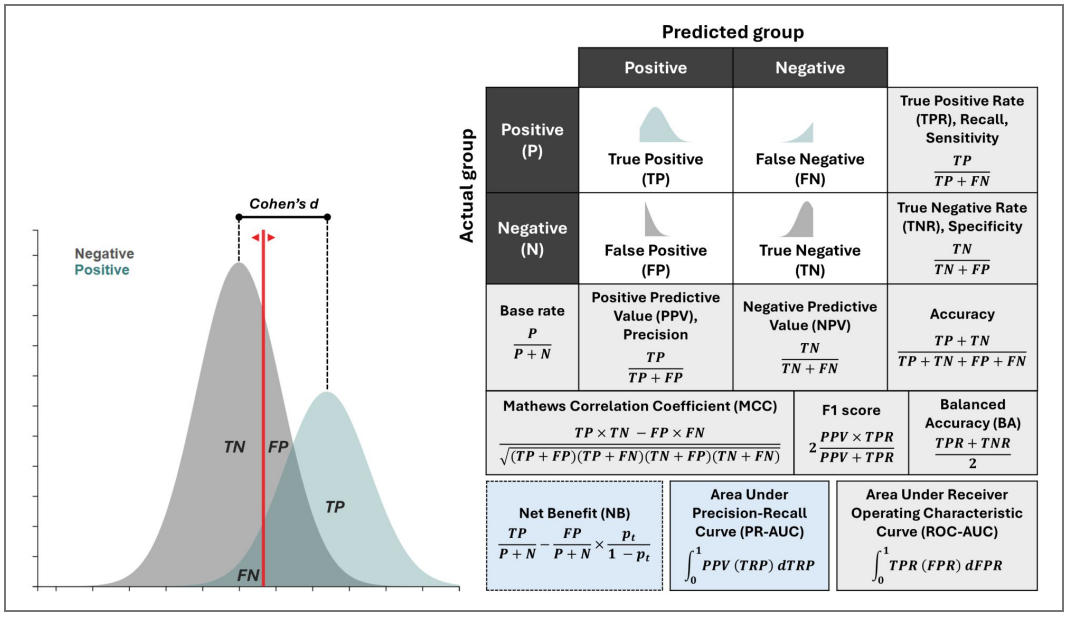


Figure 3. **Effect size vs. prediction metrics.**

Standardized mean difference measures, such as Cohen's d, capture how far the two distributions are from each other. If we place a decision threshold (red) to classify cases into two groups, there are four possible outcomes: True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). These four outcomes are used to derive all classification metrics, as presented on the right. In predictive utility analysis, we place a special emphasis on PR-AUC and NB (highlighted in blue), as they can convey both overall and context-specific model performance while accounting for real-world base rate, making them highly informative for understanding the translational potential of research findings.

where Φ is the base rate.

When base rates are low, achieving high PPV (also known as precision) is only possible at the expense of sensitivity (also known as recall). This trade-off is more transparently summarized by the area under the Precision-Recall curve (PR-AUC), which has been argued to be a more informative metric for imbalanced classification contexts (Saito and Rehmsmeier, 2015; Cartus et al., 2023; Ozenne et al., 2015; Pinker, 2018), although see Van Calster et al. (2025).

However, PR-AUC itself can be misleading when models are evaluated on artificially balanced groups, which is common practice (Jeni et al., 2013; Brabec et al., 2020; Foody, 2023; Guesné et al., 2024). In such cases, the reported PR-AUC reflects the base rate of the test set rather than the deployment conditions. Fortunately, PR-AUC can be rescaled analytically by simply substituting the expected real-world base rate into Eq. 4 and then recomputing the PR-AUC (Brabec et al., 2020). This means that for any ROC-AUC (or PR-AUC, if the base rate of the test set is known) reported in the literature, it is possible to estimate PR-AUC that can be expected under deployment conditions.

2.4.1 Real-world outcome base rates in Decision Curve Analysis

The challenge of accounting for real-world outcome base rates extends beyond standard prediction metrics to more clinically oriented evaluation frameworks, such as decision curve analysis (DCA) (Vickers and Elkin, 2006). DCA is based on calculating Net Benefit (NB), a clinical utility metric which balances true positives (TP) by weighing them against the potential harms of false positives (FP; e.g., overdiagnosis or overtreatment):

$$\mathrm{NB} = \frac{\mathrm{TP}}{N} - \frac{FP}{N} \cdot \frac{p_t}{1 - p_t} \tag{5}$$

where p_t represents the probability threshold used for classification (representing the relative cost of false positives compared to the benefit of true positives) and N is the total number of cases. It can also be expressed in terms of sensitivity, specificity, and base rate Φ:

$$\mathrm{NB} = \mathrm{Sensitivity} \cdot \phi - (1 - \mathrm{Specificity}) \cdot (1 - \phi) \cdot \frac{p_t}{1 - p_t}. \tag{6}$$

In DCA, NB is plotted against a range of probability thresholds, creating a decision curve that shows the net benefit of using the model compared to two alternative strategies: “all” (classifying everyone as positive) and “none” (classifying everyone as negative). The difference in NB between model-based predictions and the default “all” and “none” lines (ΔNB) conveys how much additional clinical utility the model can offer at relevant thresholds. An optimal threshold would reflect the costs of false negatives versus the costs of false positives. For example, a very low threshold (e.g., <5%) is appropriate for suicide risk screening in primary care where the intervention (e.g., safety planning) is low-cost but missing a case is fatal (Ross et al., 2021), whereas a high threshold (e.g., >50%) is required for initiating clozapine in early psychosis due to monitoring costs and side effects (Farooq et al., 2024).

DCA is an increasingly popular method for evaluating clinical prediction models, has desirable decision-theoretic properties (Van Calster et al., 2025), and is recommended by the TRIPOD guidelines (Moons et al., 2015; Collins et al., 2024). However, just as with PR-AUC, NB can be severely inflated if models are evaluated on artificially balanced case-control samples (Rousson and Zumbrunn, 2011). Fortunately, these estimates can also be adjusted post-hoc to reflect real-world base rates without requiring model retraining.

2.5 E2P Simulator and predictive utility analysis

To jointly address all these challenges, we recently developed an open-source web tool **E2P Simulator**, *www.e2p-simulator.com* (Karvelis and Diaconescu, 2025b). It is an interactive tool that provides a general framework for cross-translating across effect size metrics (e.g., Cohen’s d,

Pearson's r), discriminative measures (e.g., ROC-AUC, sensitivity, specificity), predictive value metrics (PR-AUC, PPV, NPV), clinical utility (Net Benefit), and many other metrics, highlighting how all of them describe the same underlying data distribution.

Most importantly, E2P Simulator can be used for evaluating real-world predictive utility of biomarkers and prediction models by accounting for real-world base rates (e.g., disease prevalence, response rates, incidence rates) and measurement reliability in the target populations - a procedure we term *predictive utility analysis*. Similar to how power analysis (Cohen, 1992) can help determine whether an effect can be reliably detected (statistical significance), predictive utility analysis can help determine whether an effect is likely to be useful in practice (practical significance). Following the same analogy, E2P Simulator can be thought of as software for performing predictive utility analysis, the same way G*Power is used for power analysis (Erdfelder et al., 1996).

As such, E2P Simulator can be used for both interpreting existing findings and planning studies. For instance, it can be used to determine whether investing in more reliable measurement instruments would meaningfully improve predictive performance, or to identify optimal target populations and indications with base rates that would yield clinically meaningful predictive values (see Supp. note 1 for more details). Additionally, E2P Simulator can be used for education and reporting standardization. In the next section, we illustrate its use across diagnostic, treatment response, and risk prediction.

3 Demonstration of predictive utility analysis with E2P Simulator

3.1 Diagnostic prediction

3.1.1 A cognitive marker for depression

Consider finding a cognitive biomarker with Cohen's d = 0.8 between healthy controls and a group diagnosed with depression. Such an effect size is rare in practice and would be interpreted as "large"; e.g., some of the strongest biomarkers, such as negative interpretation bias and decreased autobiographical memory specificity, are d < 0.8 (Weiss-Cowie et al., 2023; Everaert et al., 2017). But how would that translate to diagnostic prediction?

First, we have to account for the inter-rater reliability of the depression diagnosis, which is κ = 0.28 as determined by DSM-5 field trials (Regier et al., 2013). Next, given that the effect size is derived by comparing it to healthy controls, the relevant prevalence (or base rate) would be that of the general population, which is around 8% in the USA and Canada (Shorey et al., 2022). Finally, for this hypothetical example, we can assume the biomarker to have ICC = 0.6 test-retest reliability, in line with the average reliability of cognitive markers in recent studies (Karvelis et al., 2023).

With these parameters, the observed Cohen's d = 0.8 will yield ROC-AUC = 0.71 and PR-AUC = 0.19 (Fig. 4a). While conventionally 0.8 would be called "large", the predictive utility is modest as indicated by the low PR-AUC, which captures the trade-off between PPV and sensitivity. Next, we set the classification threshold to correspond to p_t = 15% risk of having depression in the DCA plot (a suitable threshold for screening or diagnosis if false positives have relatively low harms/costs). This results in PPV = 0.22 and sensitivity = 0.31, meaning that at this threshold, 78% of predicted cases would be false positives, and 69% of actual cases would still be missed. This would correspond to ΔNB = 0.009 or 9 additional true positives per 1,000 diagnoses.

Accounting for measurement reliability reveals that this observed effect would correspond to a much larger true effect, d = 1.58, and, in turn, much better predictive performance: ROC-AUC = 0.87, PR-AUC = 0.46, PPV = 0.34, sensitivity = 0.63, and ΔNB = 0.033 at p_t = 0.015, highlighting how much improvement in diagnostic prediction could be achieved simply by improving measurement reliability.

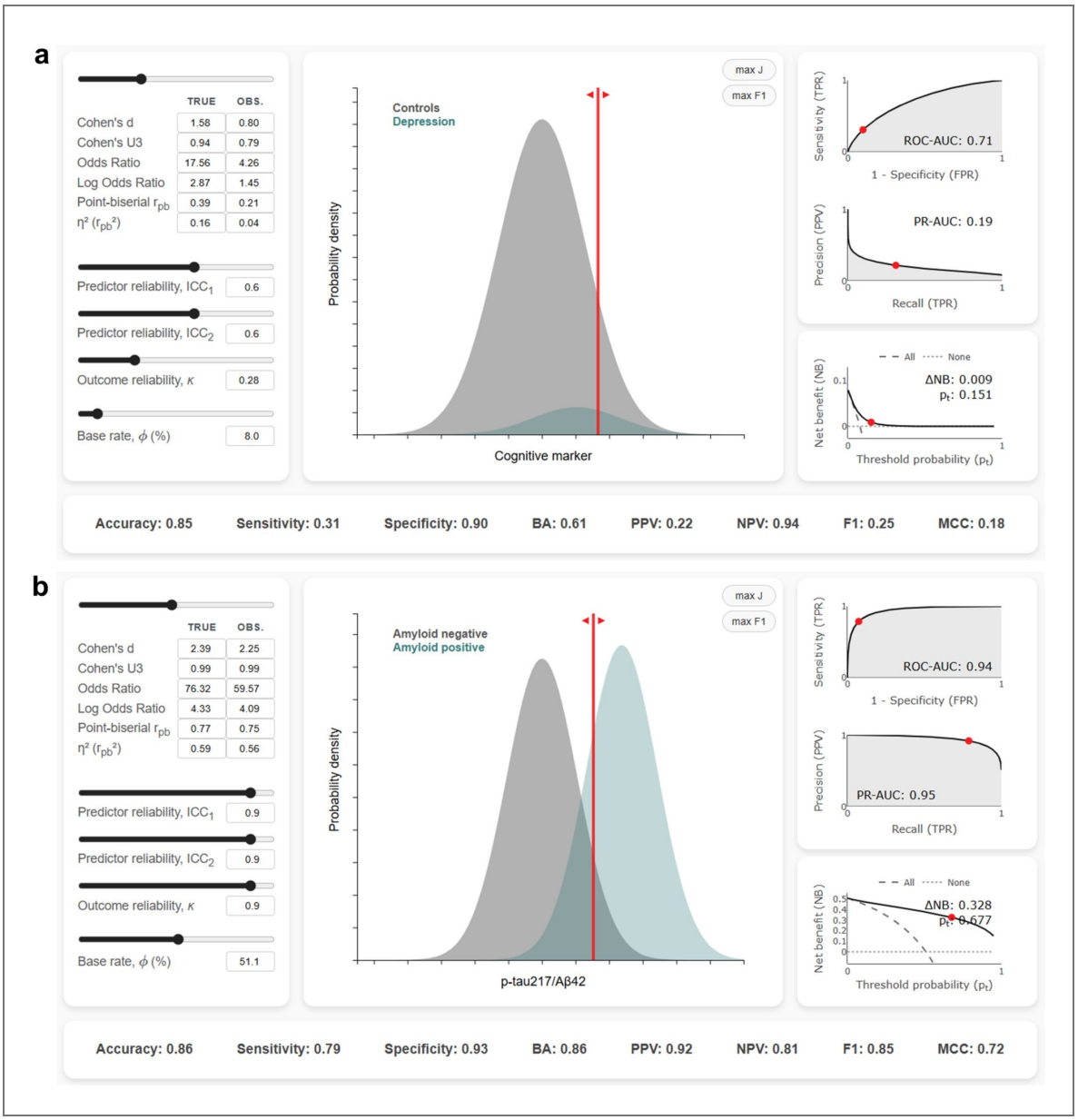


Figure 4. Diagnostic prediction examples.

(**a**) A cognitive marker for depression, (**b**) FDA-cleared diagnostic biomarker p-tau217/Aβ42 for Alzheimer's disease. Both figures are screenshots of E2P Simulator's interface showing inputs on the left (effect size metrics, reliability values, and the base rate), the resulting data distributions in the center, and all predictive and utility metrics at the bottom and on the right. The red line is the classification threshold with the corresponding red markers on ROC-AUC, PR-AUC, and DCA plots.

3.1.2 Combining multiple biomarkers to improve diagnostic prediction

E2P Simulator also includes multivariable calculators for estimating how many biomarkers would be needed to achieve a specified target performance (see Supp. note 2 for more details). Let us say our target is PR-AUC = 0.8, which at 8% prevalence corresponds to ROC-AUC = 0.96. Using the multivariable simulator, we find that with small collinearity among the predictors (0.05), this would require 20 predictors, without collinearity (0.00), it would still require 10 d = 0.8 predictors, while with stronger predictors of d = 1.35, even with a higher collinearity of 0.1, we would need only 5 to achieve the same predictive utility (Supplementary Fig. 1). This suggests that focusing on identifying larger effects (e.g., by improving measurement reliability) may be a more promising research strategy than searching for additional predictors that are weak (think multi-modal big data approaches).

3.2 FDA-cleared diagnostic biomarker p-tau217/Aβ42 for Alzheimer's disease

To provide a reference point to the previous example, let us consider an example from precision medicine, a biomarker recently cleared by the U.S. Food and Drug Administration (FDA): the plasma p-tau217/Aβ342 ratio for aiding Alzheimer's disease diagnosis (FDA, 2025b). More specifically, it was cleared to identify amyloid pathology in ≥ 50 year-olds with cognitive symptoms, based on a study reporting PPV of 91.8% and NPV of 97.3% against amyloid Positron Emission Tomography (PET) or cerebrospinal fluid (CSF) reference standards (FDA, 2025a); note, however, these results were achieved by using two classification thresholds — one for identifying positives and one for identifying negatives — leaving out approximately 20% of cases in the middle that would require confirmatory PET or CSF tests.

Let us recreate this biomarker in E2P Simulator to obtain a clear picture in terms of all the metrics. The prevalence of amyloid positivity in their sample was 51.1% (FDA, 2025a). Test-retest reliability of both p-tau217 and Aβ42 is very high, around ICC ≈ 0.9 (Della Monica et al., 2024). Inter-rater reliability of PET tracers tends to be quite high κ > 0.9 (Harn et al., 2017) and shows high agreement with CSF κ > 0.8; given that the study had a mixture of both, we can set κ ≈ 0.9 as a rough approximation. Now, we jointly set effect size and decision threshold until we achieve the reported PPV ≈ 92% and NPV ≈ 81% (here we re-calculated NPV based on the threshold used for PPV, which gave 54/255 false positives).

We find that this performance corresponds to an observed Cohen's d = 2.25, ROC-AUC = 0.94, and PR-AUC = 0.95 (similar results were also reported in a recent study in China by Wang et al. (2025), achieving ROC-AUC ≈ 0.96); the classification threshold in this case corresponds to p_t = 0.68, which gives ΔNB = 0.324 (Fig. 4b). In real-world settings, we may expect the prevalence to be lower or at least to vary: for cohorts with mild cognitive impairment, amyloid positivity ranges from ~30% at the age of 50 to ~ 60% at the age of 80 (Jansen et al., 2015). However, even at 30% prevalence, d = 2.25 would still result in an impressive PR-AUC = 0.89.

3.3 Treatment response prediction

3.3.1 Multivariable neurocognitive predictors of antidepressant response

Can we predict treatment response to antidepressants? The latest research using multivariable / machine learning models suggests that neurocognitive predictors can explain around R^2 = 0.2 variance in symptom improvement in response to antidepressants and other treatments (Karvelis et al., 2022). While this already combines multiple predictors and may be an optimistic estimate due to overfitting issues, let us consider what this would translate to in practice.

The reliability of task-evoked BOLD responses is, on average, rather low, ICC = 0.4 (Elliott et al., 2020), while the reliability of the most popular scale for assessing depression symptoms, the Hamilton Depression Rating Scale (HAMD), is quite high, ICC = 0.94 (Trajković et al., 2011). The rate of response to antidepressant treatment beyond placebo is about 15% (Stone et al., 2022).

Entering these values into E2P Simulator yields ROC-AUC = 0.73 and PR-AUC = 0.33, indicating rather modest predictive performance, as shown by the low PR-AUC (Fig. 5). At p_t = 0.2, which reflects the relative harms of antidepressant side effects, this would result in sensitivity = 0.54, PPV = 0.29, and ΔNB = 0.030, which means that 46% of those who would benefit from treatment would not receive treatment, 71% of those given treatment would not benefit from it, and we would get additional 3 true responders per 100 people who receive the treatment. Improving measurement reliability alone could improve performance quite substantially, up to ROC-AUC = 0.87 and PR-AUC = 0.57, which at p_t =0.2 would result in sensitivity = 0.74, PPV = 0.42, and ΔNB = 0.073.

Using E2P Simulator we find that to achieve PR-AUC = 0.8, it would require explaining 80% of variance (R^2 = 0.80), which is rather ambitious. This nicely demonstrates the well-known problems of dichotomizing continuous measures (Collins et al., 2016; MacCallum et al., 2002; Royston et al., 2006; Naggara et al., 2011; Streiner, 2002; Karvelis and Diaconescu, 2025a). When symptom improvement on a continuous scale gets converted to responders vs. non-responders, a lot of the information gets lost. This is important to highlight, because most research on treatment response prediction continues to use dichotomized outcome measures (Karvelis et al., 2022; Vieira et al., 2022; Amleshi et al., 2025).

3.3.2 Combining multiple biomarkers to improve treatment response prediction

When building prediction models in neuroscience, a common approach is to first show that predictors are associated with clinical variables, after which both are jointly used to train a classifier (e.g., Hauke et al., 2022; Tozzi et al., 2020; de la Salle et al., 2022; Karvelis et al., 2022). This approach guarantees high collinearity and hurts predictive performance. For example, assuming an average collinearity of 0.15 among all predictors, we would need 17 predictors of r = 0.4 to achieve R^2 = 0.8 (Supplementary Fig. 2). Interestingly, if we reduce the effect size of each predictor to r = 0.3, we would never reach R^2 = 0.8, no matter how many predictors we have (Supplementary Fig. 2).

3.4 Risk Prediction

3.4.1 Mismatch negativity (MMN) as a predictor of psychosis conversion in at-risk individuals

Most research on predictive modeling in psychiatry has thus far focused on predicting the transition to psychosis (Salazar de Pablo et al., 2021), with positive and negative symptoms and verbal memory deficits being the most prominent clinical and cognitive predictors, and mismatch negativity (MMN) emerging as the most promising biomarker (Andreou et al., 2023; Rosen et al., 2021).

How promising is MMN for predicting the transition to psychosis? The primary intended clinical application of MMN is as a prognostic tool within already identified clinical high-risk (CHR) groups. The largest longitudinal study measured MMN at baseline and, after 24 months of follow-up, found that converters showed the largest deficit in the double-deviant (duration + pitch) condition, with an effect size of d = 0.43 (Hamilton et al., 2022). Over a 24-month follow-up period, the average transition rate in CHR cohorts is about 19% (De Pablo et al., 2021). Test-retest reliability for double-deviant MMN is around ICC = 0.5 (Roach et al., 2020). The inter-rater reliability of DSM-5-based criteria for schizophrenia spectrum and other psychotic disorders is κ = 0.46 (Regier et al., 2013).

Putting this all together with E2P Simulator, we find that double-deviant MMN provides only modest predictive performance within CHR cohorts: ROC-AUC = 0.62 and PR-AUC = 0.27 (Fig. 6a). At a decision threshold of p_t = 0.15, which corresponds to relatively low-cost interventions such as increased monitoring, the net benefit is ΔNB = 0.010, with PPV = 0.22 and sensitivity = 0.81. Improving measurement reliability could improve performance up to ROC-AUC = 0.70, PR-AUC = 0.36; at p_t = 0.15 this would give PPV = 0.26, sensitivity = 0.78, ΔNB = 0.026. These numbers highlight that MMN offers incremental discrimination but remains far from sufficient as a stand-alone biomarker.

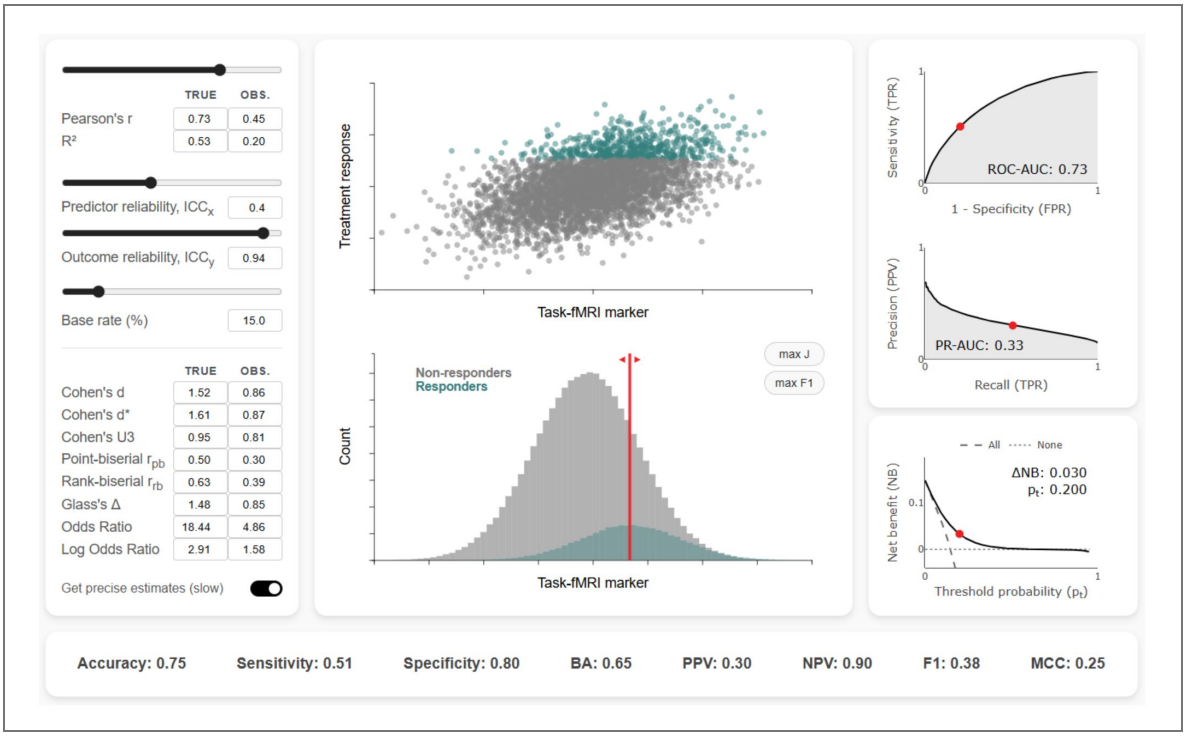


Figure 5. Treatment response prediction example: task-based fMRI for predicting response to antidepressants.

A screenshot of E2P Simulator's interface showing inputs on the left (effect size metrics, reliability values, and the base rate), the resulting data distributions in the center, and all predictive and utility metrics at the bottom and on the right. The red line is the classification threshold with the corresponding red markers on ROC-AUC, PR-AUC, and DCA plots.

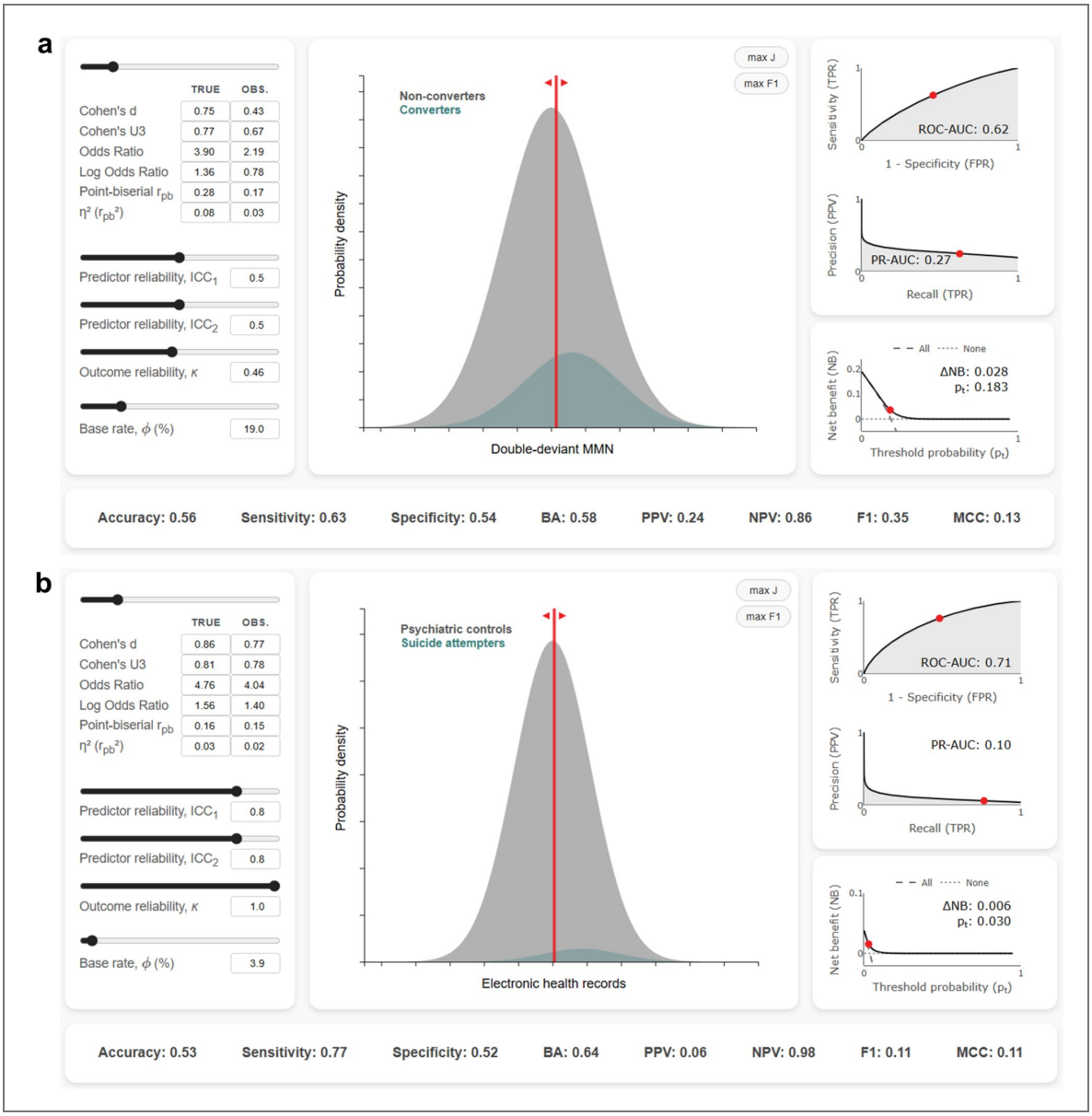


Figure 6. Risk prediction examples.

(**a**) Double-deviant MMN for predicting transition to psychosis, (**b**) electronic health records for predicting suicide attempts. Both figures are screenshots of E2P Simulator's interface showing inputs on the left (effect size metrics, reliability values, and the base rate), the resulting data distributions in the center, and all predictive and utility metrics at the bottom and on the right. The red line is the classification threshold with the corresponding red markers on ROC-AUC, PR-AUC, and DCA plots.

These estimates also need to be considered within the broader context. CHR criteria - based on subthreshold psychotic symptoms, brief episodes, family risk, functional decline - identify only about 6–7% of people who will later develop psychosis (Talukder et al., 2025), meaning that CHR criteria already miss more than 93% of people who will later develop psychosis, substantially limiting the added value of MMN in identifying psychosis risk, and suggesting that a more pressing challenge may be to develop more sensitive CHR criteria.

3.4.2 Predicting 12-month suicide risk using baseline electronic health records

Clinicians rate prediction of suicidality as the highest priority for AI tool development in mental health (Fischer et al., 2025). While many predictive models have been developed using cross-sectional suicidality data (Pigoni et al., 2024), only a few used baseline predictors to estimate prospective risk. One of the largest studies to do so (Edgcomb et al., 2021) followed women (N = 67,000) with serious mental illness for 12 months after a general medical hospitalization and trained models on pre-discharge electronic health records to predict readmission for suicide attempt or self-harm, achieving ROC-AUC of 0.73 (derivation sample) and 0.71 (external sample). A companion study in men (N = 1.4 million) reported similar (AUC ≈ 0.73; derivation sample) results (Thiruvalluru et al., 2023).

Assuming that the 3.9% prevalence reported in the men's cohort is comparable to that in the women's cohort, let us consider how ROC-AUC translates to more informative metrics of PR-AUC and NB. Outcome reliability (hospital admissions for attempts or self-harm) can be assumed to be near perfect ($\kappa \approx 1$), while the overall reliability of structured electronic health records (healthcare utilization, prior attempts, psychiatric diagnoses, etc.) can also be assumed to be rather high ($\kappa \approx 0.8$).

Using these inputs, ROC-AUC = 0.71 yields only PR-AUC = 0.10 (Fig. 6b). Using p_t = 3% absolute risk as a reasonable threshold for intervention, we find sensitivity = 0.77, PPV = 0.06, and ΔNB = 0.006. This means that while the model would capture about three-quarters of true cases, only 6% of those flagged would actually attempt suicide, and the added benefit would translate into finding six additional true cases per 1,000 individuals. Achieving PR-AUC = 0.8 in this population would require ROC-AUC = 0.98. At p_t = 0.03, this would achieve sensitivity = 0.94, PPV = 0.30, and ΔNB = 0.025.

We should also consider the fact that many studies compare attempters to healthy controls, not psychiatric controls. In such cases, a predictor or predictive model should be evaluated using the prevalence of suicide attempts in the general population, which is around 10 times lower, with the prevalence at 12 months being 0.3% (Borges et al., 2010). Such low prevalence, however, makes prediction in the general population an impossible task.

The challenges that low prevalence poses for predicting suicide attempts are well-known Carter et al. (2017); Kessler et al. (2020), so our results here are not surprising. However, we hope that it will help clarify the fundamental prediction challenges for those who continue to work on suicide risk prediction. Considering that the ultimate goal is to reduce suicide rates, improving universal prevention methods may be more productive than developing tools for individual risk assessment (Quinlivan et al., 2017; Steeg et al., 2018).

4 Discussion

In this paper, we highlighted a number of statistical pitfalls that are undermining clinical translation efforts and introduced E2P Simulator together with predictive utility analysis to help researchers address these challenges. We provided detailed examples of how this tool can help interpret research findings and plan research in three key areas: diagnostic, treatment response, and risk prediction. We further showed that even seemingly large effect sizes (Cohen's d) or strong discrimination performance (ROC-AUC) can yield modest predictive value and clinical utility when accounting for real-world outcome base rates. We also demonstrated that poor measurement reliability can be a significant factor in attenuating predictive performance - a research area that remains understudied (Gell et al., 2024; Jacobucci and Grimm, 2020; Whittle et al., 2018;

Cullen et al., 2023). Overall, these observations point to the need to improve measurement reliability and to identify larger effects in order to move closer towards translational goals. This echoes recent calls to find ways to increase effect sizes in neuroimaging and psychiatry research to improve statistical power (DeYoung et al., 2025; Makowski et al., 2024; Jahanshad et al., 2025), but here we directly connect this need to translational challenges.

Although we focused on precision psychiatry to convey our points, the challenges we address are general and apply across all health research in the areas of diagnostic test development, biomarker research, screening instruments, and prediction models (Van Calster et al., 2025; Collins et al., 2025; Leeflang et al., 2013; Maxim et al., 2014; Monsarrat and Vergnes, 2018; McGeechan et al., 2008; Hicks et al., 2022; Rutledge and Loh, 2004), and could be meaningfully applied in many other areas, including forensic psychology (Weber et al., 2025), law (Kirk, 2019; Chin, 2023), and education (Glutting et al., 1997). All these research areas grapple with the challenge of interpreting effect sizes, and using them to inform policy and real-world decision making.

4.1 No single metric tells the whole story

While in this paper we have highlighted the limitations of ROC-AUC and the informativeness of PR-AUC and Net Benefit for assessing translational value, it is important to note that all metrics come with trade-offs (Van Calster et al., 2025). For example, although the invariance of ROC-AUC to base rates can obscure a model's actual translational value, this same property makes it valuable for benchmarking discriminative ability across different contexts. Furthermore, while both ROC-AUC and PR-AUC provide convenient summary measures of performance across the entire range of classification thresholds, only a small range of those thresholds may be clinically relevant. Net Benefit, on the other hand, does require specifying clinically relevant thresholds but that is not always easy to do for non-clinical experts and may also vary based on patients' preferences or resource constraints, leading to inaccurate NB estimates. Finally, rank-based metrics like ROC-AUC and PR-AUC provide information only about relative risk. This does not guarantee that the absolute risk estimates are accurate and match observed probabilities - this is known as calibration (Van Calster et al., 2019); while we did not address it in this paper, E2P Simulator does include a module allowing researchers to explore how measurement reliability and base rates may affect calibration. Overall, a robust evaluation requires a multi-faceted approach, combining multiple complementary metrics and visualization tools. This comprehensive perspective is exactly what E2P Simulator was designed to facilitate.

4.2 Discriminative vs. causal effect sizes

To clarify the scope of E2P Simulator, we must distinguish between discriminative effects, which reflect individual or group differences (e.g., patients vs. controls), and causal effects, which quantify the impact of an intervention (e.g., treatment vs. placebo), as these categories are often conflated (Shmueli, 2010; Ramspek et al., 2021; Dyer, 2025). E2P Simulator was designed for interpreting only discriminative effects, which are relevant for building prediction models. For interpreting causal effect sizes (Kraemer and Kupfer, 2006), other interactive tools are already available: *https://rpsychologist.com/cohend*.

5 Limitations

E2P Simulator in its current form relies on idealized normal distributions. Empirical data may be skewed in different ways (Loth et al., 2021), rendering the parametric metrics less applicable and the resulting conversion across different predictive metrics less exact. To fully mitigate this, predictive utility analysis would need to be performed using the actual data - this will be included in future extensions. Furthermore, the simulations do not automatically account for sampling error, or the uncertainty around the effect size estimates, which is often a concern, especially when samples are small (Ioannidis and Panagiotou, 2011; Button et al., 2013; Ioannidis, 2008). As a workaround, however, E2P Simulator can simply take confidence intervals around the effect size (one at a time) as an input to determine uncertainty around the metrics of interest.

The current implementation is also limited to binary classification, excluding sub-typing (multiclass classification), risk stratification (multiple thresholds), or longitudinal (time-to-event) prediction - although low base rate and poor measurement reliability problems would be just as relevant in these contexts.

Supplementary material

1 Supplementary note 1: predictive utility analysis

Predictive utility analysis consists of two key components: (1) estimating predictive value and clinical utility of research findings by accounting for the real-world outcome base rate, and (2) determining how much performance is lost due to measurement reliability. It can be applied in two main scenarios: interpreting existing research findings and planning new studies. Below we outline the workflow for each and provide some caveats along the way.

1.1 Interpreting existing research findings

When evaluating published research or your own completed studies, the workflow starts with observed metrics and works forward to understand real-world utility:

1. Set measurement reliability: Ideally, reliability estimates should come from the same dataset as the observed effect sizes. If unavailable, use estimates from other relevant research. One could ignore measurement reliability (setting all reliabilities to 1), which would still allow estimating real-world utility but without assessing how much performance is lost due to measurement error.
2. Set observed metrics: This can be an effect size (e.g., Cohen's *d*, OR, Pearson's r) or a predictive performance metric (e.g., ROC-AUC, PR-AUC). Ensure you use a robust or conservative estimate (not inflated due to small samples or overfitting). We set measurement reliability first because observed effect sizes are already attenuated by it.
3. Set base rate: Use the expected prevalence in the real-world population where the predictor will be applied—not the study sample composition. For example, if the study used a balanced case-control design (50% cases, 50% controls) but the condition affects only 5% of the target population, use 5% as the base rate. Note: if using PR-AUC as the observed metric, first enter the study sample's base rate; once set, adjust to the real-world base rate to rescale PR-AUC.
4. Set classification threshold: Choose a threshold probability (p_t) that reflects the clinical context, balancing the costs of false positives against false negatives. The optimal p_t can be estimated as:

$$p_t = \frac{C_{FP}}{C_{FP} + C_{FN}} \quad (1)$$

   where C_{FP} is the cost of a false positive (unnecessary intervention) and C_{FN} is the cost of a false negative (missed case). p_t reflects the absolute risk at which clinical action is warranted, and can be informed by expert surveys, stakeholder preferences, or established guidelines.
5. Document relevant metrics: Record ROC-AUC, PR-AUC, PPV, NPV, and Net Benefit at the chosen threshold. Together, these provide a comprehensive picture of discriminative ability, predictive value, and clinical utility.

1.2 Planning studies

When planning new research, the workflow starts from a target level of real-world performance and works backward to determine what predictors are needed.

1. Identify clinically meaningful targets: Determine what level of PR-AUC, PPV, NPV, or Net Benefit would be clinically meaningful. This could come from cost-benefit analysis, existing guidelines, or benchmarking against existing clinical instruments.
2.

Determine required group separation: Use the simulator to translate clinical targets into required group separation (e.g., Cohen's d, ROC-AUC), setting the base rate to the expected real-world prevalence. Compare with typical effect sizes in the literature to assess feasibility.

3. Explore ways to improve performance: Use the simulator to explore how to get closer to your goal by:
    - Improving measurement reliability of each predictor (e.g., using more reliable assessment methods or protocols)
    - Targeting higher base rate populations: pre-screened or high-risk populations can improve PPV and PR-AUC. However, this should be guided by real-world feasibility. Pre-screening is itself a classification problem that may exclude many actual cases if it has poor sensitivity. Also, group separation found in one population (e.g., healthy controls vs. cases) should not be expected to remain the same in another population (e.g., psychiatric controls vs. cases).
    - Using multiple predictors: estimate how many predictors are needed to achieve target performance. For average effect size and collinearity, use values typical in the field or from your own data. The multivariable calculator provides a rough estimate of model performance without needing to train it.

2 Supplementary note 2: combining multiple predictors

E2P Simulator includes simulators for estimating what predictive performance can be reached by using multiple predictors, when their individual effect sizes and collinearity are known. We start with computing Mahalanobis *D* - a generalization of Cohen's *d* in multidimensional space. Given *p* predictors with Cohen's d values $d_1, d_2, \ldots, d_p$ and a $p \times p$ correlation matrix **R** among them, the general formula for Mahalanobis *D* (Mahalanobis, 1936; Del Giudice, 2009) is:

$$D^2 = \mathrm{d}^T \mathrm{R}^{-1} \mathrm{d} \tag{2}$$

where $\mathbf{d} = (d_1, d_2, \ldots, d_\mathrm{p})^\top$. To make the tool practical, we assume equal effect sizes ($d_1 = d_2 = \ldots = d_\mathrm{p} = d$) and equal pairwise correlations (r_{ij} for all pairs). This simplifies to:

$$D = d \times \sqrt{\frac{p}{1 + (p-1) \times r_{ij}}} \tag{3}$$

The numerator (p) reflects the number of predictors; the denominator adjusts for shared information. When predictors are uncorrelated ($r_{ij} = 0$), each adds fully: $D = d\sqrt{p}$. When perfectly correlated ($r_{ij} = 1$), they are redundant: $D = d$, regardless of p.

Once we obtain D, the conversion to predictive metrics follow the same relationship as for Cohen's

$$\mathrm{ROC-AUC} = \Phi\left(\frac{D}{\sqrt{2}}\right) \tag{4}$$

where Φ is the cumulative normal distribution function. Substituting the simplified *D* gives the full formula:

$$\mathrm{ROC-AUC} = \Phi\left(d \times \sqrt{\frac{p}{2\,(1 + (p-1) \times r_{ij})}}\right) \tag{5}$$

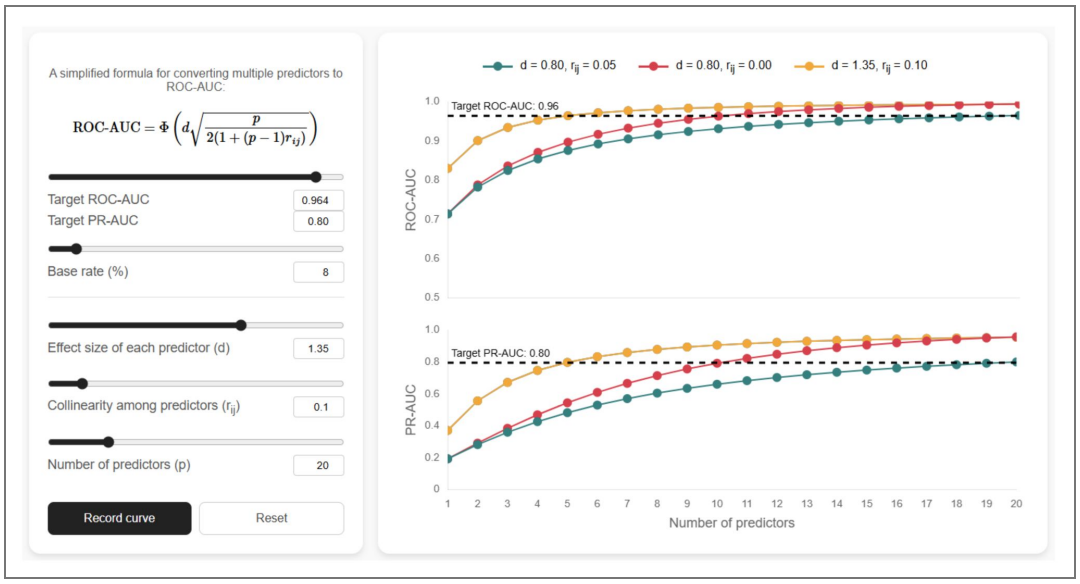
The same logic applies to continuous outcomes. With p predictors each having correlation r with the outcome, and collinearity r_{ij} among predictors:

$$R^2 = \frac{p \times r^2}{1 + (p-1) \times r_{ij}} \tag{6}$$

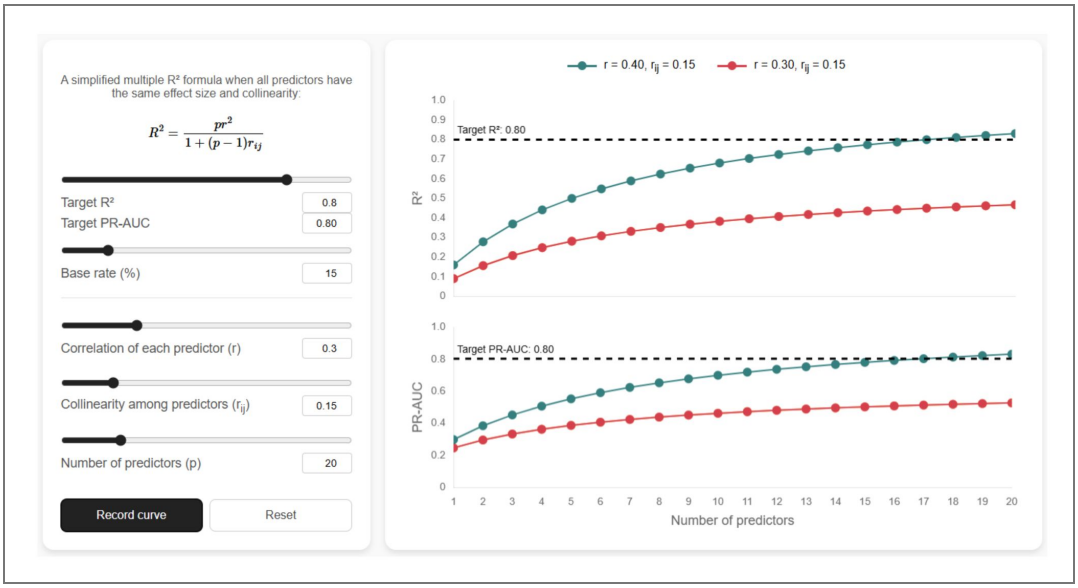
Here, the conversion to ROC-AUC is done non-analytically but by sampling the data after dichotomizing it, just like for single continuous predictors.

The additive (no interactions) assumption means these formulas best approximate linear models such as logistic regression. Research suggests that in clinical prediction, complex non-linear models generally do not outperform logistic regression (Christodoulou et al., 2019), making this a reasonable approximation for many practical settings. Some research fields, like genetics (Hill, 2010), are primarily relying on linear models to begin with. Even when real-world predictors vary in strength and collinearity, the general trends (e.g., diminishing returns with correlated predictors) remain informative for research planning.

In the main text we consider two examples to demonstrate the use of each multivariable calculator (Supp. Fig. 1 and Supp. Fig. 2).



Supplementary Figure 1. Estimating the strength and number of predictors needed for diagnostic prediction in depression. A screenshot of the binary multivariable calculator with inputs on the left and results on the right. Here we show three example settings for achieving PR-AUC = 0.8, which at 8% base rate would correspond to ROC-AUC = 0.96; we would need either 20 predictors of d = 0.8 each and 0.1 collinearity (green line); 10 predictors of d = 0.8 each and no (0.0) collinearity (red line); or 5 predictors of d = 1.35 each with 0.1 collinearity (yellow line).



Supplementary Figure 2. Estimating the strength and number of predictors needed for predicting treatment response to antidepressants A screenshot of the continuous multivariable calculator with inputs on the left and results on the right: the resulting R^2 and PR-AUC as a function of the number of parameters. Here we show two example curves for achieving R^2 = 0.8, which at 15% corresponds to PR-AUC = 0.8; we would need 17 predictors of r = 0.4 to achieve R^2 = 0.8 (green line). Interestingly, if we reduce the effect size of each predictor to r = 0.3, we would never reach R^2 = 0.8, no matter how many predictors we have (red line).

Data availability

This work presents a statistical tool and does not involve empirical data. The source code for E2P Simulator is openly available at *https://github.com/povilaskarvelis/e2p-simulator* under the MIT license. An archived version is available at *https://doi.org/10.5281/zenodo.17112626*.

Acknowledgements

Andreea Diaconescu is supported by the Krembil Foundation, Canadian Institute of Health Research and NSERC Discovery Fund.

Additional information

Funding

Funder	Grant reference number	Author
Krembil Foundation	1000824	Andreea Oliviana Diaconescu

References

Abi-Dargham A., Horga G. (2016) The search for imaging biomarkers in psychiatric disorders. *Nature medicine* **22**:1248-1255 https://doi.org/10.1038/nm.4190 | PubMed

Abi-Dargham A., Moeller S. J., Ali F., DeLorenzo C., Domschke K., Horga G., Jutla A., Kotov R., Paulus M. P., Rubio J. M., *et al.* (2023) Candidate biomarkers in psychiatric disorders: state of the field. *World Psychiatry* **22**:236-262 https://doi.org/10.1002/wps.21078 | PubMed

Amleshi R. S., Ilaghi M., Rezaei M., Zangiabadian M., Rezazadeh H., Wegener G., Arj-mand S. (2025) Predictive utility of artificial intelligence on schizophrenia treatment outcomes: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews* **169**:105968 https://doi.org/10.1016/j.neubiorev.2024.105968 | PubMed

Andreou C., Eickhoff S., Heide M., de Bock R., Obleser J., Borgwardt S. (2023) Predictors of transition in patients with clinical high risk for psychosis: an umbrella review. *Translational Psychiatry* **13**:286 https://doi.org/10.1038/s41398-023-02586-0 | PubMed

Anvari F., Kievit R., Lakens D., Pennington C. R., Przybylski A. K., Tiokhin L., Wiernik B. M., Orben A. (2023) Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters. *Perspectives on Psychological Science* **18**:503-507 https://doi.org/10.1177/17456916221091565 | PubMed

Borges G., Nock M. K., Abad J. M. H., Hwang I., Sampson N. A., Alonso J., Andrade L. H., Angermeyer M. C., Beautrais A., Bromet E., *et al.* (2010) Twelve-month prevalence of and risk factors for suicide attempts in the world health organization world mental health surveys. *The Journal of clinical psychiatry* **71**:21777 https://doi.org/10.4088/jcp.08m04967blu | PubMed

Bosco F. A., Aguinis H., Singh K., Field J. G., Pierce C. A. (2015) Correlational effect size benchmarks. *Journal of applied psychology* **100**:431 https://doi.org/10.1037/a0038047 | PubMed

Bossuyt P. M., Reitsma J. B., Bruns D. E., Gatsonis C. A., Glasziou P. P., Irwig L., Lijmer J. G., Moher D., Rennie D., de Vet H. C., *et al.* (2015) Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 351 https://doi.org/10.1136/bmj.h5527 | PubMed

Brabec J., Komárek T., Franc V., Machlica L. (2020) On model evaluation under non-constant class imbalance. In: Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3-5, 2020, Proceedings, Part IV 20. pp. 74-87

Bracher-Smith M., Crawford K., Escott-Price V. (2021) Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Molecular Psychiatry* **26**:70-79 https://doi.org/10.1038/s41380-020-0825-2 | PubMed

Button K. S., Ioannidis J. P., Mokrysz C., Nosek B. A., Flint J., Robinson E. S., Munafò M. R. (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience* **14**:365-376 https://doi.org/10.1038/nrn3475 | PubMed

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

Calin-Jageman R. J. (2018) The new statistics for neuroscience majors: thinking in effect sizes. *Journal of undergraduate neuroscience education* **16**:E21 https://doi.org/10.31234/osf.io/zvm9a | PubMed

Carey E. G., Ridler I., Ford T. J., Stringaris A. (2023) Editorial perspective: When is a 'small effect'actually large and impactful?. *Journal of Child Psychology and Psychiatry* **64**:1643-1647 https://doi.org/10.1111/jcpp.13817 | PubMed

Carter G., Milner A., McGill K., Pirkis J., Kapur N., Spittal M. J. (2017) Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. *The British Journal of Psychiatry* **210**:387-395 https://doi.org/10.1192/bjp.bp.116.182717 | PubMed

Cartus A. R., Samuels E. A., Cerdá M., Marshall B. D. (2023) Outcome class imbalance and rare events: An underappreciated complication for overdose risk prediction modeling. *Addiction* **118**:1167-1176 https://doi.org/10.1111/add.16133 | PubMed

Casscells W., Schoenberger A., Graboys T. B. (1978) Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine* **299**:999-1001 https://doi.org/10.1056/nejm197811022991808 | PubMed

Chekroud A. M., Hawrilenko M., Loho H., Bondar J., Gueorguieva R., Hasan A., Kambeitz J., Corlett P. R., Koutsouleris N., Krumholz H. M., *et al.* (2024) Illusory generalizability of clinical prediction models. *Science* **383**:164-167 https://doi.org/10.1126/science.adg8538 | PubMed

Chen J., Sahlas E., Zhou Y., Chen N., Xie J., Wadia F., Armstrong S., Caciagli L., Weil A., Dudley R., *et al.* (2025) The use of artificial intelligence in magnetic resonance imaging of epilepsy: A systematic review and meta-analysis. *bioRxiv* https://doi.org/10.1101/2025.09.19.677393

Chin J. M. (2023) Law and psychology must think critically about effect sizes. *Discover Psychology* **3**:3 https://doi.org/10.1007/s44202-022-00062-2 | PubMed

Clark S. R., Arnet V. K., Jawahar M. C., Toben C., Schubert K. O., Amare A. T. (2025) Using multi-omic signatures in the understanding of inflammation and psychosis: Can methylation-derived white blood cell proportions guide early intervention?. *Biological Psychiatry Global Open Science* **5**:100580 https://doi.org/10.1016/j.bpsgos.2025.100580 | PubMed

Cohen J. (1992) Statistical power analysis. *Current directions in psychological science* **1**:98-101

Collaboration O. S. (2015) Estimating the reproducibility of psychological science. *Science* **349**:aac4716 https://doi.org/10.1126/science.aac4716 | PubMed

Collins F. S., Varmus H. (2015) A new initiative on precision medicine. *New England journal of medicine* **372**:793-795 https://doi.org/10.1056/nejmp1500523 | PubMed

Collins G. S., Chester-Jones M., Gerry S., Ma J., Matos J., Sehjal J., Tsegaye B., Dhiman P. (2025) Clinical prediction models using machine learning in oncology: challenges and recommendations. *BMJ oncology* **4**:e000914 https://doi.org/10.1136/bmjonc-2025-000914 | PubMed

Collins G. S., Moons K. G., Dhiman P., Riley R. D., Beam A. L., Van Calster B., Ghassemi M., Liu X., Reitsma J. B., Van Smeden M., *et al.* (2024) Tripod+ ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj* 385

Collins G. S., Ogundimu E. O., Cook J. A., Manach Y. L., Altman D. G. (2016) Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Statistics in medicine* **35**:4124-4135 https://doi.org/10.1002/sim.6986 | PubMed

Correll J., Mellinger C., McClelland G. H., Judd C. M. (2020) Avoid cohen's 'small','medium', and 'large'for power analysis. *Trends in cognitive sciences* **24**:200-207 https://doi.org/10.1016/j.tics.2019.12.009 | PubMed

Cortese S., Solmi M., Michelini G., Bellato A., Blanner C., Canozzi A., Eudave L., Farhat L. C., Højlund M., Köhler-Forsberg O., *et al.* (2023) Candidate diagnostic biomarkers for neurodevelop-mental disorders in children and adolescents: a systematic review. *World Psychiatry* **22**:129-149 https://doi.org/10.1002/wps.21037 | PubMed

Council N. R., Earth D. (2011) Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. on Life Sciences, B., and on A Framework for Developing a New Taxonomy of Disease, C..

Cullen N. C., Janelidze S., Mattsson-Carlgren N., Palmqvist S., Bittner T., Suridjan I., Jethwa A., Kollmorgen G., Brum W. S., Zetterberg H., *et al.* (2023) Test-retest variability of plasma biomarkers in alzheimer's disease and its effects on clinical prediction models. *Alzheimer's & Dementia* **19**:797-806 https://doi.org/10.1002/alz.12706 | PubMed

de la Salle S., Phillips J. L., Blier P., Knott V. (2022) Electrophysiological correlates and predictors of the antidepressant response to repeated ketamine infusions in treatment-resistant depression. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **115**:110507 https://doi.org/10.1016/j.pnpbp.2021.110507 | PubMed

De Pablo G. S., Radua J., Pereira J., Bonoldi I., Arienti V., Besana F., Soardo L., Cabras A., Fortea L., Catalan A., *et al.* (2021) Probability of transition to psychosis in individuals at clinical high risk: an updated meta-analysis. *JAMA psychiatry* **78**:970-978

Della Monica C., Revell V., Atzori G., Laban R., Skene S. S., Heslegrave A., Hassanin H., Nilforooshan R., Zetterberg H., Dijk D.-J. (2024) P-tau217 and other blood biomarkers of dementia: variation with time of day. *Translational psychiatry* **14**:373 https://doi.org/10.1038/s41398-024-03084-7 | PubMed

DeYoung C. G., Hilger K., Hanson J. L., Abend R., Allen T. A., Beaty R. E., Blain S. D., Chavez R. S., Engel S. A., Feilong M., *et al.* (2025) Beyond increasing sample sizes: Optimizing effect sizes in neuroimaging research on individual differences. *Journal of Cognitive Neuroscience* 1-12 https://doi.org/10.1162/jocn_a_02297 | PubMed

Durlak J. A. (2009) How to select, calculate, and interpret effect sizes. *Journal of pediatric psychology* **34**:917-928 https://doi.org/10.1093/jpepsy/jsp004 | PubMed

Dyer B. P. (2025) The distinction between causal, predictive, and descriptive research–there is still room for improvement. *Journal of Clinical Epidemiology* 111960 https://doi.org/10.1016/j.jclinepi.2025.111960 | PubMed

Edgcomb J. B., Thiruvalluru R., Pathak J., Brooks J. O. (2021) Machine learning to differentiate risk of suicide attempt and self-harm after general medical hospitalization of women with mental illness. *Medical care* **59**:S58-S64 https://doi.org/10.1097/mlr.0000000000001467 | PubMed

Elliott M. L., Knodt A. R., Ireland D., Morris M. L., Poulton R., Ramrakha S., Sison M. L., Moffitt T. E., Caspi A., Hariri A. R. (2020) What is the test-retest reliability of common task-functional mri measures? new empirical evidence and a meta-analysis. *Psychological Science* **31**:792-806 https://doi.org/10.1177/0956797620916786 | PubMed

Enkavi A. Z., Eisenberg I. W., Bissett P. G., Mazza G. L., MacKinnon D. P., Marsch L. A., Poldrack R. A. (2019) Large-scale analysis of test-retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences* **116**:5472-5477 https://doi.org/10.1073/pnas.1818430116 | PubMed

Erdfelder E., Faul F., Buchner A. (1996) Gpower: A general power analysis program. *Behavior research methods, instruments, & computers* **28**:1-11 https://doi.org/10.3758/bf03203630

Everaert J., Podina I. R., Koster E. H. (2017) A comprehensive meta-analysis of interpretation biases in depression. *Clinical psychology review* **58**:33-48 https://doi.org/10.1016/j.cpr.2017.09.005 | PubMed

Farooq S., Hattle M., Kingstone T., Ajnakina O., Dazzan P., Demjaha A., Murray R. M., Di Forti M., Jones P. B., Doody G. A., *et al.* (2024) Development and initial evaluation of a clinical prediction model for risk of treatment resistance in first-episode psychosis: Schizophrenia prediction of resistance to treatment (spirit). *The British Journal of Psychiatry* **225**:379-388 https://doi.org/10.1192/bjp.2024.101 | PubMed

FDA (2025a) 510(k) premarket notification summary: Lumipulse G pTau217/β-Amyloid 1-42 Plasma Ratio. Technical Report K242706. FDA.

FDA (2025b) FDA clears first blood test used in diagnosing Alzheimer's disease.

Felsky D., Cannitelli A., Pipitone J. (2023) Whole person modeling: a transdisciplinary approach to mental health research. *Discover mental health* **3**:16 https://doi.org/10.1007/s44192-023-00041-6 | PubMed

Fernandes B. S., Williams L. M., Steiner J., Leboyer M., Carvalho A. F., Berk M. (2017) The new field of 'precision psychiatry'. *BMC medicine* **15**:1-7 https://doi.org/10.1186/s12916-017-0849-x | PubMed

Fischer L., Mann P. A., Nguyen M.-H. H., Becker S., Khodadadi S., Schulz A., Edwin Thanarajah S., Repple J., Hahn T., Reif A., *et al.* (2025) Ai for mental health: clinician expectations and priorities in computational psychiatry. *BMC psychiatry* **25**:584 https://doi.org/10.1186/s12888-025-06957-3 | PubMed

Flora D. B. (2020) Thinking about effect sizes: From the replication crisis to a cumulative psychological science. *Canadian Psychology/Psychologie Canadienne* **61**:318 https://doi.org/10.1037/cap0000218

Foody G. M. (2023) Challenges in the real world use of classification accuracy metrics: From recall and precision to the matthews correlation coefficient. *Plos one* **18**:e0291908 https://doi.org/10.1371/journal.pone.0291908 | PubMed

Friston K. J., Litvak V., Oswal A., Razi A., Stephan K. E., Van Wijk B. C., Ziegler G., Zeidman P. (2016) Bayesian model reduction and empirical bayes for group (dcm) studies. *Neuroimage* **128**:413-431 https://doi.org/10.1016/j.neuroimage.2015.11.015 | PubMed

Funder D. C., Ozer D. J. (2019) Evaluating effect size in psychological research: Sense and nonsense. *Advances in methods and practices in psychological science* **2**:156-168

García-Gutiérrez M. S., Navarrete F., Sala F., Gasparyan A., Austrich-Olivares A., Manzanares J. (2020) Biomarkers in psychiatry: concept, definition, types and relevance to the clinical reality. *Frontiers in psychiatry* **11**:432 https://doi.org/10.3389/fpsyt.2020.00432 | PubMed

Gell M., Eickhoff S. B., Omidvarnia A., Küppers V., Patil K. R., Satterthwaite T. D., Müller V. I., Langner R. (2024) How measurement noise limits the accuracy of brain-behaviour predictions. *Nature Communications* **15**:1-12 https://doi.org/10.1038/s41467-024-54022-6 | PubMed

Gignac G. E., Szodorai E. T. (2016) Effect size guidelines for individual differences researchers. *Personality and individual differences* **102**:74-78 https://doi.org/10.1016/j.paid.2016.06.069

Giner-Sorolla R., Montoya A. K., Reifman A., Carpenter T., Lewis N. A., Aberson C. L., Bostyn D. H., Conrique B. G., Ng B. W., Schoemann A. M., *et al.* (2024) Power to detect what? considerations for planning and evaluating sample size. *Personality and Social Psychology Review* **28**:276-301 https://doi.org/10.1177/10888683241228328 | PubMed

Glutting J. J., Watkins M. M., McDermott P. A., Kush J. C., Konold T. R. (1997) The base rate problem and its consequences for interpreting children's ability profiles. *School Psychology Review* **26**:176-188 https://doi.org/10.1080/02796015.1997.12085857

Guesne S. J., Hanser T., Werner S., Boobier S., Scott S. (2024) Mind your prevalence!. *Journal of Cheminformatics* **16**:43 https://doi.org/10.1186/s13321-024-00837-w | PubMed

Hamilton H. K., Roach B. J., Bachman P. M., Belger A., Carrión R. E., Duncan E., Johannesen J. K., Light G. A., Niznikiewicz M. A., Addington J., *et al.* (2022) Mismatch negativity in response to auditory deviance and risk for future psychosis in youth at clinical high risk for psychosis. *JAMA psychiatry* **79**:780-789 https://doi.org/10.1001/jamapsychiatry.2022.1417 | PubMed

Harn N. R., Hunt S. L., Hill J., Vidoni E., Perry M., Burns J. M. (2017) Augmenting amyloid pet interpretations with quantitative information improves consistency of early amyloid detection. *Clinical nuclear medicine* **42**:577-581 https://doi.org/10.1097/rlu.0000000000001693 | PubMed

Hauke D., Roth V., Karvelis P., Adams R., Moritz S., Borgwardt S., Diaconescu A., Andreou C. (2022) Increased belief instability in psychotic disorders predicts treatment response to metacognitive training. Schizophrenia Bulletin.

Hedge C., Powell G., Sumner P. (2018) The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods* **50**:1166-1186 https://doi.org/10.3758/s13428-017-0935-1 | PubMed

Hicks S. A., Strómimke I., Thambawita V., Hammou M., Riegler M. A., Halvorsen P., Parasa S. (2022) On evaluation metrics for medical applications of artificial intelligence. *Scientific reports* **12**:5979 https://doi.org/10.1038/s41598-022-09954-8 | PubMed

Hunt A., Law H., Carney R., Mulholland R., Flores A., Tudur Smith C., Varese F., Parker S., Yung A. R., Bonnett L. J. (2024) Systematic review of clinical prediction models for psychosis in individuals meeting at risk mental state criteria. *Frontiers in Psychiatry* **15**:1408738 https://doi.org/10.3389/fpsyt.2024.1408738 | PubMed

Huys Q. J., Maia T. V., Frank M. J. (2016) Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience* **19**:404-413 https://doi.org/10.1038/nn.4238 | PubMed

Ioannidis J. P. (2005) Why most published research findings are false. *PLoS medicine* **2**:e124 https://doi.org/10.1371/journal.pmed.0020124 | PubMed

Ioannidis J. P. (2008) Why most discovered true associations are inflated. *Epidemiology* **19**:640-648 https://doi.org/10.1097/ede.0b013e31818131e7 | PubMed

Ioannidis J. P., Panagiotou O. A. (2011) Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. *Jama* **305**:2200-2210 https://doi.org/10.1001/jama.2011.713 | PubMed

Jacobucci R., Grimm K. J. (2020) Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science* **15**:809-816 https://doi.org/10.1177/1745691620902467 | PubMed

Jahanshad N., Lenzini P., Bijsterbosch J. (2025) Current best practices and future opportunities for reproducible findings using large-scale neuroimaging in psychiatry. *Neuropsychopharmacology* **50**:37-51

Jansen W. J., Ossenkoppele R., Knol D. L., Tijms B. M., Scheltens P., Verhey F. R., Visser P. J., Aalten P., Aarsland D., Alcolea D., *et al.* (2015) Prevalence of cerebral amyloid pathology in persons without dementia: a meta-analysis. *Jama* **313**:1924-1938 https://doi.org/10.1001/jama.2015.4668 | PubMed

Jeni L. A., Cohn J. F., De La Torre F. (2013) Facing imbalanced data–recommendations for the use of performance metrics. In: 2013 Humaine association conference on affective computing and intelligent interaction. pp. 245-251 https://doi.org/10.1109/acii.2013.47 | PubMed

Jollans L., Whelan R. (2018) Neuromarkers for mental disorders: harnessing population neuroscience. *Frontiers in psychiatry* **9**:242 https://doi.org/10.3389/fpsyt.2018.00242 | PubMed

Kahneman D., Tversky A. (1973) On the psychology of prediction. *Psychological review* **80**:237 https://doi.org/10.1037/h0034747

Kambeitz-Ilankovic L., Koutsouleris N., Upthegrove R. (2022) The potential of precision psychiatry: what is in reach?. *The British Journal of Psychiatry* **220**:175-178 https://doi.org/10.1192/bjp.2022.23 | PubMed

Kapur S., Phillips A. G., Insel T. R. (2012) Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it?. *Molecular psychiatry* **17**:1174-1179 https://doi.org/10.1038/mp.2012.105 | PubMed

Karvelis P., Charlton C. E., Allohverdi S. G., Bedford P., Hauke D. J., Diaconescu A. O. (2022) Computational approaches to treatment response prediction in major depression using brain activity and behavioral data: A systematic review. *Network Neuroscience* 1-52 https://doi.org/10.1162/netn_a_00233 | PubMed

Karvelis P., Diaconescu A. O. (2025a) Clarifying the reliability paradox: Poor measurement reliability attenuates group differences. *Frontiers in Psychology* **16** https://doi.org/10.3389/fpsyg.2025.1592658 | PubMed

Karvelis P., Diaconescu A. O. (2025b) E2p simulator: An interactive tool for estimating real-world predictive utility of research findings. *Journal of Open Source Software* **10**:8334 https://doi.org/10.21105/joss.08334

Karvelis P., Paulus M. P., Diaconescu A. O. (2023) Individual differences in computational psychiatry: A review of current challenges. *Neuroscience & Biobehavioral Reviews* **148**:105137 https://doi.org/10.1016/j.neubiorev.2023.105137 | PubMed

Kessler R. C., Bossarte R. M., Luedtke A., Zaslavsky A. M., Zubizarreta J. R. (2020) Suicide prediction models: a critical review of recent research with recommendations for the way forward. *Molecular psychiatry* **25**:168-179 https://doi.org/10.1038/s41380-019-0531-0 | PubMed

Kirk H. (2019) Prediction versus management models relevant to risk assessment: The importance of legal decision-making context. *Clinical Forensic Psychology and Law* 347-359 https://doi.org/10.4324/9781351161565-3

Kirk R. E. (1996) Practical significance: A concept whose time has come. *Educational and psychological measurement* **56**:746-759 https://doi.org/10.1177/0013164496056005002

Kraemer H. C., Kupfer D. J. (2006) Size of treatment effects and their importance to clinical research and practice. *Biological psychiatry* **59**:990-996 https://doi.org/10.1016/j.biopsych.2005.09.014 | PubMed

Kühberger A., Fritz A., Lermer E., Scherndl T. (2015) The significance fallacy in inferential statistics. *BMC research notes* **8**:84 https://doi.org/10.1186/s13104-015-1020-4 | PubMed

Leeflang M. M., Rutjes A. W., Reitsma J. B., Hooft L., Bossuyt P. M. (2013) Variation of a test's sensitivity and specificity with disease prevalence. *Cmaj* **185**:E537–E544 https://doi.org/10.1503/cmaj.121286 | PubMed

Lionetti S., Gröger F., Gottfrois P., Gonzalez-Jimenez A., Amruthalingam L., Navarini A. A., Pouly M. (2025) Clinical uncertainty impacts machine learning evaluations. *arXiv* https://doi.org/10.48550/arxiv.2509.22242

Lo A., Chernoff H., Zheng T., Lo S.-H. (2015) Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences* **112**:13892-13897 https://doi.org/10.1073/pnas.1518285112 | PubMed

Loth E., Ahmad J., Chatham C., López B., Carter B., Crawley D., Oakley B., Hayward H., Cooke J., San Jose Caceres A., *et al.* (2021) The meaning of significant mean group differences for biomarker discovery. *PLoS computational biology* **17**:e1009477 https://doi.org/10.1371/journal.pcbi.1009477 | PubMed

Lovakov A., Agadullina E. R. (2021) Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology* **51**:485-504 https://doi.org/10.1002/ejsp.2752

MacCallum R. C., Zhang S., Preacher K. J., Rucker D. D. (2002) On the practice of di-chotomization of quantitative variables. *Psychological methods* **7**:19 https://doi.org/10.1037/1082-989x.7.1.19 | PubMed

Makowski C., Nichols T. E., Dale A. M. (2024) Quality over quantity: powering neuroimaging samples in psychiatry. *Neuropsychopharmacology* 1-9 https://doi.org/10.1038/s41386-024-01893-4 | PubMed

Marek S., Tervo-Clemmens B., Calabro F. J., Montez D. F., Kay B. P., Hatoum A. S., Donohue M. R., Foran W., Miller R. L., Hendrickson T. J., *et al.* (2022) Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**:654-660 https://doi.org/10.1038/s41586-022-04492-9 | PubMed

Maxim L. D., Niebo R., Utell M. J. (2014) Screening tests: a review with examples. *Inhalation toxicology* **26**:811-828 https://doi.org/10.3109/08958378.2014.955932 | PubMed

McGeechan K., Macaskill P., Irwig L., Liew G., Wong T. Y. (2008) Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. *Archives of Internal Medicine* **168**:2304-2310 https://doi.org/10.1001/archinte.168.21.2304 | PubMed

Meehan A. J., Lewis S. J., Fazel S., Fusar-Poli P., Steyerberg E. W., Stahl D., Danese A. (2022) Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Molecular psychiatry* **27**:2700-2708 https://doi.org/10.1038/s41380-022-01528-4 | PubMed

Meehl P. E. (1992) Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology.

Monsarrat P., Vergnes J.-N. (2018) The intriguing evolution of effect sizes in biomedical research over time: smaller but more often statistically significant. *GigaScience* **7**:gix121 https://doi.org/10.1093/gigascience/gix121 | PubMed

Moons K. G., Altman D. G., Reitsma J. B., Ioannidis J. P., Macaskill P., Steyerberg E. W., Vickers A. J., Ransohoff D. F., Collins G. S. (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Annals of internal medicine* **162**:W1-W73 https://doi.org/10.7326/m14-0698 | PubMed

Naggara O., Raymond J., Guilbert F., Roy D., Weill A., Altman D. G. (2011) Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. *American Journal of Neuroradiology* **32**:437-440 https://doi.org/10.3174/ajnr.a2425 | PubMed

Nakagawa S., Cuthill I. C. (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological reviews* **82**:591-605 https://doi.org/10.1111/j.1469-185x.2007.00027.x | PubMed

Nikolaidis A., Chen A. A., He X., Shinohara R., Vogelstein J., Milham M., Shou H. (2022) Suboptimal phenotypic reliability impedes reproducible human neuroscience. *bioRxiv* https://doi.org/10.1101/2022.07.22.501193

Onnela J.-P., Rauch S. L. (2016) Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* **41**:1691-1696 https://doi.org/10.1038/npp.2016.7 | PubMed

Ozenne B., Subtil F., Maucort-Boulch D. (2015) The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology* **68**:855-859 https://doi.org/10.1016/j.jclinepi.2015.02.010 | PubMed

Palminteri S., Wyart V., Koechlin E. (2017) The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences* **21**:425-433 https://doi.org/10.1016/j.tics.2017.03.011 | PubMed

Patzelt E. H., Hartley C. A., Gershman S. J. (2018) Computational phenotyping: using models to understand individual differences in personality, development, and mental illness. *Personality Neuroscience* **1**:e18 https://doi.org/10.1017/pen.2018.14 | PubMed

Paulus M. P., Thompson W. K. (2021) Computational approaches and machine learning for individual-level treatment predictions. *Psychopharmacology* **238**:1231-1239 https://doi.org/10.1007/s00213-019-05282-4 | PubMed

Pigoni A., Delvecchio G., Turtulici N., Madonna D., Pietrini P., Cecchetti L., Brambilla P. (2024) Machine learning and the prediction of suicide in psychiatric populations: a systematic review. *Translational psychiatry* **14**:140 https://doi.org/10.1038/s41398-024-02852-9 | PubMed

Pinker E. (2018) Reporting accuracy of rare event classifiers. *NPJ digital medicine* **1**:56 https://doi.org/10.1038/s41746-018-0062-0 | PubMed

Quinlivan L., Cooper J., Meehan D., Longson D., Potokar J., Hulme T., Marsden J., Brand F., Lange K., Riseborough E., *et al.* (2017) Predictive accuracy of risk scales following self-harm: multicentre, prospective cohort study. *The British Journal of Psychiatry* **210**:429-436 https://doi.org/10.1192/bjp.bp.116.189993 | PubMed

Ramspek C. L., Steyerberg E. W., Riley R. D., Rosendaal F. R., Dekkers O. M., Dekker F. W., van Diepen M. (2021) Prediction or causality? a scoping review of their conflation within current observational research. *European journal of epidemiology* **36**:889-898 https://doi.org/10.1007/s10654-021-00794-w | PubMed

Regier D. A., Narrow W. E., Clarke D. E., Kraemer H. C., Kuramoto S. J., Kuhl E. A., Kupfer D. J. (2013) Dsm-5 field trials in the united states and canada, part ii: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry* **170**:59-70 https://doi.org/10.1176/appi.ajp.2012.12070999 | PubMed

Roach B. J., Carrión R. E., Hamilton H. K., Bachman P., Belger A., Duncan E., Johannesen J., Light G. A., Niznikiewicz M., Addington J., *et al.* (2020) Reliability of mismatch negativity event-related potentials in a multisite, traveling subjects study. *Clinical neurophysiology* **131**:2899-2909 https://doi.org/10.1016/j.clinph.2020.09.027 | PubMed

Rosen M., Betz L. T., Schultze-Lutter F., Chisholm K., Haidl T. K., Kambeitz-Ilankovic L., Bertolino A., Borgwardt S., Brambilla P., Lencer R., *et al.* (2021) Towards clinical application of prediction models for transition to psychosis: a systematic review and external validation study in the pronia sample. *Neuroscience & Biobehavioral Reviews* **125**:478-492 https://doi.org/10.1016/j.neubiorev.2021.02.032 | PubMed

Ross E. L., Zuromski K. L., Reis B. Y., Nock M. K., Kessler R. C., Smoller J. W. (2021) Accuracy requirements for cost-effective suicide risk prediction among primary care patients in the us. *JAMA psychiatry* **78**:642-650 https://doi.org/10.1001/jamapsychiatry.2021.0089 | PubMed

Rothman K. J., Greenland S. (2018) Planning study size based on precision rather than power. *Epidemiology* **29**:599-603 https://doi.org/10.1097/ede.0000000000000876 | PubMed

Rousson V., Zumbrunn T. (2011) Decision curve analysis revisited: overall net benefit, relationships to roc curve analysis, and application to case-control studies. *BMC medical informatics and decision making* **11**:45 https://doi.org/10.1186/1472-6947-11-45 | PubMed

Royston P., Altman D. G., Sauerbrei W. (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in medicine* **25**:127-141 https://doi.org/10.1002/sim.2331 | PubMed

Rutledge R. B., Chekroud A. M., Huys Q. J. (2019) Machine learning and big data in psychiatry: toward clinical applications. *Current opinion in neurobiology* **55**:152-159 https://doi.org/10.1016/j.conb.2019.02.006 | PubMed

Rutledge T., Loh C. (2004) Effect sizes and statistical testing in the determination of clinical significance in behavioral medicine research. *Annals of Behavioral Medicine* **27**:138-145 https://doi.org/10.1207/s15324796abm2702_9 | PubMed

Saito T., Rehmsmeier M. (2015) The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10**:e0118432 https://doi.org/10.1371/journal.pone.0118432 | PubMed

Salazar de Pablo G., Studerus E., Vaquerizo-Serrano J., Irving J., Catalan A., Oliver D., Baldwin H., Danese A., Fazel S., Steyerberg E. W., *et al.* (2021) Implementing precision psychiatry: a systematic review of individualized prediction models for clinical practice. *Schizophrenia bulletin* **47**:284-297 https://doi.org/10.26041/fhnw-10896

Schafer T., Schwarz M. A. (2019) The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in psychology* **10**:813 https://doi.org/10.3389/fpsyg.2019.00813 | PubMed

Schnack H. G., Kahn R. S. (2016) Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. *Frontiers in psychiatry* **7**:50 https://doi.org/10.3389/fpsyt.2016.00050 | PubMed

Shmueli G. (2010) To explain or to predict?. *Statistical science* 289-310 https://doi.org/10.1214/10-sts330

Shorey S., Ng E. D., Wong C. H. (2022) Global prevalence of depression and elevated depressive symptoms among adolescents: A systematic review and meta-analysis. *British Journal of Clinical Psychology* **61**:287-305 https://doi.org/10.1111/bjc.12333 | PubMed

Singh I., Rose N. (2009) Biomarkers in psychiatry. *Nature* **460**:202-207 https://doi.org/10.1038/460202a | PubMed

Spearman C. (1904) The proof and measurement of association between two things. *The American Journal of Psychology* **15**:72-101 https://doi.org/10.2307/1412159

Steeg S., Quinlivan L., Nowland R., Carroll R., Casey D., Clements C., Cooper J., Davies L., Knipe D., Ness J., *et al.* (2018) Accuracy of risk scales for predicting repeat self-harm and suicide: a multicentre, population-level cohort study using routine clinical data. *BMC psychiatry* **18**:113 https://doi.org/10.1186/s12888-018-1693-z | PubMed

Stephan K. E., Mathys C. (2014) Computational approaches to psychiatry. *Current opinion in neurobiology* **25**:85-92 https://doi.org/10.1016/j.conb.2013.12.007 | PubMed

Stephan K. E., Penny W. D., Daunizeau J., Moran R. J., Friston K. J. (2009) Bayesian model selection for group studies. *Neuroimage* **46**:1004-1017 https://doi.org/10.1016/j.neuroimage.2009.03.025 | PubMed

Stone M. B., Yaseen Z. S., Miller B. J., Richardville K., Kalaria S. N., Kirsch I. (2022) Response to acute monotherapy for major depressive disorder in randomized, placebo controlled trials submitted to the us food and drug administration: individual participant data analysis. *Bmj* 378 https://doi.org/10.1136/bmj-2021-067606 | PubMed

Streiner D. L. (2002) Breaking up is hard to do: the heartbreak of dichotomizing continuous data. *The Canadian Journal of Psychiatry* **47**:262-266 https://doi.org/10.1177/070674370204700307 | PubMed

Szucs D., Ioannidis J. P. (2017) When null hypothesis significance testing is unsuitable for research: a reassessment. *Frontiers in human neuroscience* **11**:390 https://doi.org/10.3389/fnhum.2017.00390 | PubMed

Talukder A., Kougianou I., Healy C., Lång U., Kieseppä V., Jalbrzikowski M., O'Hare K., Kelleher I. (2025) Sensitivity of the clinical high-risk and familial high-risk approaches for psychotic disorders–a systematic review and meta-analysis. *Psychological Medicine* **55**:e46 https://doi.org/10.1017/s0033291724003520 | PubMed

Thiruvalluru R. K., Edgcomb J. B., Brooks J. O., Pathak J. (2023) Risk of suicide attempts and self-harm after 1.4 million general medical hospitalizations of men with mental illness. *Journal of psychiatric research* **157**:50-56 https://doi.org/10.1016/j.jpsychires.2022.10.035 | PubMed

Thompson B. (2007) Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools* **44**:423-432 https://doi.org/10.1002/pits.20234

Tornero-Costa R., Martinez-Millana A., Azzopardi-Muscat N., Lazeri L., Traver V., Novillo-Ortiz D., et al. (2023) Methodological and quality flaws in the use of artificial intelligence in mental health research: systematic review. *JMIR Mental Health* **10**:e42045 https://doi.org/10.2196/42045 | PubMed

Tozzi L., Goldstein-Piekarski A. N., Korgaonkar M. S., Williams L. M. (2020) Connectivity of the cognitive control network during response inhibition as a predictive and response biomarker in major depression: evidence from a randomized clinical trial. *Biological psychiatry* **87**:462-472 https://doi.org/10.1016/j.biopsych.2019.08.005 | PubMed

Trajković G., Starčević V., Latas M., Leštarević M., Ille T., Bukumirić Z., Marinković J. (2011) Reliability of the hamilton rating scale for depression: a meta-analysis over a period of 49 years. *Psychiatry research* **189**:1-9 https://doi.org/10.1016/j.psychres.2010.12.007 | PubMed

Tversky A., Kahneman D. (1974) Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* **185**:1124-1131 https://doi.org/10.1126/science.185.4157.1124

Van Calster B., Collins G. S., Vickers A. J., Wynants L., Kerr K. F., Barreñada L., Varoquaux G., Singh K., Moons K. G., Hernandez-Boussard T., *et al.* (2025) Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. *The Lancet Digital Health* 100916 https://doi.org/10.1016/j.landig.2025.100916 | PubMed

Van Calster B., McLernon D. J., Van Smeden M., Wynants L., Steyerberg E. W. (2019) diagnostic tests, T. G. and prediction models' of the STRATOS initiative Bossuyt Patrick Collins Gary S. Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. *BMC medicine* **17**:230

Van Calster B., van Smeden M., van Amsterdam W., Coemans M., Wynants L., Steyerberg E. W. (2025) The enemies of reliable and useful clinical prediction models: A review of statistical and scientific challenges. *Annual Review of Statistics and Its Application* 13

Van Dellen E. (2024) Precision psychiatry: predicting predictability. *Psychological medicine* **54**:1500-1509 https://doi.org/10.1017/s0033291724000370 | PubMed

Vickers A. J., Elkin E. B. (2006) Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* **26**:565-574 https://doi.org/10.1177/0272989x06295361 | PubMed

Vidal-Piñeiro D., Sørensen Ø., Strømstad M., Amlien I. K., Anderson M., Baaré W. F., Bartrés-Faz D., Brândmaier A. M., Brathen A. C., Garrido P., *et al.* (2025) Reliability of structural brain change in cognitively healthy adult samples. *Imaging Neuroscience* **3**:imag_a_00547 https://doi.org/10.1162/imag_a_00547 | PubMed

Vieira S., Liang X., Guiomar R., Mechelli A. (2022) Can we predict who will benefit from cognitive-behavioural therapy? a systematic review and meta-analysis of machine learning studies. *Clinical Psychology Review* **97**:102193 https://doi.org/10.1016/j.cpr.2022.102193 | PubMed

Wang J., Huang S., Lan G., Lai Y.-J., Wang Q.-H., Chen Y., Xiao Z.-S., Chen X., Bu X.-L., Liu Y.-H., *et al.* (2025) Diagnostic accuracy of plasma p-tau217/aβ42 for alzheimer's disease in clinical and community cohorts. *Alzheimer's & Dementia* **21**:e70038 https://doi.org/10.1002/alz.70038 | PubMed

Wasserstein R. L., Lazar N. A. (2016) The asa statement on p-values: context, process, and purpose. *The American Statistician* **70**:129-133 https://doi.org/10.1080/00031305.2016.1154108

Wasserstein R. L., Schirm A. L., Lazar N. A. (2019) Moving to a world beyond "p < 0.05". *The American Statistician* **73**:1-19 https://doi.org/10.1080/00031305.2019.1583913

Weber M. A., Schnyder N., Kirschstein M. A., Graf M., Endrass J., Rossegger A. (2025) The key role of base rates: systematic review and meta-analysis of the predictive value of four risk assessment instruments. *Swiss Medical Weekly* **155**:3517-3517 https://doi.org/10.57187/s.3517 | PubMed

Weiss-Cowie S., Verhaeghen P., Duarte A. (2023) An updated account of overgeneral autobiographical memory in depression. *Neuroscience & Biobehavioral Reviews* **149**:105157 https://doi.org/10.1016/j.neubiorev.2023.105157 | PubMed

Whittle R., Peat G., Belcher J., Collins G. S., Riley R. D. (2018) Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported. *Journal of clinical epidemiology* **102**:38-49 https://doi.org/10.1016/j.jclinepi.2018.05.008 | PubMed

Wilson R. C., Collins A. G. (2019) Ten simple rules for the computational modeling of behavioral data. *eLife* **8**:e49547 https://doi.org/10.7554/eLife.49547 | PubMed

Yarkoni T. (2022) The generalizability crisis. *Behavioral and Brain Sciences* 45 https://doi.org/10.1017/s0140525x20001685 | PubMed

Christodoulou E., Ma J., Collins G. S., Steyerberg E. W., Verbakel J. Y., Van Calster B. (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology* **110**:12-22 https://doi.org/10.1016/j.jclinepi.2019.02.004 | PubMed

Del Giudice M. (2009) On the real magnitude of psychological sex differences. *Evolutionary Psychology* **7**:147470490900700209 https://doi.org/10.1177/147470490900700209

Hill W. G. (2010) Understanding and using quantitative genetic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**:73-85 https://doi.org/10.1098/rstb.2009.0203 | PubMed

Mahalanobis P. C. (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India* **2**:49-55

Peer reviews

Reviewer #1 (Public review):

Summary:

The manuscript provides a well-argued discussion of the misalignment between common predictive performance evaluations reported in the literature and actually measuring clinical utility in the context of predictive psychiatry. Specifically, the authors discuss measurement reliability and prevalence as two neglected factors which can substantially inflate the assessment of model performance for clinical practice. To mitigate this, the authors offer a concrete framework and an accompanying web tool, with which to adjust performance metrics and additional predictive-value and decision-analytic measures.

Strengths:

The manuscript speaks convincingly about the risk of face validity and the practical irrelevance of seemingly promising predictive models in psychiatry. The authors outline how predictive performance estimations often fail to generalize to clinical contexts and thereby potentially mislead scientific efforts. In the face of ubiquitous biomarker models and incremental improvements in the literature, the reader is reminded that, irrespective of the glory of the proposed model, low reliability of clinical measurements fundamentally affects (and limits) both effect sizes and predictive performance ("garbage in, garbage out"), and that neglecting this can ultimately lead to misinformed decisions in the treatment of individual patients. The provision of an online tool with a user-friendly interface and clearly worked

examples is a major practical asset that will facilitate the adoption of the proposed framework beyond quantitative methodologists.

Weaknesses:

While the outlined issues highlight important aspects in the translational gap, the suggested solutions remain somewhat theoretical. For example, the use of prevalence might not reflect what a model would see in practice, assuming that population prevalence and the composition of actual clinical cohorts are aligned. Accounting for who presents to care, and under which referral or triage patterns, is a crucial determinant of effective base rates. While the authors do acknowledge the importance of using base rates from the target population, these nuances could be emphasized more prominently at the points where practical recommendations are made. Relatedly, the analytical context and the methodological assumptions are not clearly specified. Many arguments and demonstrations are derived in univariate, group-comparison settings and then discussed in a way that can be read as broadly applicable.

https://doi.org/10.7554/eLife.110646.1.sa2

Reviewer #2 (Public review):

Summary and strengths:

The authors present a description of their online tool to estimate real-world performance of predictive models. The authors bring together different calculations to make better-informed implementation choices. It is a very nice tool to go from effect sizes to base rates to decision curve analysis. The paper describes the background and use of the tool with examples and seems like an extended version of their online how-to. The methods themselves are not new, but I think the tool will be valuable for researchers from different fields. Tools already exist for the conversion of effect sizes (my current favorite is https://www.escal.site/), but I haven't seen measurement noise being incorporated previously. The main benefit is the evaluation of performance under different real-world scenarios. Code is available on GitHub, and the manuscript is well-written.

Weaknesses:

While comprehensive explanation and examples are important for correct use of the tool, I don't really see the added value above their online how-to guide, as the software itself has already been published (Karvelis, P. and Diaconescu, A. O. (2025b). E2p simulator: An interactive tool for estimating real world predictive utility of research findings. Journal of Open Source Software, 10(114):8334.)

https://doi.org/10.7554/eLife.110646.1.sa1

Reviewer #3 (Public review):

Summary:

This important work provides a web-based tool to contextualize effect sizes in psychiatry with respect to reliability and base rates (collectively referred to as predictive utility analysis). The methods for the tool incorporate established psychometric principles that I think are of use for multiple fields in this seemingly easy-to-use tool. I agree with the critical importance of this tool and the methodological points made in this manuscript. Enthusiasm for the manuscript is weakened by a lack of clarity on the formulation of the paper and stated goals of the examples used, with the inferences and impact on clinical decision making from various parameterizations via this tool left open-ended.

Strengths:

This paper presents a well-considered and, what I think will be highly useful, web-based tool to contextualize effect sizes with respect to reliability and base rates. As the authors rightly point out, such a tool could be used in conjunction with widespread analytic power analysis tools in study planning. The paper also well contexualizes the need for such a tool in the relatively recent history of concerns of power, reliability, and inference in psychiatry specifically, and more general meta-scientific debates in psychology and neuroscience.

Weaknesses:

My primary feedback on this manuscript is the lack of clarity in what the paper itself, specifically, separate from the tool, is hoping to achieve. There is a central, but unresolved, tension in whether the reader is supposed to:

(1) focus on the specifics of the examples used and whether to reevaluate the substantive claims from the studies, (2) buy in to how various reliability and base rate parameters impact modeling outcomes, (3) receive an introduction to the tool itself.

In my estimation, the largest contribution to the field here is in (2) and (3), but currently much of the real estate of the paper is dedicated to several examples of (1). While these specific examples may be illustrative to some degree, I think given the number and brevity of such, they are unlikely to incidentally achieve points (2) and (3) above. Specific examples include the assertion of kappas for DSM diagnoses, without much nuance (e.g., see *https://psycnet.apa.org/buy/2015-27500-001*). Given the relatively limited space given to this example, however, it's hard to be entirely certain what the reviewer should take away.

A second point of concern is where this tool would be situated in the research pipeline. I agree with the authors that this tool could be used in ways that parallel power analysis. With that in mind, it seems the most common use of this tool for an individual investigator is likely to be in a priori study planning. In contrast, and with my point above in mind, the use of the tool for existing results is likely best done with multiple estimates of effect sizes, reliability, and base rates, as is common in meta-analysis or consensus reviews. Nevertheless, there is no real example or guidance around how this influences new study planning.

A third point is that more nuance would be useful in the introduction about the current state of psychiatry research. For example, I share many of the authors' concerns about reliability, power, reproducibility, and barriers to translation. That said, it is the case that while effect sizes should be considered considerably more, they are widely considered in psychiatry research via the common place of meta-analysis and other data pooling approaches. Another such example that the authors state in the context of reliability: "However, this [reliability] attenuation is rarely accounted for in routine analyses in psychiatry". This is true in practice, but somewhat misleading insofar as the method by which to do this remains unclear. For example, should we all report disattenuated associations, assuming there is no error and everything is perfectly reliable? This, of course, would be unrealistic to expect zero error. That we can achieve this with the new tool is clear, but the nuance of how and under what circumstances it should be done is not clear, and such nuance should be better reflected in the framing of the problem. That is, there is also a lack of clarity on what ought to be best practices and field-wide goals, rather than simply the lack of an ability to model these factors.

Minor point

For conceptual clarity, it would benefit the manuscript to at least briefly mention the role of validity in translational importance. Of course, the current psychometric issues of reliability, base rate, power, etc are critical, but it should at least be mentioned, given the potential wide audience of this manuscript, validity is important as well. For example, highly reliable

measures may not be valid indicators of underlying disease etiology (e.g., fMRI head motion is a highly reliable trait-level feature, but typically not considered an important predictor or consequence of mental health worth investing translational resources in). Relatedly, confounding as a general topic would be useful to mention just briefly, to help with the spirit of considering underlying issues in translation.

https://doi.org/10.7554/eLife.110646.1.sa0