The Journal of Open Source Software

E2P Simulator: An Interactive Tool for Estimating Real-World Predictive Utility of Research Findings

Povilas Karvelis [1] **and Andreea O. Diaconescu** [1,2,3,4]

1 Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Toronto, ON, Canada **2** Department of Psychiatry, University of Toronto, Toronto, ON, Canada **3** Institute of Medical Sciences, University of Toronto, Toronto, ON, Canada **4** Department of Psychology, University of Toronto, Toronto, ON, Canada

DOI: 10.21105/joss.08334

Software

- Review
- Repository
- Archive

Editor: Julia Romanowska

Reviewers:

- @Annchovy
- @blackdotbug

Submitted: 05 April 2025
Published: 23 October 2025

Summary

E2P Simulator (Effect-to-Prediction Simulator) allows researchers to interactively and quantitatively explore the relationship between effect sizes (e.g., Cohen's d, Odds Ratio, Pearson's r), their discriminative ability (e.g., ROC-AUC, Sensitivity, Specificity, etc.), and real-world predictive value and clinical utility (e.g., PPV, NPV, PR-AUC, Net Benefit, etc.), while accounting for measurement reliability and outcome base rates (Figure 1).

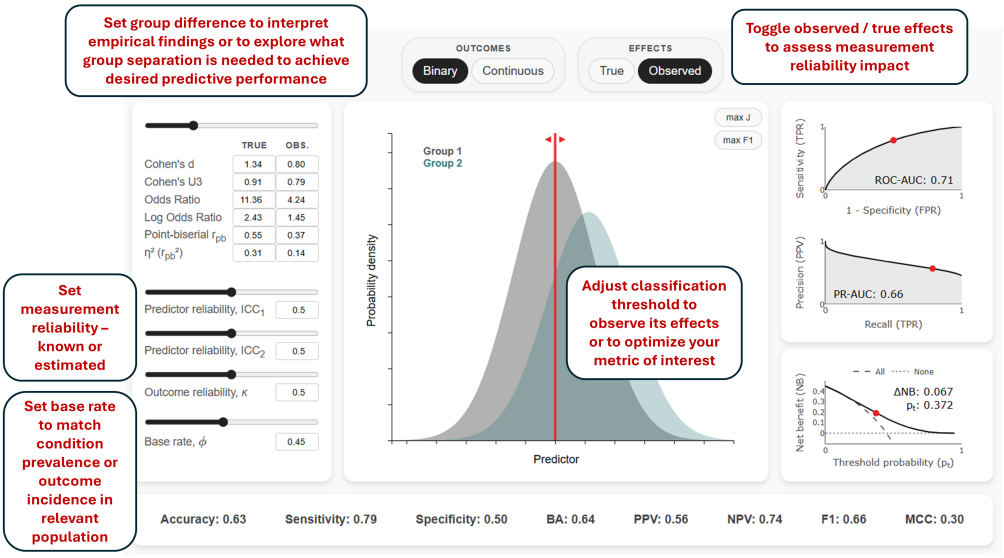

Figure 1: E2P Simulator interface with a high-level summary of its use. The left panel allows users to adjust parameters such as effect size, reliability, and base rate. The middle panel displays the resulting distributions with the adjustable threshold. The panel on the right displays Receiver Operating Characersitic (ROC) and Precision-Recall (PR) curves with corresponding Area Under the Curve (AUC) metrics, as well as Decision Curve Analysis (DCA) plot. The panel at the bottom displays the most common predictive metrics.

E2P Simulator has several potential applications:

1. **Interpretation of findings**: It helps researchers move beyond arbitrary "small/medium/large" effect size labels and misleading predictive metrics by grounding their interpretation in estimated real-world predictive utility.

2. **Research planning**: Being able to easily derive what effect sizes and predictive performance are needed to achieve a desired real-world predictive performance allows researchers to plan their studies more effectively and allocate resources more efficiently.
3. **Education**: The simulator's interactive design makes it a valuable teaching tool, helping researchers develop a more intuitive understanding of how different abstract statistical metrics relate to one another and to real-world utility.

This tool has been designed with biomedical and behavioral sciences in mind, particularly areas such as biomarker research, precision medicine, epidemiology, and biostatistics. However, it may be just as useful for any area of research that can is focused on predictive modelling and personalization, such as within forensic, education, and sports sciences.

Statement of Need

In biomedical and behavioral sciences, the predominant focus on statistical significance, which reflects the likelihood that an observed effect is not due to chance, often comes at the expense of sufficient attention to effect sizes, which quantify the practical significance of the effect (Wasserstein & Lazar, 2016). This emphasis, combined with a methodological disconnect between classical statistics and predictive modeling, frequently leads researchers to misinterpret any statistically significant finding as clinically meaningful, regardless of its effect size (Funder & Ozer, 2019; Wasserstein et al., 2019). This misinterpretation is particularly problematic in areas such as biomarker research, precision medicine, and precision psychiatry, where the goal is to find robust predictors of disease state or treatment response for individual patients (Abi-Dargham et al., 2023; Monsarrat & Vergnes, 2018). Furthermore, factors such as measurement reliability, which attenuates effect sizes (Karvelis et al., 2023; Karvelis & Diaconescu, 2025), and outcome base rates, which limit predictive power in real-world contexts (Abi-Dargham & Horga, 2016; Baldessarini et al., 1983; Brabec et al., 2020; Large, 2018; Ozenne et al., 2015), are often overlooked in evaluating both individual predictors and predictive models, leading to unrealistic expectations, ineffective research planning, and resource misalocation.

E2P Simulator addressess these challenges by providing an interactive platform where researchers can explore the relationships among all of these factors. Similar to how GPower (Faul et al., 2007) is used to explore the relationships between effect size, sample size and significance levels to perform *power analysis*, E2P Simulator can be used to explore the relationships between effect size, predictive performance, and real-world predictive utility to perform *predictive utility analysis*. By making the relationships between effect sizes, reliability, base rates, and predictive metrics explicit, E2P Simulator enables researchers to interpret findings more accurately, design more impactful studies, and communicate results more clearly to broader audiences.

Implementation

E2P Simulator is implemented using HTML, CSS, and JavaScript. The tool leverages several open-source libraries:

- D3.js for data visualization (Bostock et al., 2011)
- Plotly.js for interactive plots (Inc., 2015)
- Chart.js for additional charting capabilities (Contributors, 2013)
- MathJax.js for rendering mathematical expressions (Consortium, 2009)

The application is designed to be accessible without installation, running entirely on the web (www.e2p-simulator.com). This implementation ensures broad accessibility across different operating systems and devices. Alternatively, the tool can also be run on a local node. The instructions and examples of usage are included in the tool on the Get Started page.

Acknowledgements

AOD is supported by the Canadian Institutes of Health Research and the Krembil Foundation.

References

Abi-Dargham, A., & Horga, G. (2016). The search for imaging biomarkers in psychiatric disorders. *Nature Medicine*, *22*(11), 1248–1255. https://doi.org/10.1038/nm.4190

Abi-Dargham, A., Moeller, S. J., Ali, F., DeLorenzo, C., Domschke, K., Horga, G., Jutla, A., Kotov, R., Paulus, M. P., Rubio, J. M., & others. (2023). Candidate biomarkers in psychiatric disorders: State of the field. *World Psychiatry*, *22*(2), 236–262. https://doi.org/10.1002/wps.21078

Baldessarini, R. J., Finklestein, S., & Arana, G. W. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry*, *40*(5), 569–573. https://doi.org/10.1001/archpsyc.1983.01790050095011

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309.

Brabec, J., Komárek, T., Franc, V., & Machlica, L. (2020). On model evaluation under non-constant class imbalance. *International Conference on Computational Science*, 74–87. https://doi.org/10.1007/978-3-030-50423-6_6

Consortium, M. (2009). *MathJax: Beautiful math in all browsers*. https://www.mathjax.org/

Contributors, Chart. js. (2013). *Chart.js: Simple yet flexible JavaScript charting for designers & developers*. https://www.chartjs.org/

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

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*(2), 156–168. https://doi.org/10.1177/2515245919847202

Inc., P. T. (2015). *Plotly JavaScript open source graphing library*. https://plotly.com/javascript/

Karvelis, P., & Diaconescu, A. O. (2025). Clarifying the reliability paradox: Poor measurement reliability attenuates group differences. *Frontiers in Psychology*, *16*(1592658). https://doi.org/10.3389/fpsyg.2025.1592658

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/10.31234/osf.io/bvjzn

Large, M. M. (2018). The role of prediction in suicide prevention. *Dialogues in Clinical Neuroscience*, *20*(3), 197–205. https://doi.org/10.31887/DCNS.2018.20.3/mlarge

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*(1), gix121. https://doi.org/10.1093/gigascience/gix121

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*(8), 855–859. https://doi.org/10.1016/j.jclinepi.2015.02.010

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process,

and purpose. In *The American Statistician* (No. 2; Vol. 70, pp. 129–133). Taylor & Francis. 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". In *The American Statistician* (No. sup1; Vol. 73, pp. 1–19). Taylor & Francis. https://doi.org/10.1080/00031305.2019.1583913