Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Output-Distribution Divergence as a Pre-Interpretation Gate for Mental Health AI

Saurabh Anand[†,*]
[†] Independent Researcher

Abstract

Machine learning models in mental health are widely used to generate risk scores from observational data, yet their outputs are frequently interpreted in causal or intervention-oriented terms without explicit checks on whether such interpretations transport across demographic contexts. We propose a simple governance-oriented diagnostic that compares predicted-probability distributions across contexts against within-context baselines using standard divergence measures, functioning as a pre-interpretation gate rather than a causal estimator. We operationalize this protocol using Jensen–Shannon divergence and Wasserstein distance, calibrated via bootstrapped intra-context baselines, and evaluate it on depression risk prediction using PHQ-9 data from the National Health and Nutrition Examination Survey (NHANES) 2017–2020. Across age and sex contexts, we find that cross-context divergence consistently exceeds baseline variation, particularly for age-group transfers, even when discrimination metrics such as AUC remain stable. These results demonstrate that performance-based validation alone can mask substantial distributional instability in predicted probabilities, with implications for calibration and interpretability. We argue that output-distribution divergence provides a low-cost, model-agnostic diagnostic for identifying transportability risk prior to deploying or interpreting mental health prediction models in intervention-relevant settings.

Keywords: mental health AI, distributional shift, transportability, governance, divergence metrics, depression screening, responsible AI, calibration drift

1. Introduction

Mental health machine learning systems are increasingly used to produce risk scores, severity estimates, or prioritized factors from observational data such as surveys, electronic health records, or digital traces. While these models are typically trained and validated for predictive accuracy, they are often consumed in intervention-flavored language—e.g., "target sleep to reduce depression," or "improving social support will lower relapse risk." This interpretive slide from *what the model predicts* to *what should be done* reflects a deeper tension: prediction answers what tends to happen, whereas intervention requires understanding *why* it happens. In high-stakes mental health settings, this distinction matters acutely, because models that appear reliable under one population, time period, or measurement regime may encode fragile associations that do not generalize. When such fragility is ignored, even well-performing predictive models can motivate actions whose causal justification is far weaker than their apparent confidence suggests. For example, Chekroud et al. [1] used machine learning on observational clinical trial data to predict antidepressant remission and proposed "matching patients to interventions"—framing that implies treatment selection capability from a purely predictive model, without explicit verification that the learned associations transport across the demographic contexts in which such a model would be deployed.

This problem can be clarified through Pearl's causality ladder [2], which distinguishes associational reasoning ("seeing") from interventional ("doing") and counterfactual ("imagining") reasoning. Most mental health prediction models operate firmly at the associational level, yet their outputs are frequently interpreted as if they supported intervention or

* anandsaurabh17@gmail.com

policy decisions. This represents an implicit and often unjustified ascent up the ladder—particularly in mental health, where randomized experimentation is constrained and confounding is pervasive.

Our position is not that divergence measures establish causality, nor that they move models upward on the ladder. Rather, we argue that instability in a model's output distributions across contexts should serve as a governance-oriented warning signal: if even associational behavior is not stable, then causal or intervention-style interpretation is unlikely to be transportable. We operationalize this idea as a simple diagnostic protocol that compares predicted-probability distributions across populations using standard divergence measures (Jensen–Shannon divergence and Wasserstein distance), and flags unusually large divergence relative to within-context baselines as a signal of interpretive and deployment risk.

2. Background

Prediction vs. causation. The distinction between prediction and causal inference is well-established [2, 3], yet consistently blurred in applied mental health research. Predictive models answer what *tends to co-occur* with an outcome; causal models require understanding what *would happen* under intervention. When confounding or distributional shift underlies an association, intervening on the identified variable can be ineffective or harmful [4]. In mental health, where randomized experiments are ethically constrained and confounding is pervasive, this gap is particularly wide [1].

Transportability. A model's validity depends not only on within-distribution performance but on whether predictions *transport* to new populations or measurement contexts. The transportability literature [5, 6] formalises conditions under which conclusions generalise. In practice, transportability failures manifest as distributional shift in model outputs across contexts.

Shift monitoring and divergence. Distribution shift detection [7], covariate shift correction [8], and calibration assessment [9] are established ML practices. We do not claim novelty in detecting distributional shift itself. However, while prior work typically applies divergence measures to *input* distributions or *label* distributions [10], our focus on the full *predicted-probability* distribution—with baseline-normalized flagging framed as an epistemic governance gate—is underexplored in high-stakes clinical domains. Jensen–Shannon divergence (JSD) and Wasserstein distance capture complementary aspects of instability: JSD is sensitive to support differences and mode shifts, while Wasserstein responds to translational displacement even when distributional shape is preserved [10].

3. Protocol

To illustrate, consider a depression model trained on older adults (55+) and applied to younger adults (18–34). Even with similar depression rates, the model may produce a qualitatively different probability distribution. Our protocol asks: *does cross-context divergence exceed the variation expected from resampling within the training context?*

Formally, given a classifier f producing $\hat{p} = f(x) \in [0, 1]$ and context labels $C = \{c_1, \ldots, c_k\}$:

Step 1: Intra-context baseline. For each context c_i, perform B=200 stratified random half-splits. Train f on one half, predict on both, compute divergence between the two prediction distributions. This yields a baseline distribution $\mathcal{D}_i$ characterising normal within-context variation.

Step 2: Cross-context divergence. For each *ordered* pair $(c_i \to c_j)$: train f on c_i with probability calibration, predict on both c_i and c_j, compute divergence $d_{i \to j}$. Pairs are

directional because training context determines the learned associations; training on c_i and testing on c_j may yield different instability than the reverse.

Step 3: Comparison and flagging. Compare $d_{i \to j}$ against $\mathcal{D}_i$ via z-score and empirical percentile. Pairs where divergence exceeds the 95th percentile of the baseline distribution are flagged as transportability risks.

We compute JSD using 50 fixed bins over $[0, 1]$ with epsilon smoothing (10^{-10}) and Wasserstein distance on raw probability samples. AUC, Brier score, and Expected Calibration Error (ECE) are reported alongside divergence to distinguish discrimination from calibration degradation.

A governance flag does *not* indicate a model is wrong. It indicates that output distributions are unstable beyond normal stochastic variation—a necessary (though not sufficient) condition for transportability failure.

4. Experiments

4.1. Data

We use NHANES 2017–March 2020 [11], merging the Depression Screener (DPQ; PHQ-9) and Demographics (DEMO) on respondent sequence number. The PHQ-9 is a validated 9-item depression screener [12] widely used in clinical and epidemiological settings.

Outcome: PHQ-9 total ≥ 10 (depression screen positive). Items coded 0–3; values 7 (refused) or 9 (don't know) treated as missing; respondents with incomplete items excluded.

Sample: N=8,276 adults ($\geq$18), depression prevalence 9.3% (770 cases).

Contexts: Age group (18–34: n=2,218; 35–54: n=2,469; 55+: n=3,589) and sex (Female: n=4,205, prevalence 11.3%; Male: n=4,071, prevalence 7.2%).

Models. We use L2-regularised logistic regression (C=1.0), Platt-calibrated. Logistic regression is chosen deliberately as a conservative test: if distributional instability is detectable with a simple, well-calibrated linear model, it would likely be at least as pronounced with more complex classifiers. Two variants are evaluated: *(A) Demographics Only*—age, education, poverty-income ratio, sex, race/ethnicity, marital status (AUC $\approx$ 0.67–0.71); *(B) Items + Demographics*—adds 9 PHQ-9 item responses (AUC $\approx$ 1.0). Model A is primary; Model B confirms divergence persists even with near-perfect discrimination.

4.2. Results: Age Contexts

Table 1 presents cross-context results for the Demographics-Only model. All six directed age-group pairs were flagged: JSD exceeded the 95th percentile of the intra-context baseline in every case.

Train → Test	**JSD**	**JSD z**	**Wass**	**ΔAUC**	**ECE$_{\text{cross}}$**
18–34 → 35–54	.083	47.8	.029	−.002	.036
18–34 → 55+	.183	108.7	.044	+.033	.048
35–54 → 18–34	.021	13.2	.012	+.064	.005
35–54 → 55+	.043	30.3	.039	+.062	.042
55+ → 18–34	.254	297.4	.121	+.053	.126
55+ → 35–54	.048	53.9	.037	−.010	.034

Table 1. Demographics-only model, age-group contexts. JSD z: z-score relative to intra-context baseline (note: extreme values reflect very tight baselines with SD $\approx$ 0.001; raw JSD is more interpretable). ΔAUC: within minus cross AUC (positive = degradation). ECE$_{\text{cross}}$: cross-context expected calibration error (baseline ECE $\approx$ 0.004–0.013). All pairs exceed the 95th percentile of baseline JSD.

The highest instability occurred for 55+ → 18–34 (JSD = 0.254, Wasserstein = 0.121), with calibration degradation from ECE 0.004 to 0.126—a 30× increase.

Critically, the 55+ → 35–54 pair *improved* on AUC (−0.010 drop) yet exhibited JSD 21× the baseline mean and ECE rising from 0.004 to 0.034. This illustrates a key finding: **distributional instability and discrimination degradation are dissociable**. A model can maintain rank-ordering while systematically miscalibrating probabilities—arguably more dangerous for governance, because AUC-based validation alone would not detect it.

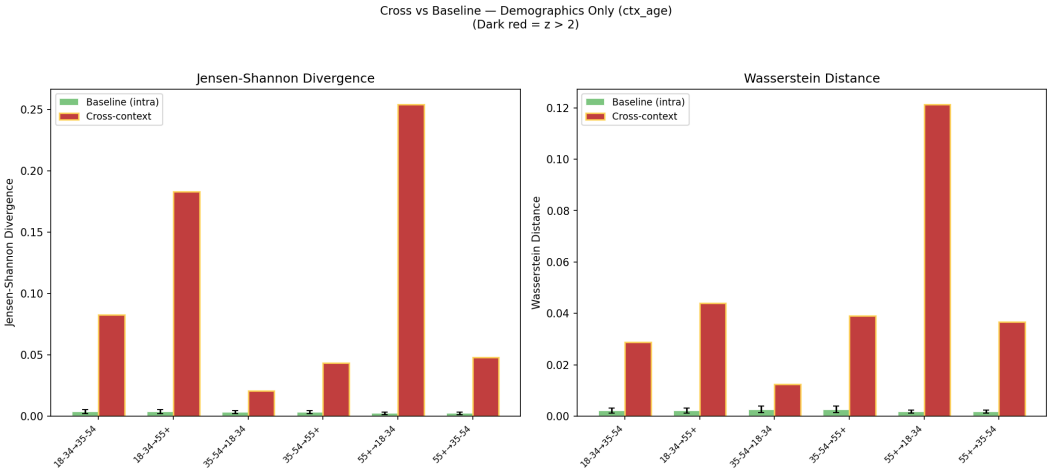

Figure 1. Cross-context JSD and Wasserstein versus intra-context baseline (mean ± SD) for the demographics-only model across age groups. Dark red bars (z-score > 2) denote governance-flagged pairs. Every cross-context pair exceeds the baseline distribution, with the 55+→18–34 pair showing divergence two orders of magnitude above baseline.

4.3. **Results: Sex Contexts**

Sex-based analysis revealed asymmetric instability (Table 2). Male → Female was flagged on JSD (percentile = 98%; Brier shifted from 0.065 to 0.099), while Female → Male was not flagged on JSD (percentile = 48%) but was flagged on Wasserstein (percentile = 100%). This metric disagreement reflects their complementary sensitivity profiles: Wasserstein detects the translational shift in predicted probabilities driven by different base rates (Female: 11.3% vs. Male: 7.2%), while JSD—comparing binned histogram shapes—is less sensitive to uniform displacement.

Pair	**JSD**	**JSD %ile**	**Wass**	**Wass %ile**	**ΔAUC**	**$\mathbf{ECE_{cross}}$**
F → M	.002	48	.004	100	+.001	.038
M → F	.003	98	.005	100	+.024	.036

Table 2. Demographics-only model, sex contexts. Percentiles relative to intra-context baseline. Baseline ECE ≈ 0.009. Metric disagreement between JSD and Wasserstein reflects their complementary sensitivity to different types of distributional shift.

4.4. **Robustness: Items + Demographics Model**

The Items + Demographics model (AUC ≈ 1.0) confirmed that divergence persists even with near-perfect discrimination. For age contexts, 55+ → 18–34 was flagged (JSD z =

4.4, percentile = 100%; Wasserstein $z = 3.0$). For sex, JSD showed no divergence beyond baseline, but Wasserstein was dramatically elevated (Female → Male: $z = 33.9$; Male → Female: $z = 35.5$), again reflecting base-rate differences. Calibration shift was small but consistently present (ECE cross: 0.013–0.027 versus baseline 0.016–0.022), confirming that distributional instability operates independently of discrimination performance.

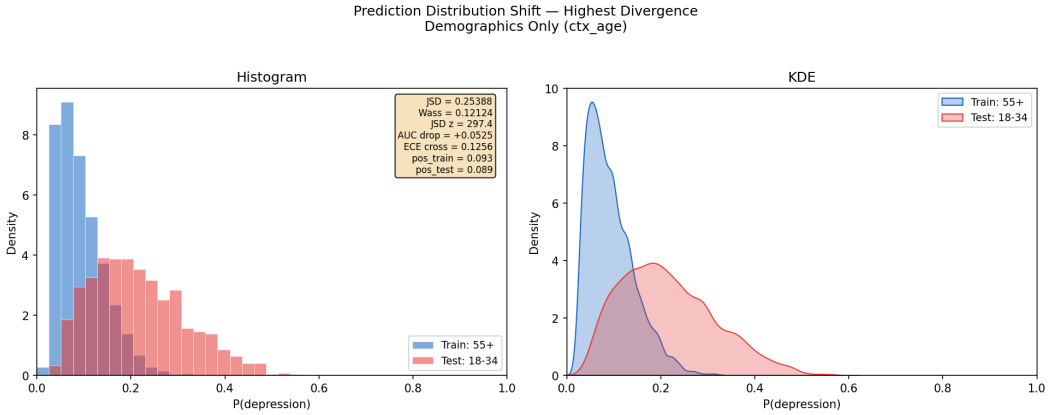

Figure 2. Predicted probability distributions for the highest-divergence pair (55+ → 18–34, demographics-only model). Left: histogram overlay; right: KDE. The model trained on older adults concentrates predicted mass near extremes, while the younger-adult distribution spreads across the probability range—producing overconfident low-risk assignments despite similar underlying prevalence (∼9% in both groups).

5. Discussion

What the gate does. The protocol provides a necessary-condition check: if output distributions are not stable across demographic contexts—exceeding normal within-context variation—the epistemic burden for interpreting those outputs as intervention-ready is elevated. This framing is deliberately modest; we propose a computationally cheap diagnostic for governance workflows, not a causal discovery method.

Dissociation of AUC and calibration. Our most actionable finding is that distributional instability can be invisible to AUC-based validation. The 55+ → 35–54 pair maintained AUC while exhibiting significant distributional shift and calibration degradation. This mirrors concerns raised in the fairness literature [13] about reliance on aggregate performance metrics. A governance framework monitoring only AUC would miss the instability our protocol detects.

Metric complementarity. JSD and Wasserstein capture different instability types: Wasserstein was more sensitive to base-rate-driven translational shift (sex contexts), while JSD better captured shape differences (age contexts). A governance protocol should employ both to avoid blind spots.

Limitations and sensitivity. Our analysis uses a single survey instrument (PHQ-9) from one country. The demographics-only model has modest discrimination (AUC ≈ 0.67), though the diagnostic concerns the output *distribution*, not clinical utility. Preliminary tests with alternative bin counts (25, 100) and a 90th-percentile threshold produced identical flagging outcomes. Future work should extend to temporal contexts and EHR data [6].

Practical implications. For clinicians and deployers, the protocol provides a pre-deployment checklist item: run the divergence diagnostic across demographic strata before using a risk model for triage or intervention decisions. Flagged pairs indicate that risk scores

should not be treated as calibrated without additional validation, consistent with emerging governance frameworks [14, 15].

6. Conclusion

We have proposed and empirically demonstrated a governance diagnostic: comparing output-distribution divergence across demographic contexts against within-context baselines as a pre-interpretation gate for mental health AI. Using NHANES PHQ-9 data, we showed that cross-context divergence systematically exceeds baseline variation, particularly across age groups, and that distributional instability can occur independently of discrimination degradation. These findings support using divergence metrics as a practical, low-cost governance mechanism for flagging when model predictions may not safely bear causal or intervention-oriented interpretation.

AI Use Statement

Generative AI tools were used for language editing and polishing during preparation of this paper. The author takes full responsibility for all content.

References

[1] A. M. Chekroud, R. J. Zotti, Z. Shehzad, R. Gueorguieva, M. K. Johnson, M. H. Trivedi, T. D. Cannon, J. H. Krystal, and P. R. Corlett. "Cross-trial prediction of treatment outcome in depression: a machine learning approach". In: *The Lancet Psychiatry* 3.3 (2016), pp. 243–250.

[2] J. Pearl. *Causality: Models, Reasoning, and Inference*. 2nd. Cambridge University Press, 2009.

[3] M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

[4] M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian. "Causal inference and counterfactual prediction in machine learning for actionable healthcare". In: *Nature Machine Intelligence* 2 (2020), pp. 369–375.

[5] E. Bareinboim and J. Pearl. "Causal inference and the data-fusion problem". In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7345–7352.

[6] J. Pearl and E. Bareinboim. "External validity: From do-calculus to transportability across populations". In: *Statistical Science* 29.4 (2014), pp. 579–595.

[7] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

[8] M. Sugiyama, M. Krauledat, and K.-R. Müller. "Covariate shift adaptation by importance weighted cross validation". In: *Journal of Machine Learning Research* 8 (2007), pp. 985–1005.

[9] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift". In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[10] A. Ramdas, N. García Trillos, and M. Cuturi. "On Wasserstein two-sample testing and related families of nonparametric tests". In: *Entropy* 19.2 (2017), p. 47.

[11] Centers for Disease Control and Prevention. *National Health and Nutrition Examination Survey, 2017–March 2020 Pre-Pandemic*. https://wwwn.cdc.gov/nchs/nhanes/. 2021.

[12] K. Kroenke, R. L. Spitzer, and J. B. Williams. "The PHQ-9: Validity of a brief depression severity measure". In: *Journal of General Internal Medicine* 16.9 (2001), pp. 606–613.

[13] A. Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big Data* 5.2 (2017), pp. 153–163.

[14] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act)*. 2024.

[15] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. 2023.

Appendix A. Additional Experimental Results

This appendix presents additional visualisations. The main paper is self-contained.

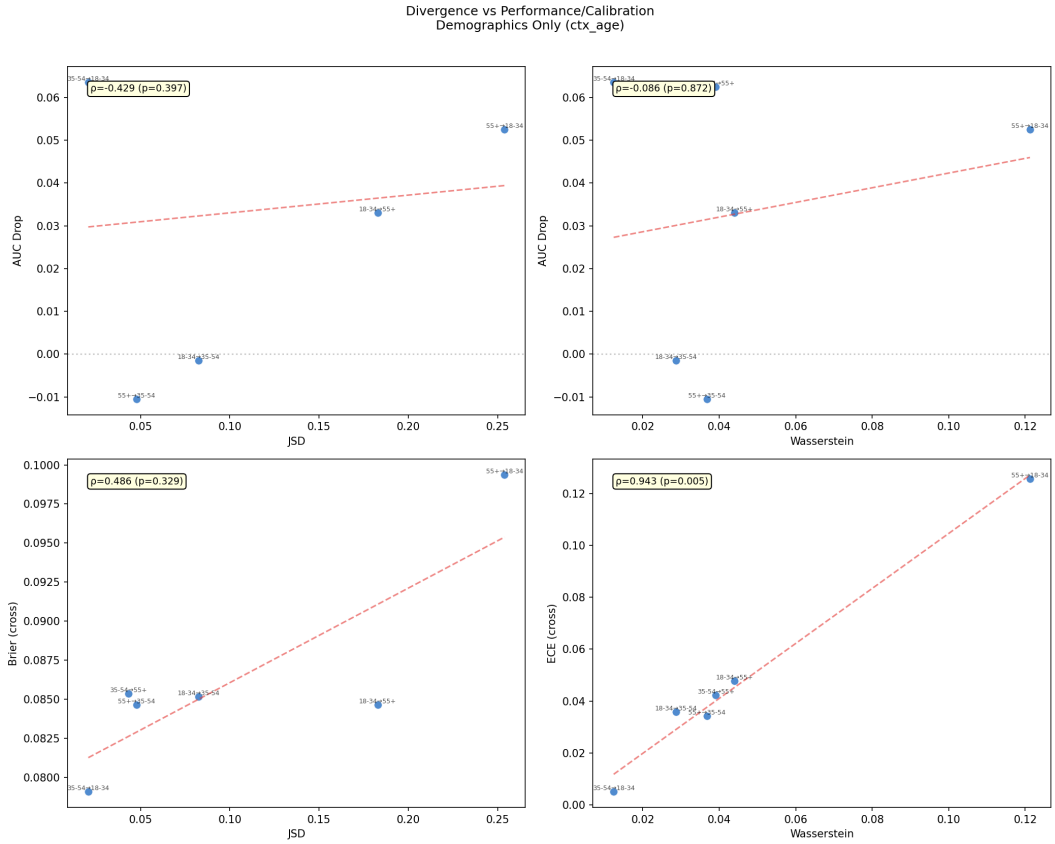

Figure 3. Divergence versus performance for age-group pairs (demographics-only model). The weak divergence–AUC correlation confirms that distributional instability primarily manifests as calibration degradation.

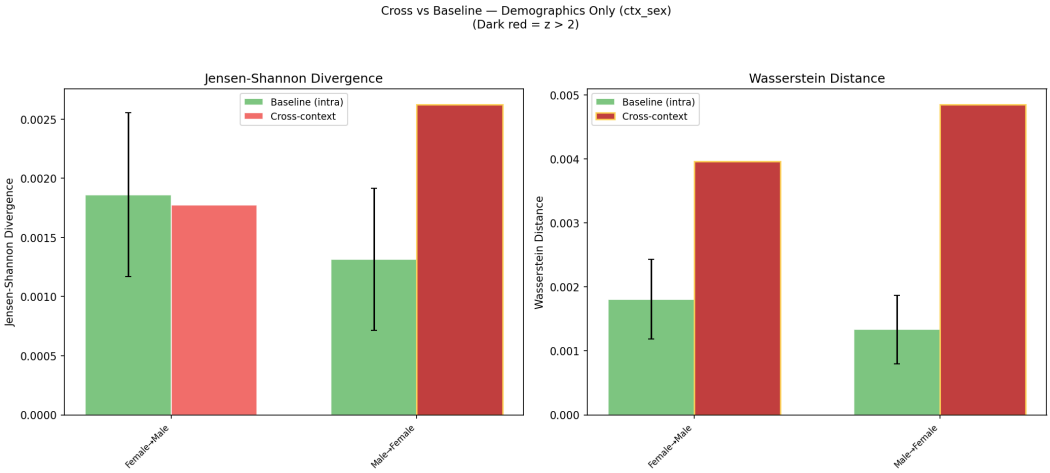

Figure 4. Cross-context versus baseline divergence for the demographics-only model, sex contexts. The asymmetric flagging pattern demonstrates JSD and Wasserstein complementarity.