Resilience Outcomes Benchmark: Toward an Outcome-Labeled Coping Strategy Dataset for Precision Mental Health

Saurabh Anand
Independent Researcher
Canada
anandsaurabh17@gmail.com

Abstract

Most AI benchmarks still measure *static competence*—accuracy on fixed math, coding, and knowledge-recall tasks. But intelligence that matters in care is *adaptive effectiveness*: knowing *which actions help which people, at what dose, and on what timeline*. Mental health AI today lacks the foundational resource that transformed vision (ImageNet) (1) and language (Common Crawl): outcome-labeled supervision. We propose the **Resilience Outcomes Benchmark (ROB)**, a two-phase, openly shareable dataset that operationalizes outcome-supervised learning for recovery after major stressors (bereavement, divorce, job loss, illness). **Phase 1** releases 10k+ expert-labeled vignettes linking context to coping strategies with effectiveness and harm-risk ratings (PHI-free), enabling contextual strategy ranking. **Phase 2** is a governed outcomes cohort capturing consented, real-world strategy use with dose/adherence and validated outcomes at 30/90 days (PHQ-9, GAD-7, WHO-5) (2; 3; 4), evaluated via a *models-to-data* server (no row-level export). ROB turns *context→strategy→outcome* into measurable supervision with benchmarks for NDCG@k, dose–response, and calibrated 30/90-day forecasts. By filling this gap, ROB could catalyze precision mental health—a domain with $1T+ global costs (5).

1 AI Task Definition

Scientific question: Given a person's context (demographics, stressor type/severity, supports, time since onset) and candidate coping strategies, can AI predict (a) which strategies will be most effective, (b) at what dose/intensity, and (c) the expected recovery trajectory?

Primary tasks: *(i) Contextual Strategy Ranking*—input $(x, \mathcal{S}) \rightarrow$ a ranking over strategies; *(ii) Dose–Response Prediction*—estimate optimal frequency/duration (minimum effective dose); *(iii) Trajectory Forecasting*—predict ΔPHQ-9/ΔGAD-7/ΔWHO-5 at 30/90 days with calibrated prediction intervals.

Metrics: NDCG@k; harm-penalized top-k (penalty λ on expert "risk" labels); dose–response via isotonic/GP fits (minimum effective dose); forecasting RMSE and Expected Calibration Error (ECE); time-to-threshold via survival C-index.

2 Dataset Rationale: Why This Is the Bottleneck

Gap: Existing datasets (e.g., CCMH, UK Biobank, MIMIC-III) are large but lack *linked intervention→outcome* supervision (6; 7; 8). Current mental-health AI can *sound* supportive yet cannot answer: "Will 20-minute daily walks help this grieving person more than weekly friend calls?" We lack linked *strategy→outcome* supervision. **Why now:** Digital tools show heterogeneous

outcomes; validated measures exist (Brief COPE→coping; PHQ-9/GAD-7/WHO-5→outcomes) but are not connected (9; 2; 3; 4); theory (Dual Process Model) suggests loss- and restoration-oriented mixes that require personalization (10).

Data types & labels: Phase 1 (open). 10k composite vignettes across stressors; 3–6 strategies per vignette mapped to *public strategy taxonomy codes* (e.g., Brief COPE categories); expert ratings: effectiveness (0–5), harm risk (0–5), cultural fit, expected latency; PHI-free. **Phase 2 (governed).** 3k–5k consented participants logging chosen strategies with dose/adherence and outcomes at T0/T30/T90; raw data in a secure enclave; a *models-to-data* server returns metrics only. **Resolution.** Short-horizon trajectories (T0/T30/T90) enable dose–response and recovery-curve modeling beyond single-timepoint associations.

3 Acceleration Potential

Model development: Makes outcome-supervised, risk-aware learning first-class; supports combination/sequence recommendations and confidence-calibrated forecasts (11). **Science enabled:** (1) resilience phenotypes (who responds to what) (12); (2) minimum effective doses (dose–response per strategy) (13); (3) cultural interactions (cross/within-culture effects) (14); (4) sequences (timing/ordering) (15). **Cross-disciplinary uses:** Beyond psychiatry, ROB could inform workplace wellness (burnout prevention), education (student resilience), disaster recovery, and healthcare planning. **Impact:** Even a 10% improvement in strategy matching could save $100B+ annually via fewer ineffective interventions and better adherence, within a domain exceeding $1T in global costs (5). **Analogy:** As ImageNet catalyzed computer vision (1), ROB can catalyze precision mental health.

4 Data Creation Pathway (Practical & Ethical)

Phase 1 (Months 0–3): 15-person clinical panel; diverse vignettes; three raters per strategy with adjudication; inter-rater reliability. **Cost:** $80–120k. **Phase 2 (Months 3–12):** Secure 2–3 LOIs (university counseling center, digital platform, NGO); integrate micro-surveys into existing care pathways; IRB/REB oversight; *models-to-data* evaluation server. **Cost:** $150–200k. **Ethics & safety:** Independent REB/IRB advisor; pre-registered protocol; adverse-event escalation; subgroup fairness reports; DP-sanitized excerpts only; no raw clinical notes. Expert harm-risk labels (conservative thresholds); red-teaming; subgroup harm audits. **No crisis use:** research-only, not a substitute for emergency support.

5 Feasibility, Originality, Shareability & Openness

Feasible: Synthetic Phase 1; enclave-scored Phase 2. **Original:** First unified, *outcome-labeled* dataset for context→strategy→outcome. **Shareable:** Phase 1 CC-BY 4.0; code Apache-2.0; Phase 2 exposes aggregates + evaluation API. **Docs:** Datasheet & Labeler Guidelines (Phase 1); Model Card (baselines).

Baselines & tracks: Reference baselines: (a) majority-strategy, (b) empathy-only LLM, (c) retrieval-augmented + expert labels, (d) dose-aware isotonic regression. Leaderboard tracks: *Open* (Phase 1) and *Governed* (Phase 2); subgroup metrics (age/culture/SES) required for listing.

Timeline & budget (MVP): Months 0–2: taxonomy, schema, tooling. 2–4: Phase 1 release (10k vignettes) + baselines. 4–6: eval server + first leaderboard. 6–12: Phase 2 cohort start + initial aggregates. **Total:** $250–350k.

Limitations: Associations (not causality) in v1; cultural generalization needs stratified sampling/reporting; adherence is noisy. ROB logs dose/adherence and co-interventions to reduce confounding and reports subgroup results by design.

Appendix A: Minimal schema (vignette, public)

```
{ "vignette_id":"V-2847", "stressor":"bereavement_parent",
  "context":{"age_band":"25-34","culture_region":"South_Asia",
             "supports":["faith_community","siblings"],
             "severity":3,"time_since_days":45},
  "strategies":[
    {"code":"SOC_calls","desc":"Two 30-min calls/week"},
    {"code":"BH_walks","desc":"20-min daylight walks"},
    {"code":"EF_journaling","desc":"10-min guided journaling"}],
  "expert_ratings":{
    "SOC_calls":{"effectiveness":4.7,"risk":0.3,"latency_days":7},
    "BH_walks":{"effectiveness":4.2,"risk":0.1,"latency_days":7},
    "EF_journaling":{"effectiveness":3.9,"risk":0.5,"latency_days":10}}
}
```

Appendix B: Minimal schema (outcomes, enclave)

```
{ "participant_id":"P-7c2f","event_id":"E-9a10","stressor":"bereavement",
  "context":{"age_band":"25-34","culture_region":"South_Asia","severity":4,
             "supports":["family","faith_community"]},
  "baseline":{"PHQ9":14,"GAD7":11,"WHO5":32},
  "strategies_used":[
    {"date":"2025-05-01","code":"SPIR_prayer","dose_per_week":5,"adherence":0.9},
    {"date":"2025-05-03","code":"SOC_calls","dose_per_week":2,"adherence":0.7}],
  "co_interventions":{"therapy_sessions":2,"medication_change":false},
  "outcomes":[
    {"t_days":0,"PHQ9":14,"GAD7":11,"WHO5":32},
    {"t_days":30,"PHQ9":10,"GAD7":8,"WHO5":44},
    {"t_days":90,"PHQ9":6,"GAD7":5,"WHO5":60}]}
```

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[2] K. Kroenke, R. L. Spitzer, and J. B. W. Williams. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

[3] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe. A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10):1092–1097, 2006.

[4] C. W. Topp, S. D. Østergaard, S. Søndergaard, and P. Bech. The WHO-5 Well-Being Index: A systematic review of the literature. *Psychotherapy and Psychosomatics*, 84(3):167–176, 2015.

[5] The Lancet Commission on Global Mental Health. Global mental health and sustainable development. *The Lancet Psychiatry*, 7(10):893–924, 2020.

[6] Center for Collegiate Mental Health (CCMH). 2023 Annual Report. Pennsylvania State University, 2023.

[7] K. A. S. Davis, M. C. Cullen, J. Adams, et al. Mental health in UK Biobank. *BJPsych Open*, 6(2):e18, 2020.

[8] A. E. W. Johnson, T. J. Pollard, L. Shen, et al. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[9] C. S. Carver. You want to measure coping but your protocol's too long: Consider the brief COPE. *International Journal of Behavioral Medicine*, 4(1):92–100, 1997.

[10] M. Stroebe and H. Schut. The Dual Process Model of coping with bereavement: Rationale and description. *Death Studies*, 23(3):197–224, 1999.

[11] J. Delgadillo, A. Huey, and S. Bennett. Targeted prescription of psychological therapies to reduce treatment waste. *Psychological Medicine*, 50(1):1–10, 2020.

[12] I. R. Galatzer-Levy and G. A. Bonanno. Beyond normality in the study of bereavement: Heterogeneity in depression outcomes. *Social Science & Medicine*, 74(12):1987–1994, 2012.

[13] I. Nahum-Shani, S. N. Smith, B. Spring, et al. Just-in-time adaptive interventions (JITAIs) in mobile health. *Annals of Behavioral Medicine*, 52(6):446–462, 2018.

[14] L. J. Kirmayer, A. Narasiah, M. Munoz, et al. Common mental health problems in immigrants and refugees: General approach in primary care. *CMAJ*, 183(12):E959–E967, 2011.

[15] D. C. Mohr, M. Burns, S. Schueller, G. Clarke, and M. K. Klinkman. Behavioral intervention technologies: Evidence review and recommendations for future research. *General Hospital Psychiatry*, 35(4):332–338, 2013.

[16] E. I. Fried and D. J. Robinaugh. Systems all the way down: Embracing complexity in mental health research. *BMC Medicine*, 18:205, 2020.

[17] Lyra Health. Clinical outcomes report, 2023.

[18] B. Inkster, K. Sarda, and C. Subramanian. An empathy-driven, conversational AI agent (Wysa) for digital mental well-being. *JMIR mHealth and uHealth*, 6(11):e12106, 2018.

[19] National Alliance of Trauma Recovery Centers. Impact Report, 2023.

[20] L. van der Krieke, M. Jeronimus, N. Blaauw, et al. How to assess momentary mental states? EMA in psychiatry. *JMIR Research Protocols*, 4(3):e100, 2015.

[21] S. Graham, M. Depp, E. Lee, and E. Nebeker. Artificial intelligence for mental health and mental illnesses. *Current Psychiatry Reports*, 21:116, 2019.

[22] Z. D. Cohen and R. J. DeRubeis. Treatment selection in depression. *Annual Review of Clinical Psychology*, 14:209–236, 2018.

Funding Disclosure

This research received no external research funding.