

OH•AI•SYS™ versus VA PTSD Self-Screen: A Pilot Validation Study Using LLM-Emulated Behavioral Digital Twins

2-Dooz Research Labs

Technical Report: Validation Series (PTSD-071426)

Tony Clark, CSO/PI

Published: July 20, 2026

Abstract

Diagnostic screening for Post-Traumatic Stress Disorder (PTSD) faces a classic operational trade-off: rapid screening tools reduce immediate friction and user burden, but often introduce a high rate of false positives, or miss nuanced cases: the “**Triage Burden**.” This pilot validation study compares the diagnostic screening performance of the multimodal **OH•AI•SYS™ AI Screening Platform**—utilizing a 55-question distress-tolerance baseline survey—against the **VA PTSD Self-Screening Tool** (a 5-item¹ adaptation of the PC-PTSD-5).

A pioneering **Behavioral Digital Twin (BDT)** methodology, comprising Large Language Model (GPT-5.4 nano and Gemini 3.5 Flash) generated cohort members ($N = 12$), was used to evaluate both screening methods against the clinical gold standard Post-Traumatic Stress Disorder Checklist for DSM-5. PCL-5 ≥ 33 was used as the cutoff for determining an apparent PTSD diagnosis that would warrant clinical assessment. The OH•AI•SYS platform demonstrated **100% diagnostic alignment** (12/12 correct classifications), whereas the VA PTSD Self-Screen achieved **91.7% accuracy** (11/12 correct classifications), yielding a single false positive.

This paper considers the systemic costs associated with the Triage Burden and outlines the statistical architecture for a fully powered, 120–150 BDT subjects follow-on validation trial.

1. Background & Clinical Dilemma

Screening tools are essential diagnostic gatekeepers, designed to identify individuals who may benefit from structured clinical assessments. In mental health workflows, however, short screening instruments often prioritize high sensitivity (minimizing false negatives) at the steep expense of specificity (increasing

¹ The online survey, which can be found at [VA PTSD Self-Screen](#), contains a conditional structure: a gating question followed by five additional items, administered only if the gating criterion is met.

false positives). This design philosophy is built on the assumption that secondary assessments will filter out false alarms.

In practice, this imbalance produces a significant Triage Burden:

- **Clinical Bottlenecks:** A high false-positive rate floods restricted mental health care pipelines with unnecessary, long-form clinical evaluations for healthy individuals.
- **Patient Experience Degradation:** False alarms cause unnecessary anxiety, stigma, or "diagnostic fatigue" for individuals who screen positive but do not require intervention.
- **Systemic Inefficiency:** While a shorter tool is faster to complete, it shifts the temporal and financial burden downstream, requiring clinicians to conduct 45-to-60-minute diagnostic interviews to resolve screening errors.

This pilot study directly compares a rapid, triage-focused screening approach with a comprehensive, robust, and multimodally augmented screening system.

2. Screening Architectures Compared

2.1 The OH•AI•SYS™ Platform

The [OH•AI•SYS](#) (pronounced *oasis*) screening engine is designed to prioritize diagnostic precision and minimize misses without generating excessive false-positive noise. Its dual-layer architecture includes:

1. **Clinical Assessment:** A robust, 55-question baseline assessment covering stress, anxiety, depression, trauma, and suicidality: structurally designed to mirror or exceed the diagnostic breath of the PCL-5. The accuracy of the OH•AI•SYS baseline assessment was evaluated in this pilot study relative to the VA PTSD Self-Screen.
2. **Wearable Integration:** Real-time physiological telemetry ingested via a **Resilience Technology Protocol (RTP) API** compliant smart ring (such as the 2-Dooz™ Resilience Smart Ring), processing Finger Skin Temperature (FST) to deliver real-time distress alerts and continuous distress-tolerance tracking. The real-time distress-tolerance tracking is seeded from the baseline assessment.

The platform categorizes distress-tolerance into one of three color-coded tiers:

- **Red (Fragile):** Acute distress; immediate professional intervention recommended.
- **Yellow (Existing):** Average-to-below-average resilience with episodic symptoms; a professional assessment recommendation *may* accompany this classification.
- **Green (Thriving):** Excellent emotional regulation, strong coping skills, and robust mental fitness.

2.2 The VA PTSD Self-Screening Tool

The [VA PTSD Self-Screen](#) is an adaptation of the National Center for PTSD's **Primary Care PTSD Screen for DSM-5 (PC-PTSD-5)**. It is a brief **5-item screen** engineered for rapid, high-throughput triage. Operating strictly on the triage principle, it treats over-flagging as an acceptable operational trade-off to maximize speed.

3. Methodology & Digital Twin Cohort

3.1 The Behavioral Digital Twin Paradigm

To evaluate and compare these tools under highly standardized, repeatable conditions, this study utilized a **Behavioral Digital Twin (BDT)** validation protocol. A BDT is a high-fidelity synthetic emulation of a human persona constructed from aggregate statistical patterns of human experience.

The Pilot Study comprised **12 behavioral digital twins**, each generated by a large language model (LLM) based on a [randomly](#) generated PCL-5 total score (see Exhibit). The set of randomly selected PCL-5 total scores for the study are as follows: 7, 45, 67, 56, 46, 23, 11, 79, 42, 71, 54, and 66.

- **GPT-5.4 nano** generated 7 personas (PCL-5 scores: 7, 67, 56, 11, 79, 42, 54).
- **Gemini 3.5 Flash** generated 5 personas (PCL-5 scores: 45, 46, 23, 71, 66).

The synthetic cohort (N=12) comprised 8 women and 4 men, with ages ranging from 22 to 55 (mean age: 47 years). Ethno-demographic composition included 7 Caucasian, 3 Asian, and 2 Hispanic profiles. Eight of the 12 twins had a PCL-5 score ≥ 33 , meeting the clinical threshold indicating a "need for additional assessment."

BEHAVIORAL DIGITAL TWIN COHORT DEMOGRAPHICS (N=12)	
=====	
Age Range: 22 - 55 (Mean: 47) Gender: 8 Female, 4 Male	
Ethnicity: 7 Caucasian, 3 Asian, 2 Hispanic	
Gold Standard Status: 8 PCL-5 positive (≥ 33), 4 PCL-5 negative	
=====	

3.2 Evaluation and Statistical Analysis

Each digital twin was prompted to complete three assessments: the PCL-5, the OH•AI•SYS survey, and the VA PTSD Self-Screen.

Screening outputs were classified relative to the PCL-5 gold standard (PCL-5 ≥ 33):

- **Correct:** The screen's clinical recommendation (treatment vs. no treatment) matched the PCL-5 gold-standard classification.
- **Incorrect:** The screen's recommendation conflicted with the PCL-5 indication.

3.3 Cohort Responses

The complete list of responses across all 12 behavioral digital twins (BDT) is detailed below.

Cohort Member	BDT 1	BDT 2	BDT 3	BDT 4	BDT 5	BDT 6	BDT 7	BDT 8	BDT 9	BDT 10	BDT 11	BDT 12
Age	29	30	40	30	22	22	34	32	34	55	42	48
PCL-5 Total	7	45	67	56	46	23	11	79	42	71	54	66
PCL-5 Q1	0	2	4	3	3	1	0	4	3	4	3	4
PCL-5 Q2	0	2	4	3	3	1	0	4	3	4	3	4
PCL-5 Q3	0	0	4	3	2	1	0	4	3	3	3	4
PCL-5 Q4	1	3	3	3	3	2	1	4	3	4	3	4
PCL-5 Q5	0	1	3	3	4	2	0	4	3	4	3	4
PCL-5 Q6	1	3	3	3	2	2	1	4	3	4	3	4
PCL-5 Q7	0	3	3	3	4	1	1	4	3	4	3	4
PCL-5 Q8	0	0	3	2	1	0	0	4	2	0	2	0
PCL-5 Q9	1	4	4	3	2	1	1	4	3	4	3	4
PCL-5 Q10	0	4	3	2	2	1	1	4	2	4	2	4
PCL-5 Q11	0	2	4	3	3	1	0	4	3	4	3	4
PCL-5 Q12	0	2	3	3	2	0	1	4	2	4	2	3
PCL-5 Q13	0	1	3	3	3	1	0	4	3	4	3	4
PCL-5 Q14	0	2	4	3	1	1	1	4	3	4	3	3
PCL-5 Q15	0	1	3	3	1	1	0	4	3	4	3	3
PCL-5 Q16	0	0	2	2	0	0	0	3	2	0	2	0
PCL-5 Q17	1	1	4	3	3	2	1	4	3	4	3	3
PCL-5 Q18	1	0	4	3	4	2	1	4	2	4	2	3
PCL-5 Q19	1	2	3	3	2	1	1	4	2	4	2	3
PCL-5 Q20	1	2	3	2	1	2	1	4	2	4	3	4
Therapy Indication	BDT 1	BDT 2	BDT 3	BDT 4	BDT 5	BDT 6	BDT 7	BDT 8	BDT 9	BDT 10	BDT 11	BDT 12
PCL-5 (>=33)	NO	YES	YES	YES	YES	NO	NO	YES	YES	YES	YES	YES
OHAISYS Screen	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct	Correct
VA PTSD Self Screen	Correct	Correct	Correct	Correct	Correct	Incorrect	Correct	Correct	Correct	Correct	Correct	Correct

Defining the OH•AI•SYS Screen as “C1” and the VA PTSD Self-Screening Tool as “C2,” the following results were observed:

- **OH•AI•SYS Screen (C1):** Correct in **12/12** cases (100% Sensitivity, 100% Specificity).
- **VA PTSD Self-Screening Tool (C2):** Correct in **11/12** cases (100% Sensitivity, 75% Specificity), producing **one false positive** on BDT 6 (PCL-5 = 23).

3.4 Contingency Matrix (2X2)

	C2 (VA) Correct	C2 (VA) Incorrect
C1 (OHAISYS) Correct	11 (a)	1 (b)
C1 (OHAISYS) Incorrect	0 (c)	0 (d)

The standard asymptotic McNemar's Chi-square test is insufficiently powered given only **one discordant pair** (b=1, c=0).

1. McNemar with Edwards' Continuity Correction:

$$\chi^2_{\text{corrected}} = \frac{(|b - c| - 1)^2}{b + c} = \frac{(|1 - 0| - 1)^2}{1} = 0 \implies p = 1.0$$

2. Binomial Exact Test:

Under the null hypothesis ($p = 0.5$), the probability of observing a 1 vs 0 split in discordant cells is:

$$P(\text{Result} \geq 1) = 2 \times (0.5)^1 = 1.0 \implies p = 1.0$$

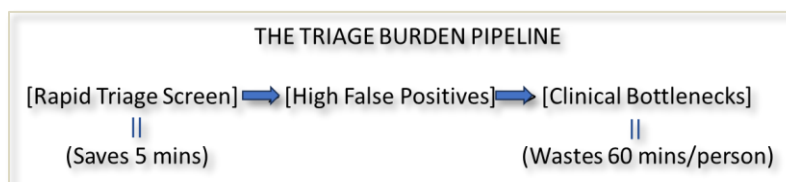
As shown immediately above, both tests yield a **p= 1.0**. Consequently, while the pilot's trend directionally favors the OH•AI•SYS Screen, the study is statistically underpowered to reject the null hypothesis of equal performance at this scale.

4. Discussion: The Clinical Cost of False Positives

The pilot study results validate the real-world performance differences dictated by the tools' underlying architectures. The VA Self-Screen's false positive on BDT 6 (an individual with mild, sub-clinical symptoms; PCL-5 = 23) exemplifies how rapid screening filters generate downstream noise.

While a single false positive may seem minor in a 12-subject sample, scaling this error rate to larger populations reveals severe administrative and clinical challenges:

- **Clinical Bottlenecks:** A high false-positive rate increases the workload for mental health professionals who must perform unnecessary follow-up assessments on healthy individuals. For example, in a cohort of 1,000 individuals, an 8.3% false-positive error rate would generate approximately 83 unnecessary clinical referrals, consuming ~83 hours of clinician time (at 60 minutes per assessment) and potentially affecting patient psychology.
- **Patient Experience:** False positives can cause unnecessary anxiety, stigma, or "diagnostic fatigue" for patients who are screened as positive when they are not.



In contrast, the more comprehensive, 55-question OH•AI•SYS platform successfully matched the gold-standard PCL-5 result in every case, demonstrating that its multi-domain design successfully protects against false-positive inflation.

5. Conclusion & Next Steps

This proof-of-concept pilot demonstrates the feasibility and accuracy of utilizing synthetic **Behavioral Digital Twin** methodologies for executing rapid, highly controlled comparative evaluations of diagnostic software.

To confirm these results with statistical significance, 2-Dooz Research Labs is initiating a fully powered follow-on study:

- **Cohort Size:** Planned scaling up to **120 to 150 Behavioral Digital Twins**.
- **Statistical Power:** Sized to achieve $\geq 80\%$ power (at $\alpha = 0.05$) based on the pilot's observed 8.3% discordance rate.
- **Primary Analysis Plan:** Standard McNemar's Chi-square if discordant pairs are sufficient (≥ 20); otherwise, the Binomial Exact Test or McNemar's with Continuity Correction will be utilized to handle low-discordance margins rigorously.

6. References

- [1] OH•AI•SYS Distress-Tolerance Survey, [OH•AI•SYS](#), accessed June 2026.
- [2] DSM-5 (PC-PTSD-5), <https://www.ptsd.va.gov/professional/assessment/documents/pc-ptsd5-screen.pdf>, accessed June 2026.
- [3] PCL-5, <https://www.ptsd.va.gov/professional/assessment/adult-sr/ptsd-checklist.asp>, accessed June 2026.
- [4] GPT-5.4 nano, <https://duck.ai/chat?t=ffab&atb=v344-1>, accessed June 2026.
- [5] Gemini 3.5 Flash, <https://gemini.google.com/app>, accessed June 2026.