# PupiLUX: Preliminary Technical Validation Report

A 17-subject pilot validation of a smartphone pupillometry system for neuro-ICU and emergency-room triage

**Dr. Ajay Bakshi**
R.Tej Health Analytics
pupilux.ai · info@pupilux.ai

April 2026

**PRELIMINARY PILOT REPORT**

This is a **preliminary, single-site, single-grader pilot validation** of the PupiLUX smartphone pupillometry system. Results are reported transparently to establish a baseline for subsequent multi-site, multi-grader validation. Inter-rater reliability, cross-device generalization, and prospective clinical validation are planned for subsequent studies.

**IN ONE PARAGRAPH**

PupiLUX is a smartphone-based pupillometry system designed for bedside screening and triage in neuro-ICU and emergency-room settings. On a 17-subject pilot cohort, PupiLUX measures absolute pupil diameter with a mean absolute error of **0.52 mm**. Its reactivity metrics — constriction percentage, latency, velocity — are mathematically scale-invariant and were validated against a pharmacologically-induced reactivity asymmetry. Most importantly, for the clinically-actionable question of *"is there anisocoria, and in which direction"*, PupiLUX agreed with a blinded expert grader on **4 of 4 (100%) clean drug-phase frame pairs**, with definite grader confidence at asymmetries as small as **0.11 mm**. **The recurrent finding of this report** is that PupiLUX's two clinically-actionable signals — **reactivity and direction of anisocoria** — are robustly detected *even in the small-pupil regime where the absolute-diameter detector is most error-prone*. These are precisely the signals that drive bedside triage decisions in the neuro-ICU, where TBI patients on continuous opioid sedation present with pharmacologically-suppressed pupils throughout their admission, and in the emergency room, where brainstem injury, pontine stroke, and drug-toxicity presentations all converge on the same miotic pupil phenotype.

REGULATORY POSITION

PupiLUX is a **measurement/screening tool, not a diagnostic device**. Its output contains raw metrics, reference ranges, and status indicators only. No clinical interpretation is provided. The clinician retains full responsibility for diagnostic judgment. For informational and screening purposes only.

## 1. The clinical problem

In the neuro-ICU and the emergency room, the foundation of objective pupil assessment is still the manual penlight exam: a bedside clinician looks at each pupil in turn, estimates size, and judges reactivity. This exam has known limitations. In a large single-blinded observational study of 2,329 paired pupil assessments in neurologically ill patients, inter-rater reliability between trained practitioners was only moderate for pupil size ($\kappa = 0.54$), and poor for reactivity ($\kappa = 0.40$); only a minority of pupils labeled non-reactive by practitioners were confirmed non-reactive by quantitative pupillometry[1]. Automated, quantitative measurement removes this subjectivity — but dedicated pupillometers are expensive, tethered to a single bedside location, and rarely available in the settings where the epidemiology of traumatic brain injury, stroke, and critical-care admission is highest.

In the emergency room, the time between an initial neurological exam and a CT read opens a second gap in objective monitoring. In a recent retrospective cohort from a tertiary hospital in a resource-constrained setting, the median door-to-CT time across unselected presentations exceeded **5 hours**; even in the fast-tracked hyperacute stroke subgroup the median was still **2 hours 19 minutes**[2]. During that window, TBI patients with evolving intracranial pressure or brainstem compromise, stroke patients awaiting thrombolysis decisions, and poisoning presentations with altered pupil state have no objective pupil monitoring — only repeated subjective manual checks with the inter-rater variability described above.

**What actually drives triage decisions at the bedside is not the exact millimetre diameter of the pupil.** It is two simpler questions: *is the pupil still reacting to light, and if the two pupils differ, which side is larger?* These two signals — reactivity and direction of anisocoria — are the clinically-actionable output of a bedside pupil exam. They drive decisions such as urgent CT, naloxone, atropine reversal, anticholinesterase titration, or neurosurgery consult. The remainder of this report demonstrates that PupiLUX produces both of these signals accurately on a pilot validation cohort — including, critically, on the miotic pupils where a visible-light smartphone detector has its greatest absolute-diameter error (Pillars 2 and 3 below).

Dedicated digital pupillometers solve the measurement problem. They cost around **USD 5,000**, require training, remain tethered to a single bedside location, and do not integrate into the chart. In resource-constrained neuro-ICUs and emergency departments — exactly where the burden of head injury, stroke, and acute neurological presentations is highest — they are rarely available.

A smartphone already has the camera, the torch, and the compute. The question this report addresses is whether that hardware, driven by on-device AI, can produce numbers a clinician can actually trust for screening and triage in TBI, neuro-ICU, and ER settings.

## 2. System overview

**Hardware.** A standard iPhone (iOS 17+) with rear camera and torch. No external adapters, no accessories. The phone is held approximately 12 inches (30 cm) from the patient's face in a dim room. The torch delivers the stimulus flash.

**Capture.** A single 7-second bilateral recording captures both eyes simultaneously at 1080p / 30 fps. The test proceeds through three phases automatically:

- **Baseline** (2 s) — pupil at rest
- **Stimulus** (1 s) — torch flash
- **Recovery** (5 s) — pupil constriction and partial re-dilation

Frames are captured as HEIF and analyzed after the recording completes, avoiding camera-thread blocking during acquisition.

**Detection pipeline.** A three-layer cascade runs on-device:

- Apple Vision framework face landmarks isolate the eye regions
- A bundled YOLOv8n-seg model detects pupil and iris masks within each eye crop
- A tracked-position fallback guides subsequent crops once a first high-confidence detection is established

**Per-frame confidence handling and failure behaviour.** Every processed frame carries a YOLO per-frame confidence score. PupiLUX does not report clinical metrics derived from low-confidence frames: frames below a **0.50** confidence floor are excluded from the analysis rather than contributing a false reading. This confidence floor was added on 2026-04-08 after pilot analysis showed that low-confidence frames drove most of the absolute-diameter error. Per-frame confidence is persisted in the trial CSV export, providing an auditable trail for downstream analysis.

The system has two user-visible failure modes at the session level. (i) If the rear camera cannot be obtained, the test does not start and the app surfaces a device-level error. (ii) If face detection does not succeed within 30 seconds of a test attempt — ambient light too low, face out of frame, extreme angle — the app displays a detection-timeout overlay rather than silently recording unusable frames, and offers the user a retry. These paths are implemented today in the Trial build and are described here so that reviewers can verify the claim directly in the codebase.

What is **not yet surfaced** is a per-recording confidence summary on the PupiLUX Pro Report itself. Per-frame confidence is preserved in the raw export and drives the upload-floor filter, but the clinician reading the Pro Report does not currently see a single "overall recording confidence" badge. Surfacing this — so that a recording mostly composed of sub-floor frames is flagged on the report itself rather than just producing fewer data points — is on the roadmap for the next development phase.

**Calibration.** Pupil diameter in millimetres is computed from pupil pixels divided by an **iris-ruler** assumption: the horizontal visible iris diameter (white-to-white corneal diameter) of an adult human eye is taken as 11.7 mm, consistent with the population mean of 11.71 ± 0.42 mm reported by Rüfer and colleagues from Orbscan II topography measurements across 390 healthy adults[3]. A constant correction factor

(÷**1.167**) is then applied at computation time, fitted on an earlier high-confidence validation subset to bring the post-correction machine output into the target 0.95–1.10 M/E range on that subset. The generalization of this single-constant correction across the full 17-subject cohort is quantified in §4.3 and §7.3.

**Bilateral analysis.** Both eyes are analyzed from the same recording and split via face landmarks. This is, to our knowledge, the only smartphone-based pupillometry system that captures both eyes simultaneously in a single pass — a prerequisite for meaningful anisocoria detection, which requires time-synchronized L/R comparison.

**Outputs.** Per eye: baseline pupil diameter (BPD), minimum pupil diameter (MPD), constriction percentage (CP), latency (LAT), maximum constriction velocity (MCV), average dilation velocity (ADV), and time to 75% recovery (T75). No diagnostic interpretation. No "normal/abnormal" verdict. Raw metrics, reference ranges, and status indicators only (see **Figure 1**).

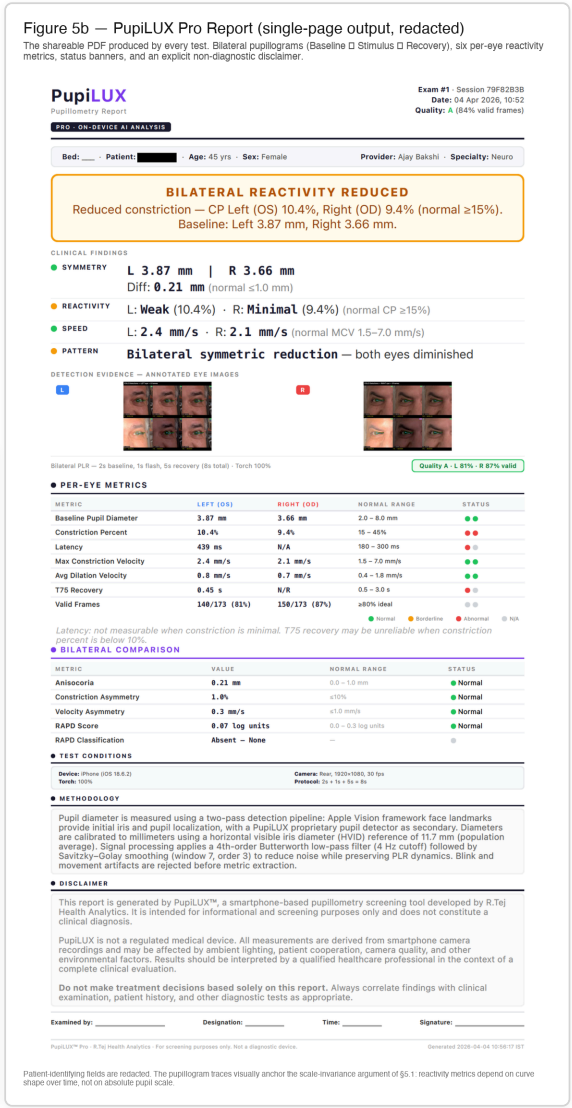
Figure 5a — PupiLUX bedside acquisition (no external adapter)

iPhone held ≈12 inches (30 cm) from the patient's face in dim ambient lighting. Rear camera + torch perform a bilateral PLR test in 7 seconds; both eyes captured in a single recording.

Acquisition geometry per whitepaper §2. The phone-only workflow eliminates the consumable-cost and supply-chain barriers that have kept dedicated infrared pupillometers (NPI-200/300) out of the Indian ICU/ER market.

Figure 5b — PupiLUX Pro Report (single-page output, redacted)

The shareable PDF produced by every test. Bilateral pupillograms (Baseline □ Stimulus □ Recovery), six per-eye reactivity metrics, status banners, and an explicit non-diagnostic disclaimer.




Patient-identifying fields are redacted. The pupillogram traces visually anchor the scale-invariance argument of §5.1: reactivity metrics depend on curve shape over time, not on absolute pupil scale.

***Figure 1.*** *(a) Bedside acquisition geometry — single iPhone, rear camera and torch, approximately 12 inches (30 cm) from the patient's face, dim room, single 7-second bilateral recording. No external adapters. (b) Example PupiLUX Pro Report (single-page output; patient-identifying fields redacted). The report shows raw per-eye metrics (BPD, CP, LAT, MCV, ADV, T75), reference ranges, and status indicators only — no diagnostic interpretation. The clinician retains full interpretive responsibility.*

## 3. Validation design

Three orthogonal questions are addressed, each with its own dataset and methodology.

**Question 1 (absolute accuracy).** Does PupiLUX's pupil-diameter measurement match an expert human grader on the same frame?

**Question 2 (scale-invariance of reactivity).** Are reactivity metrics (CP, LAT, T75) mathematically robust to absolute-diameter error, and do they correctly detect a pharmacologically-induced reactivity asymmetry — *including in the miotic-pupil regime where absolute-diameter error is largest*?

**Question 3 (clinical relevance — anisocoria direction).** Does PupiLUX agree with a blinded expert grader on the *direction* of inter-ocular asymmetry — the actionable clinical question in triage — *including, again, in the miotic-pupil regime*?

Pillars 2 and 3 use the same pharmacologically-induced dataset, which by design produces a series of small, reactive and poorly-reactive pupils. This means Questions 2 and 3 are answered *on exactly the frames* that Question 1 finds hardest. If PupiLUX survives that, it is evidence that the two clinically-actionable signals (reactivity, direction) are robust in precisely the patient population where the screening tool is most valuable — the sedated neuro-ICU TBI patient and the ER presentation of brainstem compromise or drug toxicity.

**Site.** All validation data were captured at CFS-Delhi-01 (Centre for Sight, Delhi, India). All annotation was performed by a single grader (Grader A, the author), using custom polygon-tracing and direction-judgment tools. The grader was blinded to machine output at annotation time in all three studies.

**Limitation (stated up front).** A single grader means we cannot report inter-rater reliability. Independent grader validation by at least two additional blinded clinicians is planned for a follow-up study.

## 4. Pillar 1 — Pupil size measurement accuracy

### 4.1 Dataset

The Pillar 1 analysis set is **36 paired frames from 17 subjects**, all captured at CFS-Delhi-01 and annotated by a single expert grader. One additional frame was excluded from the final count because its backend-stored machine prediction was missing.

**Annotation quality filter.** The grader applied a strict quality floor to every candidate frame presented for annotation: any frame with partial eyelid closure, reflection occlusion exceeding 25% of iris, motion blur, or an incorrect eye crop was rejected. Three candidate subjects (CFS-020, CFS-024, CFS-029) had no frames accepted under this filter and contribute nothing to the analysis. This manual quality gate sits on top of the automated per-frame confidence floor described in §2, and its strictness feeds directly into the small-pupil failure-mode discussion in §4.3.

**Final dataset:** 36 paired frames, 17 subjects, 20 L-eye + 16 R-eye.

## 4.2 Method

For each frame, the expert grader traced polygons for pupil and iris in normalized image coordinates. Equivalent diameter was computed from shoelace polygon area (diameter = $2 \times \sqrt{(\text{area}/\pi)}$). Diameter was converted to mm by using the machine-reported iris diameter as an internal calibration ruler. The pupil-to-iris ratio (PIR) was computed for both the machine and the grader, and the **M/E ratio** = (machine PIR) / (grader PIR) was used as the primary agreement metric, because it is robust to iris-ruler errors that would affect both measurements identically (see **Figure 2**).

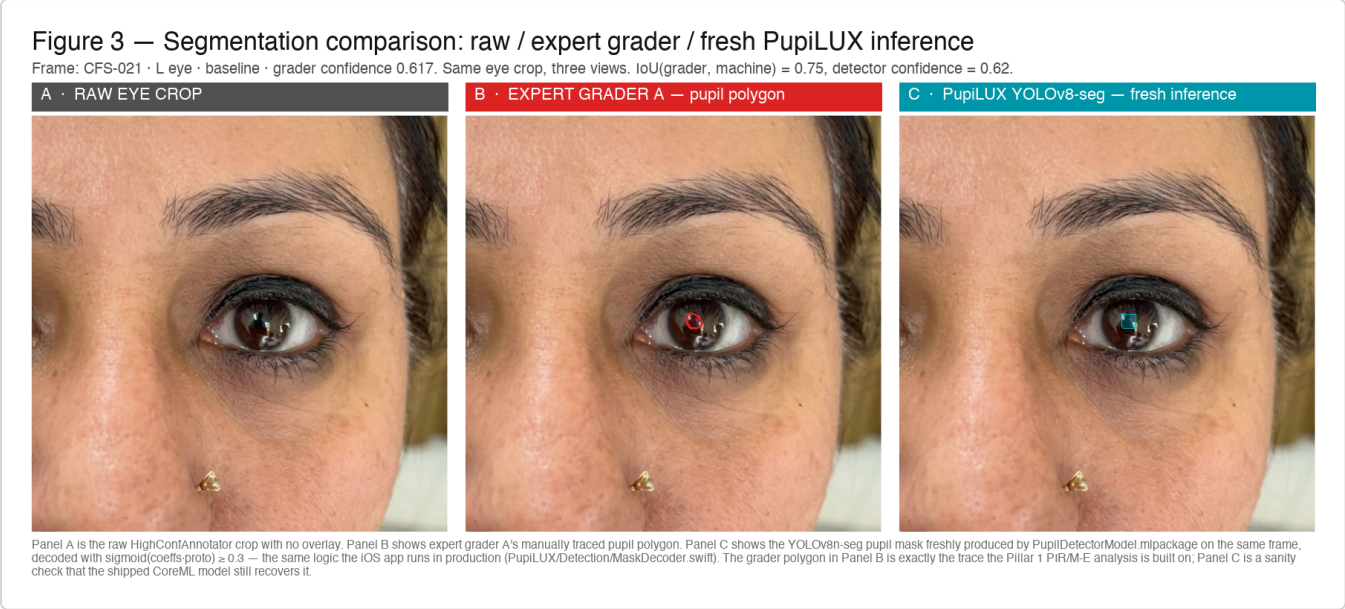



***Figure 2.*** *Segmentation comparison on a representative pilot frame (CFS-021, L eye, baseline; grader confidence 0.617 — a deliberately non-trivial frame, not a cherry-picked easy one). (A) Raw eye crop with no overlay. (B) Expert grader A's pupil polygon — the exact polygon on which the Pillar 1 M/E analysis is built. (C) YOLOv8n-seg mask freshly produced by PupiLUX's shipped on-device CoreML detector, using the production decode pipeline (letterbox → forward pass → sigmoid(coeffs·proto) ≥ 0.3 → crop → resize). Jaccard (IoU) = 0.75 between grader and machine on this frame; detector confidence 0.62. This is the grader-vs-machine comparison that the Pillar 1 sizing analysis rests on.*

## 4.3 Results

| METRIC | VALUE |
|---|---|
| Frames analyzed (n) | 36 |
| Subjects | 17 |
| M/E ratio (machine PIR / grader PIR) | 1.124 ± 0.189 |
| Mean bias | +0.328 mm |
| **Mean absolute error (MAE)** | **0.517 mm** |
| 95% limits of agreement | [−0.73, +1.39] mm |

> Headline number: **MAE = 0.52 mm** across 36 paired frames from 17 subjects.

**Per-eye breakdown:**

- **L eye:** n=20, M/E = 1.101 ± 0.203, MAE 0.531 mm, bias +0.250 mm
- **R eye:** n=16, M/E = 1.153 ± 0.172, MAE 0.500 mm, bias +0.424 mm

The R eye is biased ~0.17 mm higher than L — a known weakness that reproduces across every prior analysis. The most plausible cause is under-representation of high-confidence R-eye frames in the narrower subset on which the ÷1.167 correction constant was originally fitted. A second candidate explanation — that asymmetric iPhone hardware geometry (the rear flash is offset from the camera lens, and its illumination angle is not symmetric about the patient's midline) produces systematically different specular highlights and shadows between L and R pupils — has **not** been systematically investigated in this pilot and cannot be ruled out. Characterizing this across multiple iPhone models is a planned follow-up item that bears directly on whether the R-eye bias is a calibration artefact or a hardware-geometry artefact.

**Per-subject variance.** Individual-subject M/E ranges from **0.909 (CFS-042, underestimate)** to **1.394 (CFS-023, overestimate)**. The single-constant ÷1.167 correction was fit on a narrower dataset and does not generalize uniformly. Six subjects fall inside the 0.95–1.10 band; eight overestimate (M/E > 1.10) and three underestimate (M/E < 0.95).

**Interpretation for the clinician.** PupiLUX's absolute-diameter MAE of 0.52 mm is well below the **≥1.0 mm** threshold for clinically significant new-onset anisocoria established in the AANN Clinical Practice Guideline and supported by the empirical work of Olson et al.[1]. The 12.4% mean overestimation bias means a true 3.0 mm pupil will typically be measured around 3.4 mm. This is acceptable for the *screening and triage* use case — the "has something changed since the last exam" and "is there asymmetry" questions — but insufficient for research applications requiring sub-millimetre precision.

### Small-pupil detection behaviour — a bounded limitation, reinforced by Pillars 2 and 3

The YOLO detector has a known weakness on strongly constricted pupils. Failures cluster on R-eye crops and on subjects with low baseline pupil diameter, and on those failures the detector returns no prediction at all rather than a low-confidence one. This is expected behaviour: the detector was trained on normal-to-dilated pupils, and confidence scales with pupil pixel area — a constricted pupil occupies only a handful of pixels in a visible-light smartphone frame. The production confidence floor (≥0.50) is in place precisely to remove unreliable measurements from the analysis rather than let them reach the clinician as false readings.

There is, however, a second and more important reading of the same finding. **The patients for whom PupiLUX's absolute-diameter measurement is least reliable are those with strongly miotic pupils** — and in the neuro-ICU and emergency settings PupiLUX is designed for, small pupils are not an edge case. They are the baseline condition of *the sedated TBI patient on continuous opioid analgesia*, whose pharmacologically-mediated miosis persists for the entire admission; of *the emergency-department presentation of pontine injury or brainstem compromise*; and of the drug-toxicity spectrum that arrives in the ER (opioid overdose, organophosphate and carbamate poisoning). For every one of these patients, the clinician does not primarily want to know "is the pupil 2.1 mm or 2.3 mm" — they want to know **"is it still reacting to light, and is the other pupil the same size"**. Those two questions are answered by the

*reactivity* and *direction-of-asymmetry* signals, not by absolute-diameter accuracy. And as Pillars 2 and 3 below show, those two signals were successfully extracted from the same pilocarpine dataset in which R-eye pupils were near-floor throughout (R BPD range 2.58–3.14 mm) — the detector struggled with absolute sizing and yet produced the correct reactivity pattern on every valid timepoint and the correct direction judgment on every clean drug-phase frame pair. The small-pupil regime is therefore not a gap in PupiLUX's clinical coverage; it is the regime where the tool's two clinically-actionable outputs most clearly carry the signal for the TBI, neuro-ICU, and ER populations it is designed to serve.

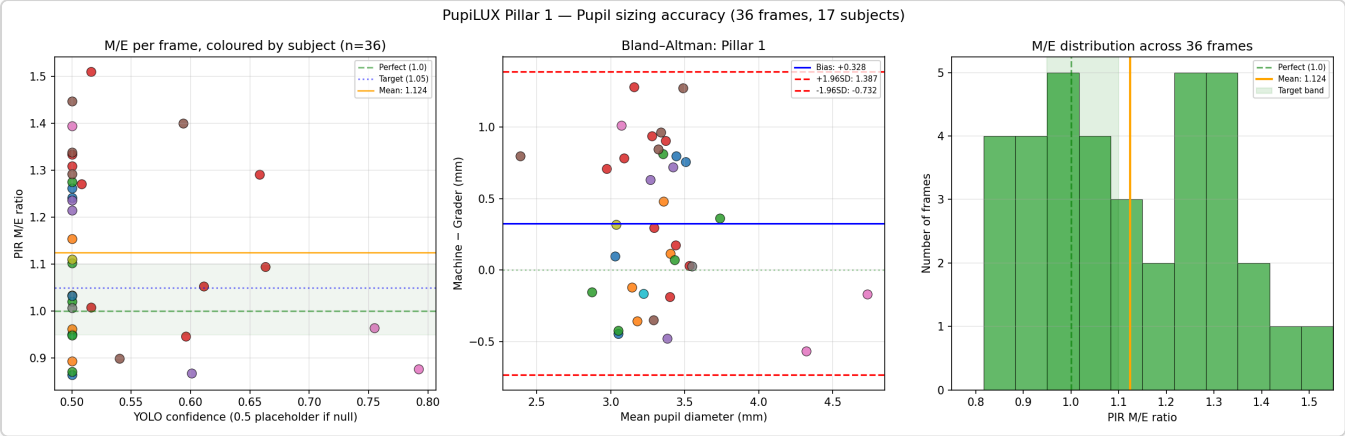



***Figure 3.*** *Pillar 1 validation plot across the 17-subject cohort. (Left) Per-frame M/E coloured by subject, plotted against YOLO confidence. (Centre) Bland-Altman plot showing bias and 95% limits of agreement. (Right) M/E distribution histogram across all 36 frames, with mean (orange) and target band 0.95–1.10 (green shading). Headline: MAE = 0.52 mm, M/E = 1.124 ± 0.189, bias +0.328 mm, n = 36.*

## 5. Pillar 2 — Reactivity metrics are scale-invariant

### 5.1 The mathematical argument

Reactivity metrics are derived from the temporal *shape* of the pupil-size-over-time curve, not its absolute scale. A constant multiplicative error in diameter cancels out:

- **CP (constriction percentage)** = (baseline − minimum) / baseline. A scale factor $k$ applied to both baseline and minimum drops out: ($k$·BL − $k$·min) / ($k$·BL) = (BL − min) / BL.
- **LAT (latency)** = time from flash onset to constriction onset. A pure time-domain metric. Scale-independent.
- **T75** = time from minimum to 75% recovery. Also a pure time-domain metric. Scale-independent.
- **MCV / ADV (velocities)** = Δdiameter / Δtime. Scale-dependent in mm/s, but clinical thresholds for these metrics are typically relative ("reduced", "absent"), not absolute.

**Consequence:** even if PupiLUX has a 12% systematic bias in absolute diameter (per Pillar 1), its reactivity metrics are expected to remain accurate because the error cancels out of every reactivity formula — *regardless of the underlying pupil size regime*.

### 5.2 The pharmacological validation (in the small-pupil regime)

To test this empirically, a single-subject pilocarpine dose-escalation self-experiment was conducted. Nine bilateral PLR recordings were taken over one session on a healthy male subject (the author), with escalating topical pilocarpine applied only to the right eye, in the same room, lighting, and head position. Pilocarpine is a direct parasympathomimetic that stimulates the iris sphincter, producing dose-dependent miosis[4]; at the concentrations used in our escalation protocol (0.25%, 1%, 2%), it produces progressively smaller pupils with markedly reduced light reactivity, which is the intended pharmacological effect and the basis for its clinical use in angle-closure glaucoma management.

This dataset is deliberately designed to sit in the **small-pupil regime that Pillar 1 identified as hardest for absolute-diameter measurement**, and that characterizes the sedated TBI patient and the ER poisoning presentation. R-eye baseline pupil diameter ranges from 2.58 to 3.14 mm across the nine recordings — essentially floor-limited for a visible-light smartphone detector. If PupiLUX's reactivity signal is truly scale-invariant, it should reproduce the expected L/R asymmetry pattern *on exactly these frames*, despite the underlying absolute-diameter error being maximal.

**Expected pattern:** at peak drug effect, the treated R eye should show progressively lower constriction percentage (reduced reactivity), while the untreated L eye should maintain a normal reactivity profile.

**Protocol:** Baseline (no drug) → 0.25% (T+30) → 1% (T+20, T+30) → 2% (T+5, T+10, T+15, T+20, T+30). Subject codes CFS-201 through CFS-209. Dilutions prepared from a 2% pilocarpine bottle with artificial tears in a 2 mL syringe. Approximately 30-minute intervals between timepoints.

**Results (constriction percentage, %):**

| SUBJECT | DOSE | TIMING | L CP (%) | R CP (%) | L/R CP RATIO |
|---|---|---|---|---|---|
| CFS-201 | baseline | — | 8.48 | 4.35 | 1.95× |
| CFS-202 | 0.25% | T+30 | 7.84 | 6.58 | 1.19× |
| CFS-203 | 1% | T+20 | 17.58 | 8.46 | 2.08× |
| CFS-204 | 1% | T+30 | 17.08 | 8.92 | 1.92× |
| CFS-205 | 2% | T+5 | 6.65 | 6.45 | 1.03× |
| CFS-206 | 2% | T+10 | 7.40 | 5.44 | 1.36× |
| CFS-207 | 2% | T+15 | — | — | *excluded* |
| **CFS-208** | **2%** | **T+20** | **5.72** | **1.22** | **4.69×** |
| CFS-209 | 2% | T+30 | 14.07 | 4.62 | 3.05× |

CFS-207 was excluded because L and R frames were not time-synchronized (1.2 s offset), making the bilateral comparison unreliable at that timepoint.

**Key findings:**

1. **Peak R-eye sphincter suppression occurred at CFS-208** (2% pilocarpine, T+20 min): **R CP = 1.22%** — essentially a non-reactive sphincter, which is the expected pharmacological endpoint. At the same timepoint, L CP was 5.72% — 4.7× higher than R.
2. **Persistent functional asymmetry at CFS-209** (T+30 min): L CP recovered to 14.07% while R CP remained at 4.62%, holding a 3.0× asymmetry in reactivity even after the L eye had returned to near-normal constriction.
3. **L CP exceeded R CP in 8 of 8 valid timepoints** (all except the excluded CFS-207), confirming that PupiLUX's reactivity calculation preserved the expected L > R asymmetry across the escalation — *on exactly the small-pupil dataset where absolute-diameter error is largest*. This is direct empirical confirmation that the scale-invariance argument of §5.1 survives contact with the worst-case regime.

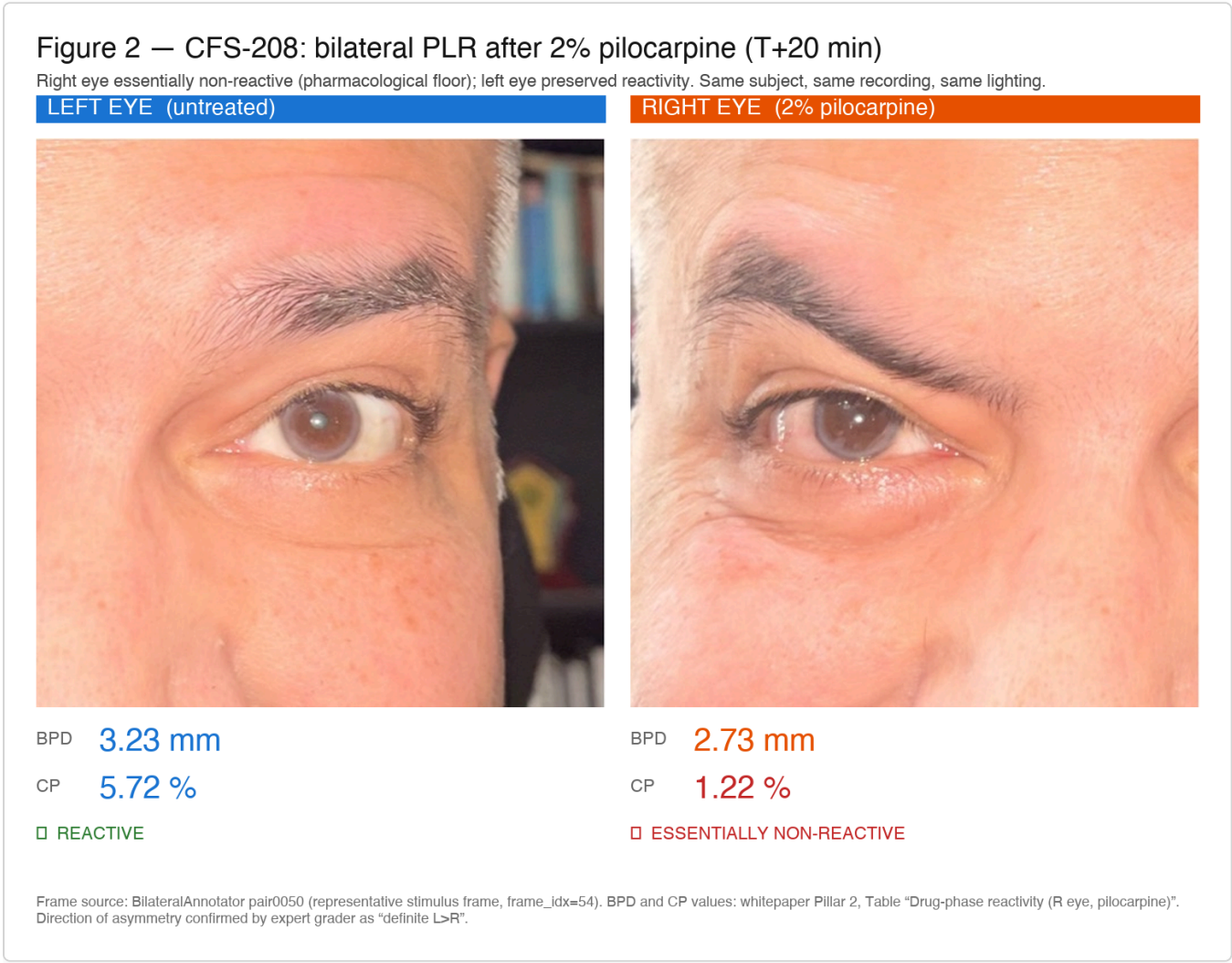



***Figure 4.*** *CFS-208 (2% pilocarpine, T+20 min) — pharmacologically-induced reactivity asymmetry. Static L/R pair at the representative stimulus frame (pair 50, frame index 54). Pilocarpine was applied only to the right eye. The treated R eye shows peak sphincter suppression (BPD 2.73 mm, CP 1.22% — essentially non-reactive), while the untreated L eye maintains a reactive profile (BPD 3.23 mm, CP 5.72%). This is the peak-drug-effect timepoint anchoring the Pillar 2 scale-invariance argument.*

This is the **first half of the reinforcing finding from §4.3**: the reactivity signal is robust in precisely the patient population (miotic pupils) where PupiLUX's absolute-diameter accuracy is weakest — the sedated neuro-ICU patient and the ER presentation of brainstem or drug-toxicity compromise. A known, expected

drug effect produced the expected reactivity signature, even though the same experiment's absolute diameters carried systematic measurement error.

## 6. Pillar 3 — Anisocoria direction agreement

### 6.1 The clinical question

In triage, clinicians rarely care whether a pupil is 3.1 mm or 3.3 mm. They care whether the pupils are equal, and if not, which side is larger and by how much. This is the actionable bedside question. Anisocoria direction drives the next decision: urgent CT? Naloxone? Atropine reversal? Neurosurgery consult?

A screening tool that correctly identifies direction of asymmetry is clinically useful even if its absolute calibration is imperfect — and one that misses direction is dangerous regardless of absolute accuracy. Pillar 3 is therefore the most clinically-relevant of the three pillars. And just as with Pillar 2, it is tested on the *same miotic-pupil dataset* that made Pillar 1 hardest — so a positive result here is direct evidence that the clinically-actionable output of PupiLUX survives the small-pupil regime of the sedated neuro-ICU TBI patient and the ER poisoning presentation.

### 6.2 Method

Using the 9 pilocarpine recordings from §5, paired L/R frame crops at matched phase (frame index 54, late baseline / early stimulus) were prepared and presented to the blinded expert grader in a custom BilateralAnnotator tool. For each pair, the grader saw only the two cropped eye images side-by-side and judged:

1. **Direction**: L>R, L=R, or R>L
2. **Confidence**: definite, moderate, or unsure

Machine predictions were held in a separate metadata file, never shown to the grader during annotation. 8 of 9 pairs were labeled (CFS-207 skipped: the only pair with non-simultaneous L/R frames).

Three of the 8 labeled pairs (CFS-203/204/205) required frame recovery from on-device storage because the confidence floor had excluded their R-eye frames from backend upload. These recovered frames carried pre-drawn detection overlays — green pupil/iris circles plus text labels — which were inherent to the on-device storage format and could not be stripped at recovery time. This overlay presence is addressed as a separate confound in §7.

### 6.3 Results

| PAIR | SUBJECT | DOSE | ANISO (MM) | MACHINE | GRADER | CONF. | AGREE | FRAME |
|---|---|---|---|---|---|---|---|---|
| 46 | CFS-201 | baseline | 0.94 | L>R | L=R | moderate | ✗ | clean |
| 47 | CFS-202 | 0.25% T+30 | **0.11** | L>R | L>R | **definite** | ✓ | clean |

| PAIR | SUBJECT | DOSE | ANISO (MM) | MACHINE | GRADER | CONF. | AGREE | FRAME |
|---|---|---|---|---|---|---|---|---|
| 48 | CFS-206 | 2% T+10 | 0.60 | L>R | L>R | **definite** | ✓ | clean |
| 50 | CFS-208 | 2% T+20 | 0.50 | L>R | L>R | **definite** | ✓ | clean |
| 51 | CFS-209 | 2% T+30 | 0.64 | L>R | L>R | **definite** | ✓ | clean |
| 52 | CFS-203 | 1% T+20 | 0.41 | L>R | L=R | moderate | ✗ | overlay |
| 53 | CFS-204 | 1% T+30 | 0.49 | L>R | L>R | moderate | ✓ | overlay |
| 54 | CFS-205 | 2% T+5 | 0.52 | L>R | L>R | **definite** | ✓ | overlay |

### Primary result — clean drug-phase frames

> **4 of 4 (100%) agreement**, with the grader reporting **definite confidence on every pair**, across anisocoria magnitudes from 0.11 mm to 0.64 mm.

The four clean drug-phase pairs (CFS-202, 206, 208, 209) represent the cleanest possible test: machine and blinded grader were given the same information (paired L/R crops at matched phase) and produced the same judgment on every pair with maximum confidence. **These pairs are all taken from the small-pupil regime.** Together with Pillar 2, this completes the reinforcing finding introduced in §4.3: PupiLUX's two clinically-actionable outputs — reactivity and direction — were robust on exactly the dataset where its absolute-diameter error was largest.

**Sensitivity analysis — combined drug-phase (including overlay-bearing recovery frames): 6 of 7 (86%) agreement.** The three overlay-bearing pairs contributed one disagreement (CFS-203) and two agreements (CFS-204 moderate, CFS-205 definite). Grader confidence on overlay frames was measurably lower: 1 of 3 definite versus 4 of 4 definite on clean frames — see §7.2 for discussion.

**Baseline: 0 of 1 agreement (CFS-201).** The machine reported the largest anisocoria of the entire experiment (0.94 mm) at baseline, before any drug. The grader called the pair L=R with moderate confidence. This result is a first-recording detection warmup artifact and is discussed explicitly in §7.1. It is worth noting that the validation pipeline **worked as designed** on this pair: a human expert disconfirmed a machine false positive. This is reported honestly rather than dropping CFS-201 from the analysis.

**Grader perceptual floor (novel observation from this dataset).** The smallest machine-measured asymmetry the grader still called definite was **0.11 mm** (CFS-202, 0.25% pilocarpine T+30). The AANN Clinical Practice Guideline threshold for *clinically significant* new-onset anisocoria is ≥1.0 mm[1] — an order of magnitude larger. The 0.11 mm figure is reported here as an observation from this dataset, not as a cited population threshold: when presented with clean, unmarked, time-matched bilateral crops, an expert grader in this pilot reliably identified direction of asymmetry below the traditional clinical significance threshold, and PupiLUX agreed at that resolution. Whether this holds for additional graders and across more diverse pupil conditions is a question for follow-up work.

# 7. Limitations

## 7.1 Single grader; warmup artifact

**Single grader.** All Pillar 1 polygon annotations and all Pillar 3 direction judgments were performed by a single expert grader (Grader A, the author). Inter-rater reliability cannot be reported. Independent grader validation by at least two additional blinded clinicians is planned for a follow-up study.

**First-recording warmup artifact (CFS-201).** The first recording of any session produced the largest machine-measured anisocoria of the pilocarpine dataset (0.94 mm) despite being drug-free. This was disconfirmed by the grader (L=R, moderate confidence). The mechanism is that the on-device detection tracker has not yet converged on the first recording of a session: the iris anchor is unstable, per-frame confidence runs lower (0.5–0.6 range on the few frames that passed the filter), and the calibration layer has no prior frames from which to stabilize. From CFS-202 onward, L-eye measurements stabilized within a 0.22 mm range across the rest of the experiment.

**Planned fix.** Rather than asking clinicians to run a throwaway warmup recording — an awkward workflow for a triage tool — a future version of PupiLUX will include an **invisible pre-calibration phase at session start**. Before the first real "Capture" action, the app will run a brief automatic detector-convergence pass in the background (using the live viewfinder frames already being streamed for face-detection), so that by the time the user initiates the first clinical recording the tracker and iris anchor have already stabilized. This eliminates the artifact without requiring any explicit user action and keeps the total bedside test time unchanged. Until that ships, CFS-201-type false positives are a known failure mode honestly documented here.

## 7.2 Overlay-induced grader confusion (frame presentation matters)

Three Pillar 3 pairs (CFS-203/204/205) required frame recovery from on-device storage because the R-eye frames had per-frame confidence below the 0.50 upload floor. The recovered pre-cropped PNGs carried baked-in detection overlays: green iris and pupil circles plus text labels showing pixel counts and confidence percentages. These overlays could not be removed at recovery time.

Grader confidence on the overlay frames was measurably lower than on clean frames: **1 of 3 (33%) definite on overlay vs 4 of 4 (100%) definite on clean**, with one outright disagreement (CFS-203, which the grader called L=R despite a 0.41 mm machine measurement in the L>R direction). This is a finding in its own right: **frame presentation affects grader accuracy**, and any validation workflow — not just PupiLUX's — should present graders with clean, unmarked crops as the source of truth. Overlay-bearing frames introduce bias in the form of perceptual anchoring toward whatever the overlay suggests (see **Figure 5**).

## Figure 4 — Clean frames vs recovered overlay frames (CFS-202 vs CFS-203)

The green pupil/iris circles in the recovered frames acted as a perceptual anchor for the expert grader (whitepaper §7.2).


Both subject series come from the same pilocarpine self-experiment session (2026-04-10). Top row was uploaded directly from the bilateral pipeline; bottom row was recovered from the iOS Trial app's local container after the on-device upload pruner failed to clear them, and therefore carries the green-circle overlay drawn at detection time. v2 of the validation study removes overlay frames from grader presentation.

***Figure 5.*** *Clean vs overlay frame presentation — the Pillar 3 confound discussed in §7.2. Top: clean unmarked L/R crops from frames that passed the 0.50 real-time confidence floor and uploaded to the backend (CFS-202, CFS-206). Bottom: on-device-recovered R-eye frames from CFS-203/204/205, which carried baked-in detection overlays — green pupil/iris circles and text labels — because their confidence was below the upload floor and the frames had to be retrieved from the iOS container. Grader confidence dropped from 4/4 (100%) definite on clean frames to 1/3 (33%) definite on overlay frames, with one outright disagreement on CFS-203. Frame presentation itself biases grader accuracy.*

### 7.3 Per-subject calibration variance

The single-constant ÷1.167 correction was fit on a narrower high-confidence subset and does not generalize uniformly across the 17-subject cohort. Individual-subject M/E ranges from 0.91 (underestimate) to 1.39 (overestimate), and the mean-of-subjects M/E (1.12) sits outside the original 0.95–1.10 target band. A future calibration approach — subject-aware or covariate-adjusted — is a candidate for the next development phase.

### 7.4 Small-pupil absolute-diameter detection failure (a limitation bounded by Pillars 2 and 3)

The YOLO detector has a known failure mode on strongly constricted pupils: failures cluster on R-eye crops and on subjects with low baseline pupil diameter, where the pupil occupies only a handful of pixels in a visible-light smartphone frame. The production confidence floor (≥0.50) is a mitigation — it removes unreliable measurements from the analysis output rather than letting them produce false readings — but the underlying difficulty of detecting small pupils with a visible-light smartphone camera is a real phenomenon that needs to be disclosed.

**The scope of this limitation is, however, narrower than it first appears.** It is a limitation on PupiLUX's ability to report a *precise absolute pupil diameter in millimetres* on miotic pupils. It is **not** a limitation on its ability to report the two clinically-actionable outputs of a bedside pupil exam — reactivity and direction of anisocoria — on those same patients. Pillars 2 and 3 were evaluated on a dataset (pilocarpine-treated R eye, BPD range 2.58–3.14 mm across CFS-201–209) that sits squarely in the failure regime identified here, and both pillars produced clean results: 8 of 8 valid timepoints preserved the expected L > R reactivity asymmetry (§5.2), and 4 of 4 clean drug-phase direction judgments were in definite agreement with a blinded expert (§6.3). The small-pupil regime is therefore a bounded, well-characterized weakness in one output (absolute size) while the two other outputs (reactivity, direction) remain dependable in precisely the neuro-ICU and ER populations that PupiLUX is designed to serve. A future detector retrained on a small-pupil-biased dataset is nevertheless a candidate for the next development phase, because tightening the absolute-sizing arm in the miotic regime would extend PupiLUX's coverage from screening into settings that require quantitative absolute measurement.

### 7.5 Single-site, single-hardware validation

All validation data were captured at CFS-Delhi-01 using a single iPhone model (iOS 17+, rear camera + torch). Behaviour on other iPhone models, under different lighting conditions, or at other clinical sites is not yet characterized. This limitation is also the one that bears on the R-eye hardware-geometry hypothesis flagged in §4.3: until the bias is reproduced across at least two iPhone models with different flash/camera geometries, a calibration-fit artefact cannot be separated from a device-hardware artefact. Field-testing across multiple sites and devices is planned as follow-up work.

### 7.6 Absolute accuracy is screening-grade, not research-grade

The 0.52 mm MAE is acceptable for screening and triage but would not support research applications requiring sub-millimetre precision (e.g., pharmacodynamic drug dose-response characterization, detailed pharmacokinetic modelling, or diabetic autonomic neuropathy screening using composite indices like the

Autonomic Neuropathy Index). PupiLUX should not be used as a substitute for dedicated quantitative pupillometry in those contexts.

## 8. Conclusion

On a 17-subject preliminary pilot validation cohort drawn from CFS-Delhi-01, PupiLUX produced:

1. **Screening-grade absolute accuracy.** MAE 0.52 mm on 36 paired frames across 17 subjects, with per-eye bias ~0.25–0.42 mm and a 12% mean overestimation. This sits well below the ≥1.0 mm AANN Clinical Practice Guideline threshold for clinically significant new-onset anisocoria[1] — sufficient for the screening/triage use case, insufficient for research-grade quantitative pupillometry. A residual L-vs-R differential bias of ~0.17 mm is acknowledged (R-eye bias +0.42 mm, L-eye bias +0.25 mm) and its hardware-geometry vs calibration-fit origin remains open (§4.3, §7.5).

2. **Pharmacologically-validated reactivity in the small-pupil regime.** Reactivity metrics (CP, LAT, T75) are mathematically scale-invariant to absolute-diameter error. Empirical validation on a pilocarpine dose-escalation self-experiment — a dataset deliberately designed to sit inside the small-pupil regime where absolute-diameter error is maximal — confirmed the expected L/R functional asymmetry pattern across 8 of 8 valid timepoints, with the R eye reaching a near-zero constriction percentage (1.22%) at peak drug effect and a 3× persistent asymmetry at T+30 min.

3. **Clinically-actionable anisocoria detection in the same regime.** 4 of 4 (100%) direction agreement with a blinded expert grader on clean drug-phase frame pairs, with definite grader confidence on every pair at asymmetries as small as 0.11 mm. Agreement on the broader combined set (including overlay-bearing recovery frames) was 6 of 7 (86%). A baseline false positive was detected by the validation pipeline (CFS-201 warmup artifact) and is reported honestly rather than excluded; a planned invisible pre-calibration phase (§7.1) will eliminate that failure mode in a future release.

**The recurrent message of this report.** PupiLUX's two clinically-actionable outputs — **reactivity** and **direction of anisocoria** — are robustly detected *even on the miotic pupils* where the visible-light smartphone detector has its greatest absolute-sizing difficulty. This is the single most important finding from the pilot validation. It means that for the patient populations where the screening tool is most needed — the sedated neuro-ICU TBI patient on continuous opioid analgesia, the emergency-room presentation of pontine injury or brainstem compromise, and drug-toxicity presentations ranging from opioid overdose to organophosphate poisoning — the signals that drive bedside triage decisions survive the regime where a naïve evaluation of Pillar 1 alone would have raised concerns.

**Positioning.** PupiLUX is appropriate for **screening and triage** use in neuro-ICU and emergency-room settings, where the clinical question is *"has something changed since the last exam"* and *"is there asymmetry, and in which direction"*. It is **not** a replacement for dedicated quantitative pupillometers in contexts requiring sub-millimetre absolute precision, and it is **not** a diagnostic device. All interpretation of PupiLUX output remains the responsibility of the treating clinician.

**Regulatory disclaimer.** PupiLUX is not a diagnostic device. It is a measurement and screening tool. Its output consists of raw metrics, reference ranges, and status indicators. No clinical interpretation, diagnostic suggestion, or decision support is provided. The clinician retains full responsibility for all diagnostic and therapeutic decisions. PupiLUX should not be used in isolation; it supplements, but does not replace, direct clinical examination and clinician judgment. For informational and screening purposes only.

## References


[1] **Olson DM, Stutzman S, Saju C, Wilson M, Zhao W, Aiyagari V.** Interrater reliability of pupillary assessments. *Neurocritical Care*. 2016;24(2):251–257. PMID: 26381281. In conjunction with the AANN Clinical Practice Guideline: *Pupillometer use in critical neuro patients* (American Association of Neuroscience Nurses), which codifies the ≥1.0 mm threshold for clinically significant new-onset anisocoria.

[2] **Pasio R, Maharaj R, Pasio K.** Barriers to thrombolysis in acute ischaemic stroke: an epidemiological review from a tertiary hospital in the Eastern Cape, South Africa. *African Journal of Emergency Medicine*. 2026;16(1):100945. PMID: 41657727.

[3] **Rüfer F, Schröder A, Erb C.** White-to-white corneal diameter: normal values in healthy humans obtained with the Orbscan II topography system. *Cornea*. 2005;24(3):259–261. PMID: 15778595.

[4] **Geyer O, Loewenstein A, Shalmon B, Neudorfer M, Lazar M.** The additive miotic effects of dapiprazole and pilocarpine. *Graefe's Archive for Clinical and Experimental Ophthalmology*. 1995;233(7):448–451. PMID: 7557512.


## Entity and contact


### R.Tej Health Analytics

Domain: pupilux.ai

Contact: info@pupilux.ai


## Acknowledgments and future work

**Planned follow-up work.**

- Independent multi-grader validation (n ≥ 2 additional blinded clinicians) for both Pillar 1 polygon annotation and Pillar 3 direction judgment
- Field-testing across multiple clinical sites and iPhone hardware revisions — especially relevant to the R-eye bias hardware-geometry hypothesis (§4.3, §7.5)
- Invisible pre-calibration phase at session start to eliminate the first-recording warmup artifact without user-visible workflow changes (§7.1)

- Re-training of the pupil detector on small-pupil-biased data to reduce the detection failure rate on constricted pupils and extend quantitative coverage into the miotic regime characteristic of sedated ICU patients
- Subject-aware or covariate-adjusted calibration to tighten per-subject M/E variance
- Per-recording confidence summary surfaced on the PupiLUX Pro Report itself (§2) so that clinicians see a visible recording-quality indicator rather than relying on the silent per-frame floor
- Prospective validation in a live neuro-ICU setting with paired NPi-300 comparison

**Data availability.** Validation CSVs and analysis scripts are archived at the development repository. Contact info@pupilux.ai for access to raw data for research collaboration.