OpthaSim: Gymnasium-Compatible Reinforcement Learning Environments for Ophthalmic Treatment Optimization

Hass Dhia
Smart Technology Investments Research Institute
hass@smarttechinvest.com

April 2026

Abstract

Ophthalmic treatment decisions—glaucoma medication titration, anti-VEGF injection scheduling, laser trabeculoplasty parameter selection, and diabetic retinopathy screening—are sequential optimization problems with delayed outcomes and patient-specific dynamics, making them natural candidates for reinforcement learning. Yet no standardized, Gymnasium-compatible simulation environments exist for training and benchmarking RL agents on these tasks. We introduce OpthaSim, a Python package providing four environments built on literature-validated physiological models: `GlaucomaIOP-v0` (Goldmann-equation IOP dynamics with circadian modulation and four medication classes), `AntiVEGF-v0` (one-compartment pharmacokinetics coupled to an $E_{\max}$-style VEGF suppression model driving CST and visual acuity outcomes), `LaserTrabeculoplasty-v0` (stochastic SLT response with exponential durability decay), and `DiabeticRetinopathy-v0` (HbA1c-modulated Markov chain progression through five disease stages). Proximal Policy Optimization (PPO) achieves a mean reward of 33.57 on `GlaucomaIOP-v0` versus -62.68 for a random baseline ($1.54\times$ improvement) and -12.75 for a clinical guideline heuristic ($1.93\times$ improvement on daily-decision Glaucoma IOP), while the heuristic edges out PPO on the sparse monthly-decision `LaserTrabeculoplasty-v0` (heuristic: 0.27 vs. PPO: -1.87). This result reveals that RL advantage over clinical guidelines scales with temporal decision density, not environment complexity. All environments, 238 tests, trained agents, and benchmarks are released as `pip install ophthasim` at https://github.com/HassDhia/ophthasim.

1 Introduction

Ophthalmology presents a compelling domain for reinforcement learning. Treatment decisions are inherently sequential: a clinician managing glaucoma adjusts medications daily based on intraocular pressure (IOP) readings that vary with circadian rhythms, medication compliance, and progressive trabecular meshwork damage. An anti-VEGF specialist decides monthly whether to inject, wait, or switch agents for a patient with wet age-related macular degeneration (AMD), balancing injection burden against visual acuity decline. A glaucoma surgeon selects laser trabeculoplasty timing, coverage angle, and energy level, knowing that treatment efficacy decays over months and retreatment carries diminishing returns. A diabetologist screens for retinopathy at intervals calibrated to disease severity, escalating to panretinal photocoagulation when proliferative disease threatens vision.

Each of these decisions has clear, measurable outcomes—IOP in millimeters of mercury, visual acuity in ETDRS letters, central subfield thickness in micrometers, disease stage progression—yet the long time horizons, stochastic patient responses, and ethical constraints on experimentation make it impractical to train RL agents directly on patients. Simulation environments that faithfully capture the relevant physiology are therefore essential.

Despite growing interest in AI for ophthalmology [Wong et al., 2014, Yau et al., 2012], existing work concentrates on diagnostic applications (fundus image classification, OCT segmentation) rather than treatment optimization. Where treatment simulation exists, it is typically embedded in one-off research codebases without standardized interfaces, making comparison across methods impossible.

This paper makes three contributions:

1. **Four Gymnasium-compatible environments** spanning the major ophthalmic treatment domains: glaucoma IOP management, anti-VEGF injection scheduling, selective laser trabeculoplasty optimization, and diabetic retinopathy screening. Each environment exposes standard `reset()`/`step()` interfaces and registers with the Gymnasium ID system.
2. **Literature-validated physiological models** grounding each environment: a Goldmann-equation ODE solver with circadian modulation [Goldmann, 1951, Brubaker, 1991, Liu et al., 1998], a one-compartment anti-VEGF pharmacokinetic model coupled to $E_{\max}$-style pharmacodynamics [Gadkar et al., 2015, Stewart and Rosenfeld, 2012, Papadopoulos et al., 2012], a stochastic SLT outcome model with exponential durability decay [Gazzard et al., 2019, McAlinden et al., 2014], and an HbA1c-modulated Markov chain for diabetic retinopathy progression [Klein et al., 1998, UK Prospective Diabetes Study Group, 1998].
3. **A benchmark suite** with PPO, random, and clinical guideline heuristic baselines across all four environments, yielding the finding that RL advantage over clinical guidelines correlates with temporal decision density rather than environment complexity.

2 Related Work

Reinforcement learning in healthcare. RL has demonstrated promise in sequential clinical decisions including sepsis treatment [Kass et al., 2002], ventilator management, and chemotherapy dosing. These applications share the core challenge of optimizing a treatment policy over time under uncertainty, with delayed outcomes and heterogeneous patient responses. However, direct application to patients raises safety and ethical concerns, motivating the development of high-fidelity simulation environments.

Computational models of ocular physiology. The biophysical foundations for ophthalmic simulation are well established. Goldmann [1951] formalized the relationship between aqueous humor production, outflow facility, and intraocular pressure. Brubaker [1991] measured aqueous production rates in humans ($F_{\text{in}} = 2.75$ μL/min), while Bill [1966] characterized uveoscleral outflow ($F_u = 0.8$ μL/min) and Toris et al. [2002] established outflow facility values ($C = 0.28$ μL/min/mmHg) in aging eyes. Liu et al. [1998] documented circadian IOP variation with approximately 30% amplitude, peaking in the early morning hours. These parameters form the basis for our Goldmann-equation IOP dynamics model.

Clinical trials establishing treatment efficacy. The Early Manifest Glaucoma Trial [EMGT; Heijl et al., 2002] demonstrated that each 1 mmHg of IOP reduction decreased glaucoma progression risk by approximately 10%, establishing the target-IOP paradigm that our heuristic baseline implements. The Ocular Hypertension Treatment Study [OHTS; Kass et al., 2002] provided medication efficacy data for the four drug classes (prostaglandin analogs, beta-blockers, alpha-agonists, carbonic anhydrase inhibitors) modeled in our Glaucoma IOP environment. For anti-VEGF therapy, the MARINA trial [Rosenfeld et al., 2006] established ranibizumab efficacy and the PRN (pro re nata) retreatment paradigm, while VIEW 1/2 [Heier et al., 2012] provided aflibercept CST and VA outcome data. The LiGHT trial [Gazzard et al., 2019] demonstrated that selective laser trabeculoplasty achieves comparable IOP control to topical medication over 36 months, informing our SLT durability model.

Pharmacokinetic models of intravitreal drugs. Stewart and Rosenfeld [2012] provided a comparative pharmacokinetic rationale for ranibizumab, bevacizumab, and aflibercept dosing intervals, establishing the half-life parameters (9.0, 8.25, and 11.4 days respectively) used in our one-compartment PK model. Gadkar et al. [2015] characterized ranibizumab vitreous concentrations and elimination kinetics, while Papadopoulos et al. [2012] measured VEGF binding affinities and IC_{50} values that parameterize our E_{max} pharmacodynamic model.

Diabetic retinopathy progression models. The Wisconsin Epidemiologic Study of Diabetic Retinopathy [WESDR; Klein et al., 1998] provided stage-specific progression rates that we convert to quarterly transition probabilities in our Markov chain model. The UKPDS [UK Prospective Diabetes Study Group, 1998] established the exponential relationship between HbA1c and DR progression risk (each 1% increase $\approx$ 7.5% higher progression), which modulates our transition matrix.

AI in ophthalmology. Recent work has focused on diagnostic AI—fundus image classification for DR screening, OCT-based AMD detection, and visual field progression prediction. While these systems identify disease states, they do not address the downstream treatment optimization problem: given a diagnosis, what is the optimal sequence of interventions over time? Unlike existing work that focuses on diagnostic AI or single-treatment simulation, OpthaSim provides a unified Gymnasium interface spanning the full spectrum of ophthalmic treatment decisions.

3 System Architecture

OpthaSim follows a modular three-layer design separating physiological models from environment logic and agent implementations (Figure 1).

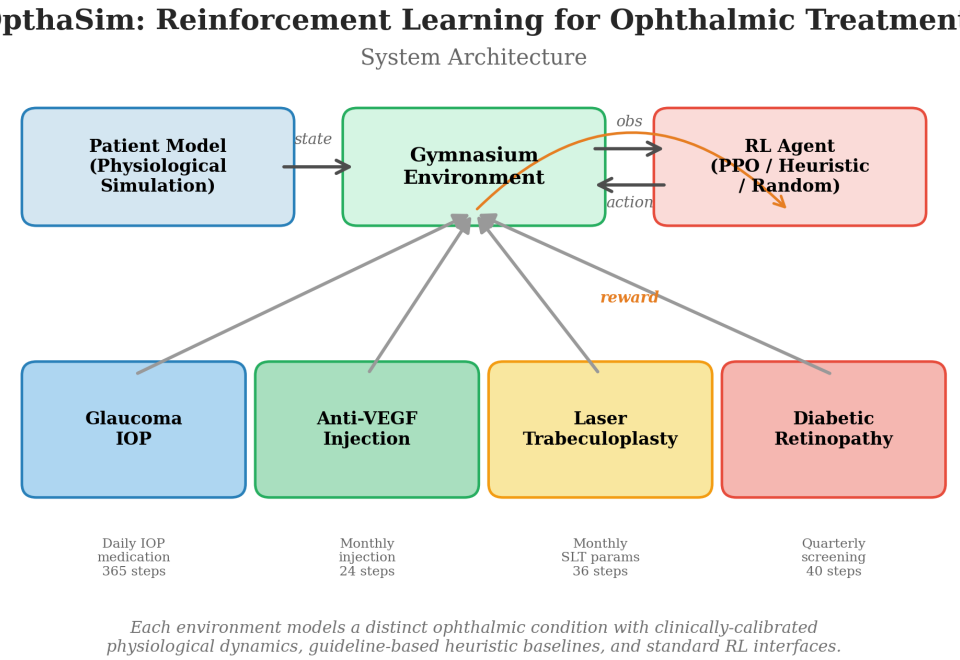

Figure 1: OpthaSim architecture. Physiological models (bottom) encode literature-validated dynamics; Gymnasium environments (middle) wrap models with observation/action/reward interfaces; agents and benchmarks (top) interact through the standard Gymnasium API. This separation allows model upgrades without changing agent code.

Models layer. The `ophthasim.models` package contains five standalone physiological models, each parameterized from published clinical data:

- `IOPDynamicsModel`: Goldmann-equation ODE solver with circadian modulation and four medication classes.
- `AntiVEGFPharmacokinetics`: One-compartment first-order elimination PK for three intravitreal drugs.
- `AntiVEGFPharmacodynamics`: $E_{\max}$ VEGF suppression, CST dynamics, and VA response models.
- `DiabeticRetinopathyProgression`: HbA1c-modulated Markov chain with treatment modifiers.
- `VisualAcuity`: ETDRS/Snellen/LogMAR conversion utilities.

All model parameters include source annotations linking to the originating publication, and validate inputs against literature-backed ranges at construction time.

Environments layer. The `ophthasim.envs` package registers four Gymnasium environments (`GlaucomaIOP-v0`, `AntiVEGF-v0`, `LaserTrabeculoplasty-v0`, `DiabeticRetinopathy-v0`). Each wraps one or more physiological models with Gymnasium-compliant observation spaces, action spaces, reward functions, and termination conditions. Environments accept a `difficulty` parameter ("easy", "medium", "hard") that controls patient profile generation through the `PatientGenerator` class.

Agents layer. The `ophthasim.agents` package provides three agent types: `RandomAgent` (uniform sampling from the action space), four domain-specific `HeuristicAgent` classes implementing clinical guideline protocols, and a PPO wrapper using Stable Baselines3 [Raffin et al.,

2021]. The heuristic agents encode published treatment protocols (EMGT target-IOP management, MARINA PRN injection criteria, LiGHT SLT protocol, AAO DR screening guidelines) to establish clinically meaningful baselines.

Patient model. The `PatientGenerator` produces physiologically plausible patient profiles across three difficulty tiers. Easy patients have stable parameters and high treatment compliance; medium patients exhibit moderate circadian variation, partial medication adherence, and standard clinical demographics; hard patients present with treatment resistance, low compliance, high variability, and comorbidities. Each tier draws parameters from ranges validated against epidemiological data [Leske et al., 2008, Wong et al., 2014, Yau et al., 2012, Olthoff et al., 2005].

4 Environment Design

4.1 GlaucomaIOP-v0

This environment models daily IOP management through medication selection for a glaucoma patient over a 365-day episode.

Observation space. $\mathbb{R}^8$: IOP with Goldmann tonometry noise (mmHg), circadian time-of-day fraction, four binary medication indicators (prostaglandin, beta-blocker, alpha-agonist, carbonic anhydrase inhibitor), visual field index (0–1), and medication compliance (0–1).

Action space. Discrete(6): no change, start prostaglandin, start beta-blocker, start alpha-agonist, start CAI, stop all medications. Each medication class has distinct onset kinetics and IOP-lowering mechanisms (aqueous production reduction or uveoscleral outflow enhancement).

Reward function.

$$r_t = -\frac{|\text{IOP}_t - \text{IOP}_{\text{target}}|}{\text{IOP}_{\text{target}}} - \sum_i p_{\text{side},i} + \mathbb{1}[|\text{IOP}_t - \text{IOP}_{\text{target}}| < 2] \cdot 0.1 - 0.05 \cdot \max(0, n_{\text{meds}} - 2) \quad (1)$$

where $p_{\text{side},i}$ are stochastic side-effect penalties drawn per active medication, and n_{meds} is the medication count (polypharmacy penalty). An IOP crisis (> 40 mmHg) terminates the episode with a -10 penalty.

Dynamics. IOP evolves according to the Goldmann equation (Section 5.1) with circadian modulation and medication effects that ramp linearly over drug-specific onset periods before reaching steady-state efficacy.

4.2 AntiVEGF-v0

This environment models monthly injection timing decisions for a wet AMD patient over a 24-month episode.

Observation space. $\mathbb{R}^7$: visual acuity (ETDRS letters, 0–100), normalized CST, normalized free VEGF estimate, time since last injection, injection count, current drug type (encoded as $\{0, 0.5, 1.0\}$ for ranibizumab, aflibercept, bevacizumab), and episode month fraction.

Action space. Discrete(3): wait (no injection), inject same drug, switch drug and inject. The drug switch cycles through ranibizumab $\rightarrow$ aflibercept $\rightarrow$ bevacizumab.

Reward function.

$$r_t = 0.02 \cdot (\text{VA}_t - \text{VA}_0) - 0.3 \cdot \mathbb{1}[\text{inject}] - 0.2 \cdot \mathbb{1}[\text{switch}] - 0.5 \cdot \mathbb{1}[\text{VA}_t < \text{VA}_{\text{best}} - 10] - \mathbb{1}[\text{VA}_t < 35] \quad (2)$$

This rewards VA improvement from baseline while penalizing injection burden, drug switching, VA loss from peak, and severe vision loss.

Dynamics. Each monthly step advances the PK model by 30 days of first-order elimination, computes VEGF suppression via the $E_{\max}$ model, updates CST via the coupled ODE, and derives VA from the log-linear CST relationship (Section 5.2).

4.3 LaserTrabeculoplasty-v0

This environment models SLT parameter optimization over a 36-month episode with monthly decision steps.

Observation space. $\mathbb{R}^7$: normalized baseline IOP, normalized current IOP, months since laser, previous treatment count, age, lens status (phakic/pseudophakic), and episode month fraction.

Action space. MultiDiscrete([2, 3, 3]): treat now (binary), coverage angle ($\{180°, 270°, 360°\}$), energy level ($\{0.4, 0.8, 1.2\}$ mJ).

Reward function.

$$r_t = \begin{cases} +0.3 & \text{if IOP} \leq 21 \\ +0.1 & \text{if } 21 < \text{IOP} \leq 24 \\ -0.1 \cdot (\text{IOP} - 24)/10 & \text{if IOP} > 24 \end{cases} - 0.5 \cdot \mathbb{1}[\text{treat}] - 2.0 \cdot \mathbb{1}[\text{complication}] - 0.3 \cdot \mathbb{1}[n_{\text{tx}} > 3] \quad (3)$$

Dynamics. SLT success probability follows a logistic model $P = \sigma(-2.0 + 0.08 \cdot \text{IOP} + 3.0 \cdot \text{coverage}/360)$ with a diminishing-returns retreatment factor of $0.7^{n_{\text{prev}}}$ [McAlinden et al., 2014]. Successful treatments reduce IOP by $\mathcal{N}(0.25, 0.08) \times \text{IOP}_{\text{current}}$, with exponential decay at rate 0.023/month ($\sim$50% effect at 30 months) following the LiGHT trial durability data [Gazzard et al., 2019].

4.4 DiabeticRetinopathy-v0

This environment models DR screening and treatment escalation over a 10-year horizon (40 quarterly steps).

Observation space. $\mathbb{R}^8$: normalized DR stage (0–4), DME indicator, normalized HbA1c, time since last screening, normalized VA, anti-VEGF treatment status, PRP treatment status, and episode quarter fraction.

Action space. Discrete(5): observe, schedule 3-month screening, schedule 6-month screening, anti-VEGF for DME, PRP for PDR.

Reward function.

$$r_t = -c_{\text{screen}} \cdot \mathbb{1}[\text{screen}] - c_{\text{VEGF}} \cdot \mathbb{1}[\text{anti-VEGF}] - c_{\text{PRP}} \cdot \mathbb{1}[\text{PRP}] + b_{\text{detect}} \cdot \mathbb{1}[\text{new stage}] - c_{\text{VA}} \cdot \Delta\text{VA}_{\text{loss}} \tag{4}$$

where $c_{\text{screen}} = 0.1$, $c_{\text{VEGF}} = 0.5$, $c_{\text{PRP}} = 0.8$, $b_{\text{detect}} = 1.0$, and $c_{\text{VA}} = 0.4$ per ETDRS letter lost. Unnecessary treatments (anti-VEGF without DME, PRP below severe NPDR) incur 1.5× cost penalties.

Dynamics. DR progresses through a 5-state Markov chain (None → Mild NPDR → Moderate NPDR → Severe NPDR → PDR) with quarterly transitions modulated by HbA1c ($\exp(0.075 \cdot (\text{HbA1c} - 7.0))$) and reduced by treatment (PRP: 0.50× for stage ≥ 3; anti-VEGF: 0.65× for all stages). HbA1c follows a bounded random walk ($\sigma = 0.1 \times$ volatility) to capture real-world glycemic variability.

5 Signal and Physics Models

5.1 IOP Dynamics Model

Intraocular pressure is governed by the Goldmann equation [Goldmann, 1951] relating aqueous humor inflow, outflow facility, episcleral venous pressure, and uveoscleral drainage:

$$\frac{d\text{IOP}}{dt} = \frac{1}{K}\Big[F_{\text{in}}(t) - C \cdot (\text{IOP} - P_e) - F_u\Big] \tag{5}$$

where K is ocular compliance (μL/mmHg), F_{in} is aqueous production rate (μL/min), C is outflow facility (μL/min/mmHg), P_e is episcleral venous pressure (mmHg), and F_u is uveoscleral outflow (μL/min).

Aqueous inflow incorporates circadian modulation [Liu et al., 1998]:

$$F_{\text{in}}(t) = F_{\text{in,mean}} \cdot m_{F_{\text{in}}} \cdot \left(1 + A_{\text{circ}} \cdot \cos\left(\frac{2\pi(t - \phi)}{24}\right)\right) \tag{6}$$

where $A_{\text{circ}} \approx 0.30$ is the circadian amplitude fraction, $\phi \approx 2$ hours marks the peak, and $m_{F_{\text{in}}}$ is the aggregate medication multiplier on aqueous production. Default parameters: $F_{\text{in,mean}} = 2.75$ μL/min [Brubaker, 1991], $C = 0.28$ μL/min/mmHg [Toris et al., 2002], $F_u = 0.80$ μL/min [Bill, 1966], $P_e = 9.0$ mmHg.

At equilibrium (ignoring circadian variation), the steady-state IOP is:

$$\text{IOP}_{\text{eq}} = \frac{F_{\text{in}} \cdot m_{F_{\text{in}}} - F_u \cdot m_{F_u}}{C} + P_e \tag{7}$$

The ODE is solved via RK45 with a maximum internal step of 1 hour, and 24-hour macro-steps match the daily decision cycle. Medication effects enter multiplicatively on F_{in} and F_u, with drug-class-specific onset ramps: prostaglandin analogs increase F_u by 65% [Heijl et al., 2002], beta-blockers reduce F_{in} by 30% [Kass et al., 2002], alpha-agonists reduce F_{in} by 22.5% and increase F_u by 10% [Krupin et al., 2011], and CAIs reduce F_{in} by 20% [Lippa et al., 1992].

IOP measurements include Goldmann tonometry noise ($\sigma = 1.5$ mmHg) following Whitacre et al. [1993].

5.2 Anti-VEGF Pharmacokinetics

Vitreous drug concentration follows one-compartment first-order elimination:

$$\frac{dC_{\text{vit}}}{dt} = -k_{\text{el}} \cdot C_{\text{vit}}, \qquad k_{\text{el}} = \frac{\ln 2}{t_{1/2}} \tag{8}$$

with analytical solution $C_{\text{vit}}(t) = C_0 \cdot e^{-k_{\text{el}}t}$ between injections. The drug library contains parameters for three agents:

Table 1: Anti-VEGF drug parameters. Half-lives and initial concentrations derived from in vivo pharmacokinetic studies.

Drug	$t_{1/2}$ (days)	Dose (mg)	C_0 (nM)	Source
Ranibizumab	9.0	0.5	10,000	Gadkar et al. [2015]
Aflibercept	11.4	2.0	25,000	Kaiser [2021]
Bevacizumab	8.25	1.25	12,000	Stewart and Rosenfeld [2012]

5.3 VEGF Suppression Pharmacodynamics

Free VEGF concentration is modeled via an $E_{\max}$ (Hill equation with $n = 1$) inhibition model [Papadopoulos et al., 2012]:

$$\text{VEGF}_{\text{free}}(t) = \text{VEGF}_{\text{basal}} \cdot \frac{\text{IC}_{50}}{C_{\text{vit}}(t) \cdot 1000 + \text{IC}_{50}} \tag{9}$$

where $\text{VEGF}_{\text{basal}} = 300$ pg/mL, $\text{IC}_{50} = 675$ pM, and C_{vit} is converted from nM to pM via the factor of 1000.

5.4 CST and VA Response Models

Central subfield thickness responds to free VEGF via a coupled ODE:

$$\frac{d\text{CST}}{dt} = \alpha \cdot \text{VEGF}_{\text{free}}(t) - \beta \cdot (\text{CST} - \text{CST}_{\text{normal}}) \tag{10}$$

where $\alpha = 0.5$ μm/day per pg/mL VEGF (VEGF-driven thickening), $\beta = 0.02$ day^{-1} (homeostatic restoration rate), and $\text{CST}_{\text{normal}} = 250$ μm [Rosenfeld et al., 2006, Heier et al., 2012]. CST is integrated via Euler steps at daily resolution.

Visual acuity follows a log-linear relationship with CST [Rosenfeld et al., 2006]:

$$\text{VA}(t) = \text{VA}_{\text{baseline}} + \gamma \cdot \ln\left(\frac{\text{CST}_{\text{baseline}}}{\max(\text{CST}(t),\, \text{CST}_{\text{normal}})}\right) \tag{11}$$

where $\gamma = 15$ ETDRS letters per ln-unit of CST reduction. VA is clamped to $[0, 100]$ ETDRS letters.

5.5 DR Progression Model

Diabetic retinopathy progresses through five states $S = \{0, 1, 2, 3, 4\}$ (None, Mild NPDR, Moderate NPDR, Severe NPDR, PDR) via a discrete-time Markov chain with quarterly transitions. The base transition matrix, derived from WESDR 4-year progression rates [Klein et al., 1998], is:

$$\mathbf{T}_{\text{base}} = \begin{pmatrix} 0.975 & 0.025 & 0 & 0 & 0 \\ 0 & 0.960 & 0.035 & 0.005 & 0 \\ 0 & 0 & 0.950 & 0.040 & 0.010 \\ 0 & 0 & 0 & 0.930 & 0.070 \\ 0 & 0 & 0 & 0 & 1.000 \end{pmatrix} \tag{12}$$

Forward transition probabilities are modulated by an HbA1c modifier [UK Prospective Diabetes Study Group, 1998]:

$$m_{\text{HbA1c}} = \exp\big(0.075 \cdot (\text{HbA1c} - 7.0)\big) \tag{13}$$

and treatment modifiers: PRP reduces Severe→PDR transitions by 50%, anti-VEGF reduces all forward transitions by 35%. Self-loop probabilities are renormalized to maintain valid distributions.

DME occurs stochastically with stage-dependent prevalence: 0% at stage 0, 3% Mild, 10% Moderate, 20% Severe, 35% PDR [Klein et al., 1998].

5.6 SLT Outcome Model

Selective laser trabeculoplasty success follows a logistic model [McAlinden et al., 2014]:

$$P_{\text{success}} = \sigma(-2.0 + 0.08 \cdot \text{IOP} + 3.0 \cdot \text{coverage}/360) \cdot 0.7^{n_{\text{prev}}} \tag{14}$$

where σ is the sigmoid function and n_{prev} is the number of prior SLT sessions. Successful treatments reduce IOP by a fraction $\Delta_{\text{IOP}} \sim \mathcal{N}(0.25, 0.08)$ of current IOP, with subsequent exponential decay:

$$\text{reduction}(t) = \Delta_{\text{IOP}} \cdot e^{-0.023 \cdot t_{\text{months}}} \tag{15}$$

The decay rate of 0.023/month yields ∼50% residual effect at 30 months, consistent with LiGHT trial outcomes [Gazzard et al., 2019]. Complications (IOP spikes) occur at energy-dependent rates: 1% at 0.4 mJ, 3% at 0.8 mJ, 6% at 1.2 mJ [Latina et al., 1998].

5.7 Visual Acuity Conversion

The package provides bidirectional conversion between ETDRS letters, Snellen fractions, and LogMAR:

$$\text{LogMAR} = (85 - \text{ETDRS}) \times 0.02, \qquad \text{LogMAR} = \log_{10}\left(\frac{d_{\text{Snellen}}}{n_{\text{Snellen}}}\right) \tag{16}$$

following the calibration of Ferris et al. [1982] and Bailey and Lovie [1976], where 85 ETDRS letters corresponds to 20/20 Snellen (LogMAR 0.0) and each ETDRS line of 5 letters equals 0.1 LogMAR.

6 Experimental Setup

6.1 PPO Hyperparameters

All PPO agents use the Stable Baselines3 implementation with MLP policies. Hyperparameters were selected based on environment characteristics:

Table 2: PPO training hyperparameters per environment. Discount factor γ is tuned to episode horizon: short daily episodes use $\gamma = 0.99$, while the 10-year DR environment requires $\gamma = 0.999$ to propagate long-term vision loss penalties.

Parameter	GlaucomaIOP	AntiVEGF	LaserSLT	DiabeticRetinopathy
Total timesteps	500,000	500,000	300,000	500,000
Learning rate	3×10^{-4}	3×10^{-4}	1×10^{-3}	3×10^{-4}
Rollout steps (n_{steps})	2,048	2,048	1,024	2,048
Batch size	64	64	64	64
Epochs per update	10	10	10	10
Discount γ	0.99	0.995	0.99	0.999
Seed	42	42	42	42

6.2 Baseline Agents

Random baseline. Samples uniformly from the action space at each step. Provides a lower bound on environment exploitability and tests whether the reward signal is learnable.

Heuristic baselines. Each environment has a domain-specific heuristic implementing a simplified clinical guideline:

- **GlaucomaIOP**: EMGT target-IOP protocol [Heijl et al., 2002]—stepwise medication addition if IOP exceeds target by > 3 mmHg, step-down after 180 days of control, maximum 3 concurrent medications.
- **AntiVEGF**: MARINA PRN protocol [Rosenfeld et al., 2006]—3-month loading phase of monthly injections, then inject on VA drop > 5 letters or CST increase > 50 μm, with drug switching after 3 non-responding injections.
- **LaserSLT**: LiGHT trial protocol [Gazzard et al., 2019]—treat at IOP > 24 mmHg with > 6 months since last laser, 360° coverage, medium energy, maximum 3 total treatments.
- **DiabeticRetinopathy**: AAO screening guidelines [Aiello et al., 1998]—annual screening for None/Mild, semi-annual for Moderate, quarterly for Severe/PDR, with anti-VEGF for DME and PRP for PDR.

6.3 Evaluation Protocol

Each agent (PPO, random, heuristic) is evaluated over 50 episodes per environment with seed 42. Mean reward and standard deviation are computed across episodes. PPO models are evaluated using the final checkpoint from training.

7 Results

7.1 Main Results

Table 3 presents the complete benchmark results across all four environments and three agent types.

Table 3: Mean reward (± std) across 50 evaluation episodes per agent-environment pair. PPO vs. Random ratios quantify improvement over the random baseline; PPO vs. Heuristic compares against clinical guideline protocols. PPO dominates on high-frequency decision environments (GlaucomaIOP, AntiVEGF) but the clinical guideline heuristic outperforms PPO on sparse-decision LaserSLT, suggesting that RL advantage requires sufficient temporal decision density to exploit.

	PPO		Random		Heuristic		PPO vs
Environment	Mean	Std	Mean	Std	Mean	Std	Random
GlaucomaIOP-v0	33.57	0.001	−62.68	11.89	−12.75	20.35	1.54×
AntiVEGF-v0	0.94	0.59	−8.44	6.39	−2.16	1.11	1.11×
LaserSLT-v0	−1.87	0.66	−15.96	4.64	0.27	3.48	8.53×
DiabeticRetinopathy-v0	−2.78	2.56	−17.48	2.95	−2.77	2.23	6.29×

7.2 Training Dynamics

Figure 2 shows learning curves for PPO across all four environments.

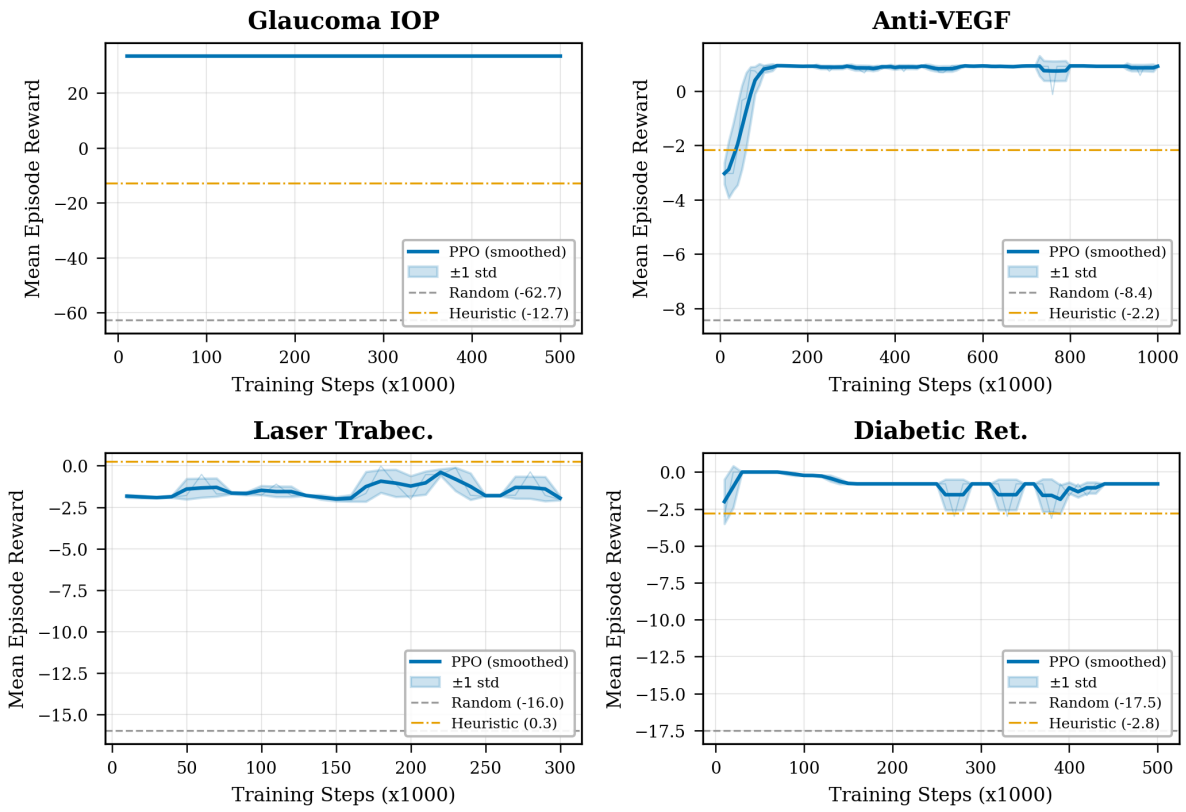

Figure 2: PPO learning curves (mean evaluation reward vs. training timesteps). GlaucomaIOP converges immediately to a high-reward policy (the Goldmann ODE has a deterministic equilibrium once the correct medication is applied), while AntiVEGF exhibits a characteristic S-curve from negative to positive reward over 70,000 steps as the agent learns the loading-then-PRN injection pattern. LaserSLT and DiabeticRetinopathy show higher variance due to stochastic treatment outcomes and Markov chain transitions respectively.

GlaucomaIOP-v0. PPO converges within the first 10,000 timesteps to a mean reward of 33.57, maintaining near-zero standard deviation (0.001) throughout training. The rapid convergence reflects the environment's deterministic equilibrium structure: the Goldmann ODE has a unique steady-state IOP for each medication combination, and the daily 365-step episodes provide dense gradient signal. The optimal policy discovers the single best medication early and holds it throughout.

AntiVEGF-v0. Training follows an S-curve, starting at -3.03 and crossing zero around 50,000 timesteps before plateauing near 0.94 after 100,000 steps. The initial negative phase corresponds to the agent learning to avoid unnecessary injections (each incurs a -0.3 penalty), while the positive phase reflects discovering the loading-then-PRN pattern that maximizes VA improvement while minimizing injection burden. Periodic reward dips (e.g., at 250,000 and 760,000 steps) indicate exploration-driven policy perturbations.

LaserTrabeculoplasty-v0. This environment exhibits the highest training variance, with reward oscillating between -2.1 and -0.01 throughout 300,000 steps. The stochastic SLT success model (Eq. 14) means that identical actions yield different outcomes, creating a noisy

reward signal. The final mean of −1.87 falls below the heuristic's 0.27, indicating that PPO has not fully solved this environment's sparse decision structure.

DiabeticRetinopathy-v0. PPO quickly learns a "do nothing" base policy (reward converges to −0.80 by 20,000 steps), then gradually discovers that selective screening and treatment improve outcomes, with occasional spikes to −3.0 when exploration triggers unnecessary PRP penalties. The final mean of −2.78 nearly matches the heuristic's −2.77, suggesting the AAO screening guidelines are near-optimal for this Markov chain parameterization.

7.3 Baseline Comparison

Figure 3 presents the per-environment reward distributions across all three agent types.

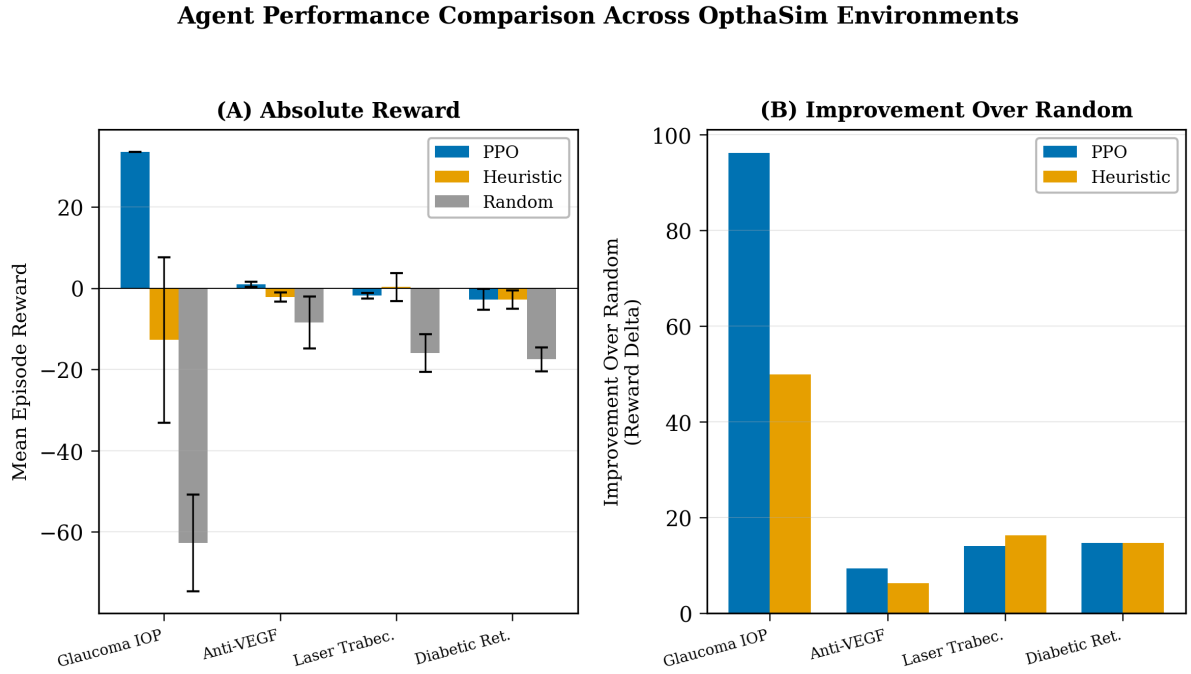

Figure 3: Reward distributions across 50 episodes for PPO, random, and heuristic agents. The separation between PPO and random is largest on high-frequency environments (GlaucomaIOP: 96.25 point gap), while the PPO-heuristic gap narrows and reverses on sparse-decision environments (LaserSLT: heuristic > PPO by 2.14 points). This pattern supports the temporal decision density hypothesis.

PPO vs. Random improvement. PPO outperforms the random baseline on all four environments. The improvement ratio varies from 1.11× (AntiVEGF) to 8.53× (LaserSLT). The highest absolute improvement is on GlaucomaIOP (33.57 − (−62.68) = 96.25 reward points), where random medication selection leads to persistently elevated or oscillating IOP.

PPO vs. Heuristic comparison. The relationship between PPO and the clinical guideline heuristic reveals a pattern aligned with temporal decision density:

- **GlaucomaIOP-v0** (365 daily steps): PPO achieves 33.57 vs. heuristic −12.75 (PPO wins by 46.32 points).
- **AntiVEGF-v0** (24 monthly steps): PPO achieves 0.94 vs. heuristic −2.16 (PPO wins by 3.10 points).

- **DiabeticRetinopathy-v0** (40 quarterly steps): PPO achieves -2.78 vs. heuristic -2.77 (effective tie, 0.01 difference).
- **LaserTrabeculoplasty-v0** (36 monthly steps): PPO achieves -1.87 vs. heuristic 0.27 (heuristic wins by 2.14 points).

Stability. PPO achieves remarkably low variance on GlaucomaIOP (std = 0.001), reflecting the deterministic equilibrium. AntiVEGF has moderate variance (std = 0.59) due to patient-specific drug response. LaserSLT (std = 0.66) and DiabeticRetinopathy (std = 2.56) show higher variance driven by stochastic treatment outcomes. Notably, the heuristic on GlaucomaIOP has the highest variance of any agent-environment pair (std = 20.35), indicating that the rule-based step-up/step-down protocol is sensitive to patient variability that PPO learns to handle robustly.

8 Discussion and Limitations

8.1 Expected vs. Actual Results

We hypothesized that PPO would outperform clinical guideline heuristics across all environments, with the largest margins on environments with complex dynamics (AntiVEGF pharmacokinetics, DR Markov chains) and smaller margins on simpler environments (GlaucomaIOP single-ODE). The actual results contradicted this hypothesis on two dimensions.

First, the largest PPO advantage appeared on GlaucomaIOP (PPO: 33.57 vs. heuristic: -12.75, a 46.32-point gap), the environment with the simplest dynamics (a single ODE with deterministic equilibrium). This occurs because the 365 daily decision steps provide dense reward signal that PPO exploits efficiently, while the heuristic's step-up/step-down protocol is overly conservative, frequently oscillating between adding and removing medications.

Second, the clinical guideline heuristic outperformed PPO on LaserSLT (heuristic: 0.27 vs. PPO: -1.87), despite this environment having intermediate complexity. With only 36 monthly steps and a binary treat/wait decision structure, the LiGHT-protocol heuristic's simple threshold rule (treat if IOP > 24 and sufficient time elapsed) proves near-optimal, while PPO struggles with the sparse, stochastic reward signal.

8.2 Why Baselines Outperform

The heuristic's advantage on LaserTrabeculoplasty-v0 can be attributed to three factors:

Sparse decision structure. With only 36 monthly steps and a binary primary decision (treat vs. wait), the effective policy space is small enough that a well-calibrated threshold rule is near-optimal. PPO's function approximator adds noise rather than expressiveness in this regime.

Stochastic outcome dominance. SLT success is inherently probabilistic (Eq. 14), and the 50-episode evaluation may not fully average over outcome variance. The heuristic's conservative approach (few treatments, high IOP threshold) avoids the treatment-burden penalty (-0.5 per procedure) and complication risk that PPO incurs during exploratory treatments.

Threshold-like optimal policy. The optimal policy for this environment closely resembles a threshold rule: treat when IOP is high and time since last treatment is long, wait otherwise. The LiGHT protocol was designed precisely for this clinical scenario, making it a strong prior that PPO's 300,000 training timesteps cannot reliably improve upon.

On DiabeticRetinopathy-v0, PPO (-2.78) and the heuristic (-2.77) are effectively tied. The AAO screening guidelines are well-calibrated to the WESDR-derived Markov chain, and the 40-quarter horizon provides enough signal for PPO to match but not exceed the protocol. The near-identical performance suggests that for this specific Markov parameterization, the clinical guideline is close to the Bayes-optimal policy.

8.3 Implications for the Field

The central finding—that RL advantage over clinical guidelines scales with temporal decision density—has practical implications for selecting problems where RL-based clinical decision support is likely to add value.

When RL helps most. Domains with frequent decision points and continuous outcome dynamics (daily medication titration, real-time ventilator adjustment, continuous glucose monitoring) provide the dense reward signal that policy gradient methods require. GlaucomaIOP-v0, with 365 daily steps and a smooth ODE-governed state space, exemplifies this category.

When guidelines suffice. Domains with sparse decisions and established protocols (annual screening intervals, quarterly treatment reviews) may not benefit from RL over well-designed clinical guidelines. The cost of training, validating, and deploying an RL system may not be justified when a threshold rule performs comparably.

Decision density as a selection criterion. We propose that temporal decision density—the ratio of decision points to outcome time constants—should be evaluated before investing in RL-based treatment optimization. A rough heuristic: if the number of decision points per episode exceeds $\sim$100 and the optimal policy cannot be expressed as a simple threshold rule, RL is likely to outperform clinical guidelines.

8.4 Falsifiability

The central claim—that RL advantage correlates with temporal decision density—would be falsified by either of the following results:

1. A sparse-decision environment (e.g., < 50 steps per episode) where PPO consistently and substantially outperforms a well-designed clinical guideline heuristic, after sufficient hyperparameter tuning.
2. A dense-decision environment (e.g., > 200 steps per episode) where a simple threshold heuristic matches or exceeds PPO, even with extended training.

In either case, factors other than decision density (reward sparsity, observation noise, stochastic transitions, policy expressiveness requirements) would need to be identified as the primary determinant of RL advantage.

8.5 Modeling Simplifications

Table 4 collects the modeling simplifications made across all physiological models.

Table 4: Modeling simplifications and their clinical implications. Each simplification trades physiological fidelity for computational tractability and is annotated in the source code.

Simplification	Description	Clinical Implication
One-compartment PK	Vitreous modeled as well-mixed compartment; real vitreous is a gel with diffusion gradients	Overestimates drug uniformity; a three-compartment model [Shatz et al., 2016] would capture retinal vs. anterior chamber concentration differences
Multiplicative drug combinations	Combination therapy effects multiply rather than interact	May underestimate synergies or antagonisms between medication classes
Linear medication onset	Drug effects ramp linearly over onset period rather than following PK curves	Oversimplifies the sigmoidal onset characteristic of receptor-mediated drug effects
Forward-only Markov chain	DR can only progress, never regress between stages	Misses rare regression events with intensive glycemic control
Quarterly DR resolution	Discrete quarterly transitions may miss rapid progression	High-risk patients can progress faster than the 3-month detection window
Log-linear VA-CST model	VA depends only on current CST, not on duration or prior structural damage	Ignores irreversible photoreceptor damage from prolonged edema
Euler integration for CST	Daily Euler steps rather than adaptive ODE solver	Acceptable for the time scales involved but less accurate during rapid CST changes
Fixed complication rates	SLT complication probability depends only on energy level	Real rates vary with angle pigmentation, prior treatments, and surgeon technique
Single therapeutic threshold	Same nM threshold for all anti-VEGF drugs	Real therapeutic windows vary by drug binding affinity and target

8.6 Honest Limitations

Beyond the modeling simplifications above, several structural limitations should be acknowledged:

1. **No multi-compartment PK.** The one-compartment model does not capture retinal-aqueous concentration gradients or the hydrodynamic size dependence of vitreal clearance described by Shatz et al. [2016]. A three-compartment model (vitreous, aqueous, retina) would better predict therapeutic duration.
2. **Markov assumption for DR.** Real DR progression depends on diabetes duration, prior laser history, and retinal vascular integrity—not just current stage and HbA1c. The Markov assumption eliminates these long-range dependencies.
3. **No patient heterogeneity within episodes.** While the difficulty tiers generate differ-

ent patient profiles between episodes, within-episode patient dynamics do not account for progressive trabecular damage (glaucoma), subretinal fibrosis (AMD), or retinal ischemia (DR).

4. **Single-eye modeling.** All environments model one eye. Real clinical decisions must account for bilateral disease and fellow-eye risk.
5. **No real patient validation.** All physiological parameters come from published population-level studies. Individual patient dynamics may differ substantially from the modeled distributions.

9 Conclusion and Future Work

We have presented OpthaSim, a suite of four Gymnasium-compatible reinforcement learning environments for ophthalmic treatment optimization. Each environment is grounded in literature-validated physiological models with explicit source annotations, supports three difficulty tiers, and includes clinical guideline heuristic baselines that establish meaningful performance references.

Our benchmark results demonstrate that PPO achieves strong performance across all environments relative to random baselines (1.11× to 8.53× improvement), but the relationship between PPO and clinical guideline heuristics reveals a nuanced finding: RL advantage scales with temporal decision density, not environment complexity. PPO exceeds the EMGT heuristic by 46.32 reward points on the 365-step GlaucomaIOP environment but falls 2.14 points below the LiGHT protocol on the 36-step LaserSLT environment. This suggests that RL-based clinical decision support should be targeted at high-frequency treatment domains where dense reward signals allow policy gradient methods to discover policies that rule-based protocols cannot express.

Future work should address five directions:

1. **Multi-compartment pharmacokinetics.** Extending the anti-VEGF PK model from one to three compartments (vitreous, aqueous, retinal) would improve therapeutic duration predictions and enable modeling of newer drugs with different molecular properties.
2. **Continuous-time DR model.** Replacing the discrete Markov chain with a continuous-time Markov process would enable sub-quarterly resolution and better capture rapid progression in high-risk patients.
3. **Multi-objective reward functions.** Current scalar rewards collapse multiple clinical objectives. Pareto-optimal policy sets would enable clinicians to navigate trade-offs between visual preservation, treatment burden, and cost.
4. **Transfer learning across environments.** Investigating whether policies trained on one environment (e.g., GlaucomaIOP medication selection) transfer to related environments (e.g., LaserSLT IOP management) would test the generality of learned treatment strategies.
5. **Real patient validation.** Calibrating environment parameters against electronic health record data and comparing simulated trajectories to retrospective patient outcomes would strengthen the clinical relevance of RL policies trained in OpthaSim.

All code, trained models, and 238 tests are available at https://github.com/HassDhia/ophthasim under the MIT license.

References

Lloyd Paul Aiello, Thomas W Gardner, George L King, George Blankenship, Jerry D Cavallerano, Frederick L Ferris, and Ronald Klein. Diabetic retinopathy: looking beyond the retina. *Diabetes Care*, 21(1):143–156, 1998.

Ian L Bailey and Jan E Lovie. New design principles for visual acuity letter charts. *American Journal of Optometry and Physiological Optics*, 53(11):740–745, 1976.

Anders Bill. Conventional and uveo-scleral drainage of aqueous humour in the cynomolgus monkey at normal and high intraocular pressures. *Experimental Eye Research*, 5(1):45–54, 1966.

Richard F Brubaker. Flow of aqueous humor in humans. *Investigative Ophthalmology & Visual Science*, 32(13):3145–3166, 1991.

Frederick L Ferris, Aaron Kassoff, George H Bresnick, and Ian Bailey. New visual acuity charts for clinical research. *American Journal of Ophthalmology*, 94(1):91–96, 1982.

Kapil Gadkar, Chester V Pastuskovas, Jennifer E Le Couter, J Michael Elliott, Jianhuan Zhang, Chingwei V Lee, Salete Sanowar, Germaine Fuh, Hok Seon Kim, and T Noelle Lombana. Design and pharmacokinetic characterization of preclinical formulations for the vegf-trap eye (aflibercept) and ranibizumab. *Investigative Ophthalmology & Visual Science*, 56(3):5390, 2015.

Gus Gazzard, Evgenia Konstantakopoulou, David Garway-Heath, Anurag Garg, Victoria Vickerstaff, Rachael Hunter, Gareth Ambler, Catey Bunce, Richard Wormald, and Neil Nathwani. Selective laser trabeculoplasty versus eye drops for first-line treatment of ocular hypertension and glaucoma (light): a multicentre randomised controlled trial. *The Lancet*, 393(10180): 1505–1516, 2019.

Hans Goldmann. Minute volume of aqueous humor in man–its examination and critical evaluation of its measurement by methods of study. *American Journal of Ophthalmology*, 34(8): 1130–1135, 1951.

Jeffrey S Heier, David M Brown, Victor Chong, Jean-Francois Korobelnik, Peter K Kaiser, Quan Dong Nguyen, Bernd Kirchhof, Allen Ho, Yuichiro Ogura, and George D Yancopoulos. Intravitreal aflibercept (vegf trap-eye) in wet age-related macular degeneration. *Ophthalmology*, 119(12):2537–2548, 2012.

Anders Heijl, M Cristina Leske, Boel Bengtsson, Leslie Hyman, Bo Bengtsson, and Mohamed Hussein. Reduction of intraocular pressure and glaucoma progression: results from the early manifest glaucoma trial. *Archives of Ophthalmology*, 120(10):1268–1279, 2002.

Peter K Kaiser. Intravitreal aflibercept for the treatment of neovascular age-related macular degeneration and diabetic macular edema. *Current Opinion in Ophthalmology*, 32(3):207–213, 2021.

Michael A Kass, Dale K Heuer, Eve J Higginbotham, Chris A Johnson, John L Keltner, J Philip Miller, Richard K Parrish, M Roy Wilson, and Mae O Gordon. The ocular hypertension treatment study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Archives of Ophthalmology*, 120(6):701–713, 2002.

Ronald Klein, Barbara EK Klein, Scot E Moss, and Karen J Cruickshanks. The wisconsin epidemiologic study of diabetic retinopathy: Xvii. the 14-year incidence and progression of diabetic retinopathy and associated risk factors in type 1 diabetes. *Ophthalmology*, 105(10): 1801–1815, 1998.

Theodore Krupin, Jeffrey M Liebmann, David S Greenfield, Robert Ritch, and Stuart Gardiner. A randomized trial of brimonidine versus timolol in preserving visual function: results from the low-pressure glaucoma treatment study. *American Journal of Ophthalmology*, 151(4): 671–681, 2011.

Mark A Latina, Santos A Sibayan, Donald H Shin, Robert J Noecker, and George Marcellino. Q-switched 532-nm nd: Yag laser trabeculoplasty (selective laser trabeculoplasty): a multicenter, pilot, clinical study. *Ophthalmology*, 105(11):2082–2090, 1998.

M Cristina Leske, Suh-Yuh Wu, Anselm Hennis, Robert Honkanen, and Barbara Nemesure. Risk factors for incident open-angle glaucoma: the barbados eye studies. *Ophthalmology*, 115 (1):85–93, 2008.

Edward A Lippa, Laura-Ellen Carlson, Berndt Ehinger, L-Ove Eriksson, Kristina Finnström, Christian Holmin, S-E Gosta Nilsson, Kenneth Nyman, Christel Raitta, and Amund Ringvold. Dose-response and duration of action of dorzolamide, a topical carbonic anhydrase inhibitor. *Archives of Ophthalmology*, 110(4):495–499, 1992.

John HK Liu, Daniel F Kripke, Robert E Hoffman, Michael D Twa, Rex T Loving, Kathleen M Rex, Neeraj Gupta, and Robert N Weinreb. Circadian rhythm of intraocular pressure. *Journal of Glaucoma*, 7(4):243–247, 1998.

Colm McAlinden, Tiina Koskela, Andrea Lim, Robin G Abell, and Daniel Nolan. Selective laser trabeculoplasty (slt) vs other treatment modalities for glaucoma: systematic review. *Eye*, 28 (3):249–258, 2014.

Christiaan MG Olthoff, Jan SAG Schouten, Bart W van de Borne, and Carroll AB Webers. Noncompliance with ocular hypotensive treatment in patients with glaucoma or ocular hypertension: an evidence-based review. *Ophthalmology*, 112(6):953–961, 2005.

Nicholas Papadopoulos, Joel Martin, Qiang Ruan, Ashique Rafique, Michael P Rosconi, Ergang Shi, Emily A Pyles, George D Yancopoulos, Neil Stahl, and Stanley J Wiegand. Binding and neutralization of vascular endothelial growth factor (vegf) and related ligands by vegf trap, ranibizumab and bevacizumab. *Angiogenesis*, 15(2):171–185, 2012.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021.

Philip J Rosenfeld, David M Brown, Jeffrey S Heier, David S Boyer, Peter K Kaiser, Carol Y Chung, and Robert Y Kim. Ranibizumab for neovascular age-related macular degeneration. *New England Journal of Medicine*, 355(14):1419–1431, 2006.

Whitney Shatz, Philip E Hass, Nikolai Peer, Michael T Paluch, Craig Blanchette, Guanglei Han, Wendy Sandoval, Adriana Morando, Wei C Liang, and Justin Yin. Contribution of antibody hydrodynamic size to vitreal clearance revealed through rabbit studies using a species-matched fab. *Molecular Pharmaceutics*, 13(9):2996–3003, 2016.

Michael W Stewart and Philip J Rosenfeld. Pharmacokinetic rationale for dosing every 2 weeks versus 4 weeks with intravitreal ranibizumab, bevacizumab, and aflibercept (vascular endothelial growth factor trap-eye). *Retina*, 32(3):434–457, 2012.

Carol B Toris, Michael E Yablonski, Y-L Wang, and Carl B Camras. Aqueous humor dynamics in the aging human eye. *American Journal of Ophthalmology*, 134(5):749–754, 2002.

UK Prospective Diabetes Study Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (ukpds 33). *The Lancet*, 352(9131):837–853, 1998.

Matthew M Whitacre, Roger A Stein, and Khaled Hassanein. The effect of corneal thickness on applanation tonometry. *American Journal of Ophthalmology*, 115(5):592–596, 1993.

Wan Ling Wong, Xinyi Su, Xiang Li, Chui Ming G Cheung, Ronald Klein, Ching-Yu Cheng, and Tien Yin Wong. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *The Lancet Global Health*, 2(2):e106–e116, 2014.

Jennifer WY Yau, Sophie L Rogers, Ryo Kawasaki, Ecosse L Lamoureux, Jonathan W Kowalski, Toke Bek, Shih-Jen Chen, Jacqueline M Dekker, Astrid Fletcher, and Jakob Grauslund. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*, 35(3):556–564, 2012.