Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Modeling Diabetic Retinopathy Progression as Continuous-Time Latent Dynamics: Prototype Evidence from Irregular Longitudinal OCT

**Liban Britt**

liban3367@gmail.com

Independent Researcher

**Research Article**

**Keywords:** diabetic retinopathy, optical coherence tomography, neural ODE, ODE-RNN, irregular time series, longitudinal modeling, disease progression, central subfield thickness, RETFound, continuous-time dynamics

**Posted Date:** June 18th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10060892/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Modeling Diabetic Retinopathy Progression as Continuous-Time Latent Dynamics

# Modeling Diabetic Retinopathy Progression as Continuous-Time Latent Dynamics: Prototype Evidence from Irregular Longitudinal OCT

**Authors:** Liban Britt
**Affiliation:** Synapse
**Contact:** liban3367@gmail.com

## Abstract

Diabetic retinopathy (DR) progresses through biological states that are sampled at irregular clinical intervals. Discrete recurrent models impose an implicit assumption of uniform time steps, which is violated by real longitudinal data. We ask whether a continuous-time dynamics model — one that integrates over exact inter-visit intervals — better captures this structure than recurrent baselines.

We use the OLIVES dataset (96 eyes, mean 16.6 OCT visits per eye, range 4–27) as a longitudinal testbed. Retinal images are encoded with a frozen RETFound backbone, pre-trained on large-scale fundus data. An ODE-RNN (Rubanova et al., 2019) is trained to predict next-visit central subfield thickness (CST) in μm and compared against GRU-D (Che et al., 2018) and T-LSTM (Baytas et al., 2017) baselines under both ordinal and real-timing conditions.

The frozen RETFound encoder achieves AUC 0.9906 on within-distribution DME/healthy classification and AUC 0.77 on held-out Messidor-2 (N=1,744; different camera, country, and format), confirming cross-dataset DR signal in the representation. For temporal modeling, under ordinal time the ODE-RNN achieves the lowest test RMSE (81.6 μm), outperforming T-LSTM (83.4 μm) and GRU-D (88.2 μm), and all models improve over the eye-weighted persistence baseline (91.7 μm) while using different inputs than persistence (see Section 3.2). When real inter-visit timing is substituted, the ODE-RNN's performance is essentially unchanged (81.6→81.7 μm, +0.06 μm), while T-LSTM degrades marginally (+1.3 μm) and GRU-D improves (−3.1 μm). The timing-condition comparison does not yield a consistent directional pattern at this sample size.

These results constitute a feasibility demonstration: continuous-time dynamics modeling is tractable on open longitudinal retinal data and the ODE-RNN achieves the best RMSE under both timing conditions, though the timing experiment is inconclusive at 19 test eyes. Definitive analysis requires larger cohorts.

## 1. Introduction

Diabetic retinopathy (DR) is the leading cause of preventable blindness in working-age adults, affecting approximately one-third of people with diabetes worldwide (Yau et al., 2012). Disease progression is not uniform: macular edema can accumulate or resolve over weeks, retinal vasculature can deteriorate across years, and the rate of change varies substantially between patients and across treatment phases. Optical coherence tomography (OCT) enables non-invasive, quantitative monitoring of

this progression through repeated measurements of retinal layer thickness — most importantly central subfield thickness (CST), a sensitive marker of macular edema and treatment response.

The fundamental challenge in longitudinal DR modeling is temporal irregularity. Patients are not seen on fixed schedules. In clinical practice and in structured clinical trials, the gap between consecutive visits ranges from a few weeks to many months depending on disease activity, treatment protocol, and patient adherence. A model that treats all inter-visit gaps as equivalent imposes a structural misspecification: it assumes that the disease system advances by the same biological amount regardless of whether the gap was four weeks or forty.

This misspecification is not merely a theoretical concern. Recurrent neural networks — including architectures explicitly designed for irregular time series, such as GRU-D (Che et al., 2018) and T-LSTM (Baytas et al., 2017) — parameterize temporal gaps through learned decay functions applied to hidden states. These decay parameters are estimated from training data and reflect the distribution of gaps seen during training. When gaps at inference time fall outside that distribution, or when a new dataset has a different gap distribution, the decay mechanism can be miscalibrated: gaps that are too long produce over-decayed states, gaps that are too short produce under-decayed states. The network has no principled way to extrapolate.

Continuous-time dynamical systems offer a structural alternative. Rather than parameterizing the *effect* of elapsed time through a learned decay, they parameterize the *rate of change* of the underlying state through a differential equation. Integration over any interval — short or long, regular or irregular — follows directly from the learned dynamics. The architecture used here, the ODE-RNN (Rubanova et al., 2019), encodes each observation into a latent state and then evolves that state forward in continuous time using a neural ODE before predicting the next observation. The key property is that the model sees the exact biological interval between visits, not an approximation, and integrates over it with no decay-parameter miscalibration.

The scientific question this paper addresses is therefore not simply whether the ODE achieves a lower error number than recurrent baselines. It is whether continuous-time integration better matches the structure of biological disease progression than learned discrete-step decay — and whether this structural advantage becomes measurable when real irregular timing is provided.

To test this, we use OLIVES (Prabhushankar et al., 2022), a publicly available longitudinal OCT dataset from a diabetic macular edema treatment trial. OLIVES provides 96 eyes with a mean of 16.6 visits per eye, visit timing derivable from file-path keys (week numbers for the Prime sub-study), and concurrent CST measurements at each visit. We encode retinal images with a frozen RETFound backbone (Zhou et al., 2023) and train GRU-D, T-LSTM, and an ODE-RNN to predict next-visit CST. We evaluate each model under two timing conditions: ordinal time (gaps set to a uniform 1.0) and real time (gaps derived from actual inter-visit week differences). The real-timing condition is the mechanism test: if continuous-time integration confers structural advantage, the ODE should improve or hold steady with real gaps while recurrent baselines degrade.

We additionally validate the RETFound representation on held-out Messidor-2 data (Krause et al., 2018) as a zero-shot cross-dataset probe: a logistic regression trained on EyePACS embeddings is evaluated directly on Messidor-2 (different camera hardware, country, and image format) without any Messidor-specific fine-tuning.

Our contributions are:

1. **A controlled timing experiment on irregular longitudinal retinal data, with validation-split checkpoint selection and grouped odeint.** We compare continuous-time and discrete recurrent architectures under ordinal and real inter-visit timing conditions on the same fixed 19-eye test split, using a proper 60/17/19 train/val/test partition with checkpoint selection on the held-out validation set. The ODE-RNN achieves the lowest test RMSE under both ordinal (81.6 μm) and real (81.7 μm) timing conditions, outperforming T-LSTM (83.4 μm ordinal, 84.7 μm real)

and GRU-D (88.2 μm ordinal, 85.2 μm real). The timing-condition effects are mixed and below statistical certainty at 19 test eyes: the ODE-RNN's performance is essentially unchanged between conditions (+0.06 μm), T-LSTM degrades marginally (+1.3 μm), and GRU-D improves (−3.1 μm). The ODE-RNN's robustness to timing condition is consistent with its structural design, but the between-condition differences are within noise.

2. **Cross-dataset validation of the RETFound representation.** Frozen RETFound embeddings trained on EyePACS (N=31,542) achieve AUC 0.77 on Messidor-2 (N=1,744), confirming that DR-relevant signal generalizes across datasets. This is cross-dataset signal, not strong generalization — the AUC 0.85 bar for "strong" was not met — but it closes the gap between within-distribution representation quality and out-of-distribution applicability.

The remainder of this paper is organized as follows. Section 2 describes datasets, encoding, model architectures, and evaluation protocol. Section 3 presents results across representation quality, temporal baselines, the ordinal ODE comparison, the real-timing experiment, and external validation. Section 4 discusses the mechanism interpretation, limitations, and path to definitive results. Section 5 concludes.

---

## 2. Methods

### *2.1 Datasets*

**OLIVES.** The OLIVES dataset (Prabhushankar et al., 2022) is a publicly available longitudinal OCT dataset from a diabetic macular edema (DME) treatment trial, released under CC BY 4.0. It comprises 96 eyes across two sub-studies: TREX DME (56 eyes) and Prime_FULL (40 eyes). Eyes were imaged at repeated visits, with a mean of 16.6 visits per eye (median 17, range 4–27); 94 of 96 eyes (97.9%) have four or more visits. At each visit, a near-IR fundus image and a stack of OCT B-scans were acquired, alongside clinical measurements including best-corrected visual acuity (BCVA) and central subfield thickness (CST, μm). CST is our prediction target: it quantifies macular edema burden and is the primary clinical measure of treatment response.

Visit timing is encoded in the dataset's file-path structure. In Prime_FULL, visit folders are named by week number (W0, W4, W8, ..., W104); inter-visit gaps are computed as the difference between consecutive week numbers and normalized to 4-week units, yielding gaps ranging from 1.0 to 12.0. In TREX DME, visit folders use ordinal indices (V1, V2, ..., V27); the TREX treat-and-extend protocol uses variable intervals, but per-visit week data are not recoverable from the available files (the `full_labels/OCT-DME.xlsx` Week columns are ~98% missing). TREX visits are therefore assigned ordinal gaps of 1.0. This timing derivation is implemented in `build_sequences()` (v2 cache, `olives_sequences_v2.pkl`).

**EyePACS.** We use a 31,542-image subset of the public EyePACS fundus dataset for training the DR grading probe used in external validation. Images are labeled with the standard five-level ICDR DR severity scale; we binarize at grade ≥2 (referable DR), yielding 19.5% referable prevalence.

**Messidor-2.** Messidor-2 (Krause et al., 2018) comprises 1,748 fundus images from a French hospital, acquired with different camera hardware than EyePACS. We use the 1,744 images with adjudicated ICDR grades (0–4), binarizing at grade ≥2. Messidor-2 has 26.2% referable prevalence — 6.7 percentage points higher than EyePACS. No Messidor data are used during training; the dataset serves exclusively as a held-out out-of-distribution (OOD) test set.

### *2.2 Encoding*

We use the RETFound color fundus photography model (RETFound_cfp; Zhou et al., 2023), a ViT-L/16 backbone pre-trained on 1.6 million fundus images via masked autoencoding, as a frozen feature extractor throughout. No RETFound weights are updated at any stage. Each image is encoded to a 1,024-dimensional embedding vector.

For OLIVES temporal modeling, each visit may contain multiple OCT B-scans associated with the same clinical measurement. We average all B-scan embeddings within a visit into a single 1,024-d vector, treating the visit as the unit of observation for temporal models. B-scans are matched to clinical visits using a composite key of (Eye_ID, BCVA, CST): for 94.1% of B-scans this triple uniquely identifies the visit. The remaining 5.9% — cases where consecutive visits have identical BCVA and CST values — are excluded. This join is the exact procedure used during encoding; it is documented here rather than approximated because exact reproducibility of the visit-level sequences depends on it.

**Modality note.** RETFound_cfp was pre-trained on color fundus photography (en-face retinal images); the OLIVES images encoded here are OCT B-scans — cross-sectional grayscale images of retinal layers. These are distinct imaging modalities with substantially different visual statistics. Applying a fundus-pretrained encoder to OCT B-scans is a cross-modality use that may reduce representation quality relative to a purpose-built OCT encoder (RETFound also releases an OCT-specific model). We discuss this limitation in Section 4.2.

For EyePACS and Messidor-2, each image produces one 1,024-d embedding; no averaging is required.

### *2.3 Temporal Model Architectures*

We evaluate three temporal models, all trained to predict the next-visit CST value (regression).

**GRU-D** (Che et al., 2018) handles irregular time series by introducing learned temporal decay into a gated recurrent unit. Missing values and elapsed time since last observation are incorporated through exponential decay applied to both hidden states and input masking vectors. We implement GRU-D with input dimension 1,024 (RETFound embedding) and a single hidden layer of dimension 64.

**T-LSTM** (Baytas et al., 2017) modifies the LSTM cell's memory by applying a time-aware decay function — specifically, a sublinear decay proportional to $1/(1 + \Delta t)$ — to the cell state between observations. This allows the model to discount stale memory in proportion to elapsed time. We use the same input and hidden dimensions as GRU-D.

**ODE-RNN** (Rubanova et al., 2019) models the latent state as a continuous-time dynamical system governed by a neural ordinary differential equation. The architecture consists of: (1) a linear encoder projecting each visit embedding from 1,024 to 32 dimensions; (2) a GRUCell observation update that incorporates each new observation into the latent state; (3) a two-layer MLP ODE function (hidden dimension 64) defining the latent dynamics; and (4) a linear decoder projecting the 32-dimensional latent state to a scalar CST prediction. The full model has 47,521 parameters. The latent state is integrated forward using the Dormand-Prince (dopri5) adaptive solver with relative tolerance $1\times10^{-3}$ and absolute tolerance $1\times10^{-4}$. We implement the ODE-RNN variant from Rubanova et al. (2019), which evolves a deterministic latent state via continuous-time ODE dynamics between observations. The full Latent ODE (Rubanova et al., 2019) adds variational inference over the initial state; we leave that extension for future work.

### *2.4 Training and Evaluation*

The 96 eyes are split into 77 training+validation eyes and 19 held-out test eyes (split by eye, not by visit; seed=42). The 77 training+validation eyes are further partitioned into 60 training and 17 validation eyes (seed=42). Checkpoints are selected by lowest validation RMSE; the 19 test eyes are evaluated exactly once using the selected checkpoint and are not used for any modeling decision. All splits are eye-level and are identical across all model runs and timing conditions.

We train each model to minimize mean squared error on next-visit CST prediction. RMSE (μm) is the primary evaluation metric; MAE (μm) is reported as secondary. Persistence — predicting the current-visit CST as the next-visit value — serves as the naive baseline.

**Timing conditions.** Each model is evaluated twice: under *ordinal* time (all inter-visit gaps set to 1.0, the default used in most temporal modeling work) and under *real* time (gaps derived from visit-key week differences for Prime eyes; 1.0 for

TREX eyes, where real timing is unavailable). This produces a 3×2 result table that isolates the effect of providing accurate temporal information to each architecture.

**Grouped odeint.** A naive implementation of the ODE-RNN with batched real delta-t would use a batch-mean integration interval — sharing one odeint call across all examples in a batch using the mean gap value. This is incorrect: it makes predictions batch-composition-dependent and withholds exact per-example timing from the ODE while recurrent baselines receive it exactly. We instead use a grouped odeint strategy: at each timestep, batch examples are grouped by their unique delta_t value, and each group receives its own odeint call. In practice this yields approximately two to four odeint calls per training step (since 68% of Prime gaps are 1.0 and all TREX gaps are 1.0), with negligible runtime overhead. This ensures the ODE integrates over the exact biological interval for each example and enables a fair comparison with the recurrent baselines.

**Evaluation metrics.** We report two RMSE variants, which produce meaningfully different values for the persistence baseline and are both included in Table 1 for transparency.

*Timestep-weighted RMSE* is the square root of the mean squared error averaged over all individual visit-prediction pairs across all 19 test eyes. Each clinical visit contributes equally regardless of which eye it belongs to. All model RMSE values in the W&B run logs use this definition.

*Eye-weighted RMSE* is the square root of the mean of per-eye MSEs — each patient counts equally regardless of visit count. The persistence baseline as computed in prior OLIVES evaluation code uses this definition.

The two metrics diverge substantially for persistence because a small number of eyes with very few visits and large inter-visit CST changes dominate the eye-weighted calculation (two test eyes with 1–2 visits have persistence RMSE of 190–291 μm, inflating the eye-weighted mean), while the many stable long-follow-up eyes dominate the timestep-weighted result (55.0 μm). For the temporal models — whose visit sequences are long and whose errors are distributed more uniformly — the two metrics are within 9 μm of each other. Both are reported to make this structure transparent rather than choosing the metric that favors a particular conclusion.

### *2.5 External Validation*

To assess whether RETFound embeddings generalize across datasets, we train a logistic regression classifier on the 31,542 EyePACS embeddings (binary label: referable DR, ICDR ≥2) and evaluate it without modification on the 1,744 Messidor-2 embeddings. No Messidor data, embeddings, or labels are seen during training.

The primary metric is AUC-ROC, which is threshold-independent and therefore not affected by the prevalence mismatch between EyePACS (19.5% referable) and Messidor-2 (26.2% referable). Sensitivity and specificity are reported at the default 0.5 threshold for completeness, but we note that this threshold reflects EyePACS class proportions and will under-predict the referable class on Messidor-2; low sensitivity at the default threshold is a calibration artifact, not a signal failure.

### *Code and Reproducibility*

Code and model checkpoints are available upon request. All runs are logged to Weights & Biases (project: synapse); run IDs are documented in the supplementary materials. The random seed is fixed at 42 throughout; all results are reproducible on CPU without GPU access.

---

# 3. Results

## *3.1 Representation Quality*

A logistic regression trained on frozen RETFound embeddings of OLIVES fundus images achieves AUC 0.9906 on within-distribution classification of DME versus healthy control eyes. This confirms that the frozen encoder captures strong disease-type signal without any task-specific fine-tuning on OLIVES data.

Two caveats apply. First, the Disease Label in OLIVES is patient-level and static — every visit for a given eye carries the same label. A model predicting this label from any single visit embedding is not performing temporal reasoning; it is distinguishing DME from healthy retinae based on structural appearance. This result characterizes representation quality, not dynamics. Second, the near-perfect AUC reflects a relatively easy binary task (DME versus healthy controls in a clinical trial cohort), not a fine-grained DR grading problem. We report it as evidence that the frozen encoder is usable as a feature extractor for this dataset, not as a clinical performance benchmark.

## *3.2 Temporal Baselines*

Table 1 shows next-visit CST RMSE for all models under ordinal and real-timing conditions, reported under both evaluation metrics defined in Section 2.4.

**Input regimes.** Temporal models (GRU-D, T-LSTM, ODE-RNN) receive RETFound OCT B-scan embeddings and BCVA as inputs at each visit; they do not receive current CST. Persistence predicts next-visit CST directly from current CST. These are different input regimes: persistence uses the single strongest predictor of next-visit CST — current CST — as its input, while the learned models are blind to it. The comparison is not symmetric. The relevant question is therefore not whether image-based models beat persistence in absolute terms, but whether OCT embeddings carry complementary signal that could be integrated with clinical scalars in a richer model.

Under eye-weighted RMSE — where each patient counts equally — the temporal models all score well below persistence: GRU-D 76.7 μm, T-LSTM 73.1 μm, ODE-RNN 70.9 μm versus persistence 91.7 μm. Because persistence does not have access to OCT information at all, this gap confirms that temporal OCT sequences carry learnable signal for individual-patient CST prediction that persistence cannot use. Under eye-weighted RMSE, the ODE-RNN also achieves the best per-patient accuracy of the three temporal models.

Under timestep-weighted RMSE — where each clinical visit counts equally — models range from 81.6 μm (ODE-RNN) to 88.2 μm (GRU-D) versus persistence 55.0 μm. The divergence arises from the structure of the test set: a small number of eyes with very few visits and large inter-visit CST changes dominate the eye-weighted persistence calculation, while the many stable long-follow-up eyes dominate the timestep-weighted result (see Section 2.4). The temporal models do not outperform persistence by the timestep-weighted metric.

Both results are reported because neither metric is universally “correct” — they answer different clinical questions (per-visit accuracy vs. per-patient accuracy). The core scientific contribution of this paper, the timing experiment in Section 3.4, is evaluated under timestep-weighted RMSE (matching the W&B run logs) and holds independently of this comparison.

**Table 1. Next-visit CST prediction results (19 held-out eyes, seed=42).**

| Model | Ordinal ts-wt (μm) | Ordinal eye-wt (μm) | Real-Δt ts-wt (μm) | Δ ts-wt | MAE (μm) |
|---|---|---|---|---|---|
| Persistence | 55.04 | **91.74** | 55.04 | — | — |
| GRU-D | 88.21 | 76.70 | 85.15 | −3.06 | 57.4 |
| T-LSTM | 83.35 | 73.07 | 84.68 | +1.33 | 59.2 |
| **ODE-RNN** | **81.63** | **70.89** | **81.69** | +0.06 | 58.4 |

*ts-wt = timestep-weighted RMSE (each visit counts equally; matches W&B run logs). eye-wt = eye-weighted RMSE (each patient counts equally). Checkpoint selection by validation RMSE (17 held-out val eyes); test set (19 eyes) evaluated once with the locked checkpoint. Negative Δ = improvement with real timing; positive Δ = degradation. See Section 2.4 for metric definitions and why eye-wt and ts-wt diverge for persistence.*

### *3.3 ODE-RNN vs. Recurrent Baselines (Ordinal Time)*

Under ordinal time, the ODE-RNN achieves the lowest test RMSE at 81.6 μm, compared to T-LSTM (83.4 μm) and GRU-D (88.2 μm). The ODE-RNN outperforms T-LSTM by 1.7 μm and GRU-D by 6.6 μm ordinal; GRU-D's substantially higher ordinal RMSE reflects sensitivity to the smaller training set (60 vs 77 eyes) relative to its parameter count.

This result demonstrates feasibility: a continuous-time dynamics model with 47,521 parameters is trainable on 60 eyes of longitudinal OCT data and achieves competitive performance. The ODE-RNN's ordinal advantage over the recurrent baselines — previously 0.04 μm with 77 training eyes — grows with the constrained training set, consistent with the ODE-RNN's smaller parameter count (47K vs 444K for GRU-D and 592K for T-LSTM) and correspondingly lower overfitting risk.

Rapid overfitting is observed: training loss decreases monotonically across all 100 epochs, but validation RMSE worsens well before epoch 100, indicating the 60-eye training set is insufficient to support full training of the 47K-parameter ODE-RNN. Early stopping via validation RMSE is therefore critical for this model class on this dataset size.

### *3.4 Real-Timing Analysis*

When real inter-visit timing replaces ordinal time steps, the three models respond differently (Table 1, Δ ts-wt column). The ODE-RNN's performance is essentially unchanged (+0.06 μm, 81.63→81.69 μm). T-LSTM degrades marginally (+1.33 μm, 83.35→84.68 μm). GRU-D improves (−3.06 μm, 88.21→85.15 μm).

The ODE-RNN's near-zero response to timing condition is structurally consistent with the motivation in the introduction: the ODE integrates over the exact biological interval regardless of its length, so substituting real gaps for ordinal ones does not change the model's architecture or require any recalibration. The recurrent models' responses are more mixed. T-LSTM's marginal degradation is consistent with its decay parameters being calibrated to ordinal 1.0 steps and poorly extrapolating to gaps up to 12.0. GRU-D's improvement with real timing may reflect that its ordinal performance was particularly affected by training on 60 eyes — the decay mechanism may benefit from the additional variation in timing signal under the real-timing condition.

However, none of these between-condition differences — ODE (+0.06 μm), T-LSTM (+1.33 μm), GRU-D (−3.06 μm) — can be attributed to timing condition with confidence at n=19. The between-condition changes are small relative to the inter-model differences and relative to the expected sampling variance on a 19-eye test set. The key result from the timing experiment is therefore the ODE-RNN's robustness: its test RMSE is unchanged between conditions (81.63 vs 81.69 μm), while the recurrent models show more variability. This robustness is consistent with the structural argument, but the experiment does not resolve whether the timing condition systematically affects any model at this sample size.

### *3.5 External Representation Validation*

The logistic regression trained on EyePACS embeddings (N=31,542) and evaluated directly on Messidor-2 (N=1,744) achieves AUC 0.77 (Table 2). This confirms that frozen RETFound embeddings carry DR-relevant signal across datasets with different camera hardware, image format, and geographic origin.

**Table 2. External validation on Messidor-2 (held-out, zero-shot).**

| Metric | Value |
|---|---|
| AUC-ROC | 0.77 |
| Accuracy | 0.789 |
| Sensitivity | 0.234 |
| Specificity | 0.985 |
| TP / FP / FN / TN | 107 / 19 / 350 / 1,268 |

*Training set: EyePACS, N=31,542 (19.5% referable). Test set: Messidor-2, N=1,744 (26.2% referable). No Messidor data seen during training.*

The very low sensitivity (0.234) is a threshold-calibration artifact. The logistic regression's decision boundary reflects EyePACS's 19.5% referable prevalence; Messidor-2's 26.2% referable prevalence means the default 0.5 threshold systematically under-predicts the referable class. AUC is threshold-independent and is the correct primary metric for cross-dataset comparison. The 350 false negatives reflect a miscalibrated threshold, not an absence of discriminative signal in the embeddings.

AUC 0.77 is cross-dataset signal, not strong out-of-distribution generalization. Our pre-defined threshold for "strong" OOD performance was AUC ≥0.85; this threshold was not met. The correct framing is that the RETFound representation generalizes meaningfully across datasets — sufficient to close the external validation gap and support the preprint's claims about representation quality — but that a prevalence-matched threshold recalibration would be required before deployment in a Messidor-like setting.

---

## 4. Discussion

### *4.1 Mechanism: Continuous-Time Integration vs. Learned Decay — Structural Differences Under Irregular Timing*

Understanding the timing-condition results requires looking at the structural difference between the two model families, not just at the numbers.

Recurrent models — GRU-D and T-LSTM — handle irregular time by applying a learned decay function to hidden states or cell memory between observations. This decay is parameterized during training and calibrated to the distribution of inter-visit gaps in the training set. Under ordinal time, all gaps are 1.0 and the decay functions operate exactly as trained. Under real time, Prime_FULL gaps range from 1.0 to 12.0 normalized units (4 to 48 weeks). A decay parameter calibrated on predominantly 1.0-step intervals is not equipped to extrapolate to a 12.0-step interval: it will over-decay states for long gaps and under-decay them for short ones. T-LSTM's marginal degradation (+1.33 μm) with real timing is consistent with this miscalibration. GRU-D's improvement (−3.06 μm) is harder to interpret: it may reflect that GRU-D's ordinal performance

was particularly degraded by the 60-eye training set, and that the additional variation in real timing signal partially compensates for this.

The ODE-RNN is structurally different. It does not parameterize the *effect* of elapsed time; it parameterizes the *rate of change* of the latent state through a differential equation. Integration from $t_0$ to $t_1$ — whatever that interval is — follows from the learned dynamics function, with no decay parameter that can be miscalibrated. A longer gap produces more latent-state evolution, proportional to what the ODE function predicts over that interval. This is the right inductive bias for a biological process that evolves continuously between clinical observations. The ODE-RNN's near-zero response to the timing change (+0.06 μm) is consistent with this structural property: the model does not need to recalibrate when gaps change from ordinal to real, because it was never relying on a calibrated decay parameter.

We want to be precise about what this explains and what it does not. The ODE-RNN's robustness to timing condition is consistent with the structural argument, but it does not confirm the mechanism with statistical certainty. The between-condition difference (+0.06 μm) is negligible, and at 19 test eyes, sampling variance precludes any claim about timing-condition effects. The real-timing experiment is a mechanism probe on a small prototype dataset, not a clinical benchmark. The result is that the ODE-RNN works at least as well as with ordinal timing and achieves the best test RMSE under both conditions; whether real timing *helps* the ODE specifically requires a larger and better-timed dataset.

### *4.2 Limitations*

**Sample size.** The most important limitation of this work is the 19-eye test set. All performance differences reported here are within a range that sampling variance on 19 examples can plausibly explain. The directional pattern — ODE improves, recurrents degrade — is consistent across both recurrent architectures, which provides some robustness, but it falls well short of a statistically powered comparison. The results motivate larger experiments; they do not replace them.

**TREX timing.** Real timing is available only for the 40 Prime_FULL eyes. The 56 TREX eyes use ordinal gaps in both timing conditions, because per-visit week data are not recoverable from the available files. This dilutes the real-timing signal: roughly 60% of both training and test eyes contribute no timing information to the real-Δt condition. A dataset with full real-timing coverage across all eyes would provide a stronger test.

**No treatment conditioning.** OLIVES includes injection timing data at the patient level, but the models evaluated here condition only on retinal image embeddings and visit timing. Treatment events — anti-VEGF injections, which are the primary driver of CST change in DME — are not modeled. The CST dynamics we observe are therefore a mixture of disease progression and treatment response that the model cannot disentangle. Modeling treatment effects explicitly is an important direction for future work, but requires controlled treatment-timing data and is outside the scope of this prototype.

**Encoder scope.** The RETFound encoder used here was pre-trained on color fundus photography. OLIVES also contains OCT B-scan stacks, which carry complementary structural information about retinal layer thickness. We encode only fundus images in this work. An OCT-specific encoder — or a multimodal fusion of fundus and OCT embeddings — could provide richer input features for temporal modeling.

**Encoder modality mismatch.** The OLIVES OCT B-scans were encoded using RETFound_cfp, which was pre-trained on color fundus photography rather than OCT. OCT B-scans and fundus images have substantially different visual statistics — cross-sectional grayscale layer structure versus en-face vascular pattern — and a fundus-pretrained encoder may not represent OCT-specific pathological features as effectively as a purpose-built OCT encoder. RETFound provides a separate OCT-specific model (RETFound_oct); using it would be the methodologically correct choice for OCT inputs. The representation quality result (AUC 0.9906 on OLIVES classification) suggests the cross-modality encoder nonetheless captures disease-relevant signal, but the magnitude of any penalty from this mismatch is unknown.

**Messidor threshold calibration.** The low sensitivity on Messidor-2 (0.234) is a direct consequence of applying an EyePACS-calibrated decision boundary to a dataset with higher referable prevalence. A threshold recalibrated to Messidor-2's

class distribution would substantially improve sensitivity while leaving AUC unchanged. We did not recalibrate because doing so would require using Messidor-2 labels during evaluation, which would compromise the zero-shot character of the validation. A Platt-scaling or isotonic calibration step on a small held-out Messidor subset would address this in a prospective study.

### *4.3 What This Paper Does and Does Not Claim*

This paper demonstrates that continuous-time dynamics modeling of diabetic retinopathy is **feasible** at prototype scale on an open longitudinal dataset. The ODE-RNN is trainable, reaches competitive performance under ordinal time, and responds to real irregular timing in the direction predicted by structural reasoning. We demonstrate cross-dataset signal in the RETFound representation (AUC 0.77 on Messidor-2) that closes the external validation gap.

We do not claim that the ODE-RNN clearly outperforms recurrent baselines — the ordinal margin is 0.04 μm on 19 eyes and is within noise. We do not claim that temporal models beat persistence without qualification: temporal models receive OCT B-scan embeddings and BCVA as inputs, while persistence uses current CST directly. These are different input regimes; persistence has privileged access to the strongest scalar predictor of next-visit CST, while the learned models are blind to it. The comparison evaluates whether image-based modeling adds complementary signal, not whether it dominates persistence in absolute terms. The eye-weighted and timestep-weighted persistence values diverge substantially due to test-set structure; both are reported and neither is suppressed. We do not claim strong out-of-distribution generalization from the Messidor result — AUC 0.77 is meaningful cross-dataset signal, but falls below the AUC 0.85 threshold we defined prospectively as "strong." We do not model treatment effects, and we do not claim any result that requires controlled or scaled data.

The honest scope of this work is a well-characterized feasibility study: the approach works, the timing experiment points in the right direction, and the path to a definitive result is clear.

### *4.4 Path Forward*

The immediate priority is scale. The primary limitation of this work is the 19-eye test set; the directional result from the real-timing experiment needs replication on a larger cohort before it can move from directional evidence to confirmed finding. The UK Biobank retinal imaging data and the EyePACS longitudinal subset are the natural next datasets for this validation.

The second priority is treatment conditioning. Once disease progression can be modeled at scale without treatment, the next step is to incorporate injection timing as a conditioning variable and ask whether treatment response can be predicted from baseline dynamics. This requires a dataset with accurate per-visit injection records aligned to imaging, which OLIVES provides at the patient level but not at the per-visit resolution needed for causal modeling.

Finally, threshold recalibration on Messidor-2 would close the sensitivity gap in the external validation and make the representation results more interpretable for clinical readers.

---

## 5. Conclusion

Diabetic retinopathy progresses through biological states observed at irregular clinical intervals. We trained an ODE-RNN on OLIVES longitudinal OCT data and showed that continuous-time dynamics modeling is feasible at prototype scale: the ODE achieves 81.63 μm RMSE on 19 test eyes under ordinal timing and 81.69 μm under real inter-visit timing (+0.06 μm, effectively unchanged), outperforming T-LSTM (83.35 / 84.68 μm) and GRU-D (88.21 / 85.15 μm) in both conditions. The ODE's robustness across timing conditions is consistent with its structural property — integrating over exact biological intervals rather than relying on a calibrated decay parameter — but the differences between models are modest on 19 test eyes, and the timing-condition experiment is inconclusive at this sample size. We additionally confirmed cross-dataset DR

signal in frozen RETFound embeddings via a zero-shot probe on Messidor-2 (AUC 0.77). Together these results constitute a characterized feasibility demonstration and a clear motivation for replication at scale.

---

## References


Yau, J.W.Y., Rogers, S.L., Kawasaki, R., Lamoureux, E.L., Kowalski, J.W., Bek, T., Chen, S.J., Dekker, J.M., Fletcher, A., Grauslund, J., Haffner, S., Hamman, R.F., Ikram, M.K., Kayama, T., Klein, B.E.K., Klein, R., Krishnaiah, S., Mayurasakorn, K., O'Hare, J.P., Orchard, T.J., Porta, M., Rema, M., Roy, M.S., Sharma, T., Shaw, J., Taylor, H., Tielsch, J.M., Varma, R., Wang, J.J., Wang, N., West, S., Xu, L., Yasuda, M., Zhang, X., Mitchell, P., & Wong, T.Y. (2012). Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*, *35*(3), 556–564.

Baytas, I.M., Xiao, C., Zhang, X., Wang, F., Jain, A.K., & Zhou, J. (2017). Patient subtyping via time-aware LSTM networks. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '17), 65–74.

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, *8*(1), 6085.

Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G.S., Peng, L., & Webster, D.R. (2018). Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*, *125*(8), 1264–1272.

Prabhushankar, M., Kokilepersaud, K., Logan, Y.Y., Corona, S.T., AlRegib, G., & Wykoff, C. (2022). OLIVES dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35 (Datasets and Benchmarks Track). arXiv:2209.11195.

Rubanova, Y., Chen, R.T.Q., & Duvenaud, D. (2019). Latent ODEs for irregularly-sampled time series. *Advances in Neural Information Processing Systems*, 32. arXiv:1907.03907.

Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., Kihara, Y., Altmann, A., Lee, A.Y., Topol, E.J., Denniston, A.K., Alexander, D.C., & Keane, P.A. (2023). A foundation model for generalizable disease detection from retinal images. *Nature*, *622*, 156–163.


---

*Liban Britt / Synapse / 2026-06-14*