The appendix is divided into four sections. Appendix A provides the dataset, labels, and benchmarking access. The dataset images and benchmarking codes are currently public, while the labels are provided privately as a link in Appendix A.1. Appendix B details some of the statistics of the dataset. This includes comparison against existing ophthalmology datasets, detailing the challenges within the OLIVES dataset, expanding on the full list of clinical labels that are available in the PRIME and TREX-DME clinical trials, and the exact procedure used to annotate the biomarkers. Appendix C provides additional medical context to all the benchmarking results from Section 4. Furthermore, experimental details including training setup, error bars, and computational resources are discussed. Finally, relevant procedural details regarding the PRIME and TREX DME clinical trials are discussed in Appendix D, along with screenshots of relevant labels.

A Dataset and Benchmarking Access

A.1 Links to Access Dataset

We provide open access to the dataset. The images and labels found in the OLIVES dataset are present at:
Image Access
Alternate access to the labels directly can be found at:
Labels Access
The benchmarks provided in the paper are accessible at the following link:
Code Access

A.2 Licenses and DOI

The code is associated with an MIT License. The DOI of the dataset is 10.5281/zenodo.6622145. The associated license with the dataset is a Creative Commons International 4 license.

A.3 Maintenance Plan

The code will be hosted within the github repository specified in Section A.1. Instructions and details regarding the dataset will be located at this same repository. Images for the dataset are located at the zenodo directory in Section A.1. Labels for these images will be included within this same zenodo dataset after acceptance of the paper. Additional data from other clinical studies will be added over time as part of our partnership with the Retinal Consultants of Texas. Within the Github repository, we will maintain a comprehensive survey of all literature that use the OLIVES dataset. This will include a unified result table and access to publicly available github repositories that benchmark on OLIVES. Furthermore, we anticipate additional applications that make use of the OLIVES dataset and its multi-modal and time-series data (Appendix C.4) and will update the Github repository with these applications.

A.4 Dataset Folder Structure

Images The dataset is split into two folders: Prime and TREX-DME. These correspond to the studies that the respective data originated from. These studies also act as labels for images with diabetic retinopathy (within PRIME folder) and DME (within TREX-DME folder) as these are the disease states studied in their respective trials. Within each clinical study directory there are folders that have the imaging data for each respective patient. Inside of each patient folder is a directory for every visit by each patient. Within every visit folder are folders containing the OCT scans and fundus image for the eye(s) associated with the patient of interest. This structure is consistent in both studies with the only difference being that the TREX DME directory is split into three subdirectories called GILA, Monthly, and TREX that identify specific cohorts of patients. Within every visit, there is a numpy file that is the 3D volume stitched together for the OCT scans of that patient. Additionally, for every patient, there is a numpy file that holds the fundus image and OCT volume generated at every visit into one data structure in the order in which the visits occurred.

Labels The labels exist within two directories called "full labels" and "ml centric labels." Full labels contains the complete clinical datasheets for both the Prime and TREX DME studies. This

directory also has a word document with additional details regarding the study. The ml centric labels directory has two csv files. The first contains full biomarker and clinical labels for the 9408 OCT scans that were labeled from the first and last visit of every eye. The other excel file contains the BCVA, CST, eye id, and patient id of all 78185 OCT scans that exist within the OLIVES dataset. These are the clinical labels that are common between both trials.

A.5 Reproducibility Statement and Attributions

We compare against three self-supervised approaches in this paper. Links to their implementations are provided here:

SimCLR
PCL
Moco v2

Results for our paper can be replicated using the code, images, and labels found in Section A.1.

B Dataset Statistics

B.1 Dataset Comparison

Dataset	Clinical Labels	Biomarker Labels	TimeSeries Data	MultiModal Images	Disease States	No. of Images	No. of Biomarkers
Kermany (13)	✓	✓	✗	✗	✓	109312	4
Farisu (14)	✓	✓	✗	✗	✓	38400	4
Srinivasan (42)	✓	✗	✗	✗	✓	3231	0
Maetschke (43)	✓	✗	✗	✗	✓	1110	0
Kaggle DR (44)	✓	✗	✗	✗	✓	35126	0
AG-CNN (45)	✓	✗	✗	✗	✓	4854	0
ODIR (46)	✓	✗	✗	✗	✓	10000	0
DeepDrid (47)	✓	✗	✗	✓	✓	2256	0
Laterality (48)	✓	✗	✗	✗	✓	18394	0
Messidor (49)	✓	✗	✗	✗	✓	1748	0
OLIVES	✓	✓	✓	✓	✓	78185	16

Table 5: Comparison of eye-related datasets along relevant medical considerations.

Dataset	Clinical Labels	Biomarker Labels	MultiModal Images	Disease States	No. of Eyes	No. of Images	No. of Biomarkers
Rotterdam (50)	✓	✗	✗	✓	70	1120	0
Rivail et. al. (51)	✗	✗	✗	✓	221	3308	0
OLIVES	✓	✓	✓	✓	96	78185	16

Table 6: Comparison of eye-related time-series datasets along relevant medical considerations.

In Table 5, we compare OLIVES against existing datasets based on 7 relevant considerations: clinical labels, biomarker labels, time-series data, multi-modal data, disease states, number of images, and number of biomarkers. Among these, biomarker labels and disease state labels have the most semantic overlap and necessitate a clear differentiation with how these are defined. Disease states refer to the overall condition of the eye. For example, an eye can have the overall disease of diabetic retinopathy or any of its variants. However, biomarkers refer to explicit features present within an OCT scan or fundus image that can act as indicators for the disease (1). For example, a biomarker such as intra-retinal fluid (IRF), is a description of the features present in an individual image, but do not make a statement of the overall disease that the eye is experiencing. Additionally, biomarkers can vary between OCT scans found at different positions within a volume and thus act as a more fine-grained description of the content of an individual image. Furthermore, we define biomarkers with respect to biological features, rather than measurements taken across the image. We deem measurements, such

as various retinal thickness values, as a type of clinical label due to its derivation from values taken from the imaging acquisition device (OCT Machine).

B.2 Challenges in Dataset

A number of challenging datasets exist for natural images and videos. These challenges include noise additions (7), background and imaging modality shifts (8), fine-grained domain shifted videos (9), and microscopic textures (52). Challenging datasets for computed images include large scale seismic datasets (53). The challenge in OLIVES and other medical datasets arises not because of interventions in data, but due to issues in data collection, inversion, representation, annotation, and analysis of minute changes within computed data. Consider Fig. 2a). A singular OCT scan sampled randomly from the 3D volume of two separate visits between treatments is shown. Notice the same disease diagnosis and minimal differences within the scans. In contrast, Fig. 2b) shows the OCT scans of three separate patients in their first visit, all of whom are diagnosed with DME. The manifestations of the DME pathololgy is noticeably different between patients. Similarly, in Fig. 2c), the CST clinical label for two separate patients with visually dissimlar OCT scans is shown. On the other hand, gradually decreasing CST values between visits for the same patient indicates a decrease in DME's manifestation in Fig. 2d).

Moreover, the ML techniques used to analyze natural images may not be applicable or sufficient for OCT scans. (51) introduced a novel pretext task that involved predicting the time interval between OCT scans taken by the same patient. (54) showed how a combination of different pretext tasks such as rotation prediction and jigsaw re-ordering can improve performance on an OCT anomaly detection task. (55) showed how assigning pseudo-labels from the output of a classifier can be used to effectively identify labels that might be erroneous. These works all identify ways to use variants of deep learning to detect important biomarkers in OCT scans. The OLIVES dataset introduces new challenges in these setups by providing biomarkers and clinical labels that correlate with image data.

B.3 Dataset Logistics

Clinical Trial Funding The initial clinical trials, PRIME and TREX-DME are published at (2) and (3; 4; 5; 6) respectively. These trials were conducted between December 2013 and April 2021 at the Retina Consultants of Texas (Houston, TX, USA). The PRIME study was supported by Regeneron Pharmaceuticals. Further financial disclosures are provided in (2). The corresponding author on (2) is also an author for this article. The TREX-DME study was supported by various grants detailed after References in (3).

Labeling The processes for the clinical trials and diagnosis is provided in (2) and (3; 4; 5; 6). For OLIVES, biomarkers are retrospectively added to $9,408$ images. The biomarkers are identified by Charles C. Wykoff with an ophthalmology experience of sixteen years and the labeling is performed by Stephanie Trejo Corona with a grading experience of one year.

B.4 Addressing Limitations of OLIVES

An issue identified in Section 5 is that the OLIVES dataset does not provide a global patient distribution. This is a common problem with medical datasets and has sparked research into strategies that can overcome this distributional bias (56). Within the corpus of ophthalmology related studies, there are several datasets that originate from different regions of the world, such as (49) from France, (13) from a collaboration of the USA and China, (45) from China, and (57) from the United Kingdom. It is possible to train with our dataset and test the resulting algorithm with these and others found at (11) to test for out of distribution performance from cohorts across the world.

Other limitations are addressed in the main paper and relate to the nature of the cohort in our studies. The cohorts chosen are from patients exhibiting some severity level of Diabetic Retinopathy (DR) or Diabetic Macular Edema (DME). As a result, there are no patients that are completely healthy. If it is desirable to guarantee healthy instances within a specific study, then it is possible to augment our dataset with healthy OCT scans or Fundus images from sources such as (18), (49), or (14).

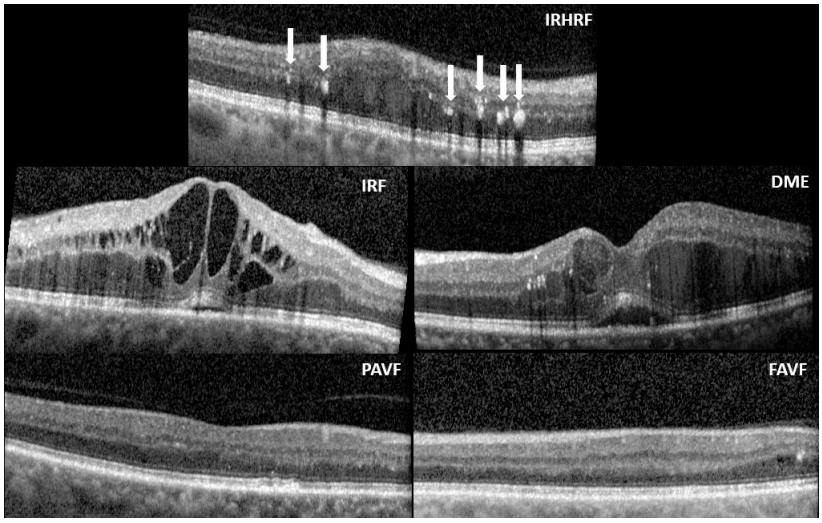

Figure 7: Cross-sectional images of graded biomarkers. Intra-Retinal Hyper-Reflective Foci (IRHRF), indicated by the six white arrows, are areas of hyperreflectivity in the intraretinal layers with or without shadowing of the more posterior retinal layers. Intra-Retinal Fluid (IRF) encompasses the cystic areas of hyporeflectivity. Diabetic Macular Edema (DME) is the apparent swelling and elevation of the macula due to the presence of fluid. A Partially Attached Vitreous Face (PAVF), with an arrow indicating the point of attachment and a Fully Attached Vitreous Face (FAVF). A discussion of these biomarkers can be found at (58).

B.5 Description of Labels

B.5.1 Biomarkers and their Generation

The authors in (1) describe biomarkers as objective indicators of medical state as observed and measured from outside the patient. They are quantifiable characteristics of biological processes. In this paper, the biological processes are diseases and biomarkers indicate the presence or absence of such diseases. Under limited circumstances, the authors in (1) suggest that biomarkers can be surrogate endpoints in clinical trials. However, they caution against doing so unless the underlying clinical trial is specifically meant for the study. As such, biomarkers indicate the presence of diseases, but are not causal to these diseases. Causality in the medical domain can be singular causality or general causality (59). Singular causality is constrained by events in a time-series linked events while general causality analyzes relationships between events. As such this is different from visual causal features from (23) or causal question-based analysis in (24) or causal factor analysis in (25).

All image interpretations were performed by a trained grader for the presence of the following parameters: atrophy or thinning of retinal layers, disruption of the ellipsoid zone (EZ), disruption of the retinal inner layers (DRIL), intraretinal (IR) hemorrhages, intraretinal hyperreflective foci (IRHRF), partially attached vitreous face (PAVF), fully attached vitreous face (FAVF), preretinal tissue or hemorrhage, vitreous debris, vitreomacular traction (VMT), diffuse retinal thickening or macular edema (DRT/ME), intraretinal fluid (IRF), subretinal fluid (SRF), disruption of the retinal pigment epithelium (RPE), serous pigment epithelial detachment (PED), and subretinal hyperreflective material (SHRM). The following describes the grading used for each morphological feature evaluated in each B-scan using the Heidelberg Spectralis HRA+OCT software.

Atrophy or thinning of retinal layers was indicated as present with evidence of RPE atrophy or thinning of the retina at the trained grader's discretion (60; 61). Disruption of the EZ was indicated as present with when the second-most posterior hyperreflective band of the retina was discontinuous. DRIL was indicated as present when the boundaries of the retinal inner layers such as the inner nuclear layer, outer plexiform layer, and ganglion cell layer were not clearly defined (27). Intraretinal hemorrhages were indicated as present when there was a small, localized lesion that caused shadowing of the more posterior retinal layers, with a corresponding lesion visible on the near-infrared fundus image. IRHRF were indicated as present with the appearance of intraretinal, highly reflective spots,

Table of Abbreviations	
Abbreviation	Full Name
CST	Central Subfield Thickness
BCVA	Best Central Visual Acuity
Eye ID	Eye Identity
EZ	Ellipsoid Zone
DRIL	Disruptionof the Retinal Inner Layers
IR	Intraretinal
IRHRF	Intraretinal Hyperreflective Foci
PAVF	Partially Attached Vitreous Face
FAVF	Fully Attached Vitreous Face
VMT	Vitreomascular Traction
DRT/ME	Diffuse Retinal Thickening or Macular Edema
IRF	Intraretinal Fluid
SRF	Subretinal Fluid
RPE	Retinal Pigment Epithelium
PED	Pigment Epithelial Detachment
SHRM	Subretinal Hyperreflective Material
DR	Diabetic Retinopathy
DME	Diabetic Macular Edema
CI-DME	Center-Involved Diabetic Macular Edema
PDR	Proliferative Diabetic Retinopathy
NPDR	Non-Proliferative Diabetic Retinopathy
OCT	Optical Coherence Tomography
AMD	Age-related Macular Degeneration
CNV	Choroidal Neovascularization
VEGF	Vascular Endothelial Growth Factor
ETDRS	Early Treatment Diabetic Retinopathy Study
DRSS	Diabetic Retinopathy Severity Scale
PRIME	Real-Time Objective Imaging to Achieve Diabetic Retinopathy Improvement
TREX-DME	Treat and Extend Protocol in Patients with Diabetic Macular Edema

Table 7: Summary clinical and biomarker abbreviations used throughout the paper.

which correspond pathologically to microaneurysms or hard exudates, with or without shadowing of the more posterior retinal layers (62). A partially attached vitreous face was indicated as present with evidence of perifoveal detachment of the vitreous from the internal limiting membrane (ILM) with a macular attachment point within a 3-mm radius of the fovea. A fully attached vitreous was indicated as present with no evidence of perifoveal or macular detachment from the ILM. Preretinal tissue or hemorrhage was indicated as present with evidence of an hyporeflective preretinal tissue, epiretinal membrane, or hemorrhage over the surface of the ILM (63). Vitreous debris was indicated as present with evidence of hyperreflective foci in the vitreous or shadowing of the retinal layers in the absence of an intraretinal hemorrhage. VMT was indicated as present with evidence of perifoveal vitreous separation, vitreomacular attachment, and foveal anatomic distortions (64). Diffuse retinal thickening or macular edema was indicated as present when there was increased retinal thickness of 50 µm above the otherwise flat retina surface with associated reduced reflectivity in the intraretinal tissues (65). Intraretinal fluid was indicated as present when intraretinal hyporeflective areas or cysts had a minimum fluid height of 20 µm (65). Subretinal fluid was indicated as present when hyporeflective areas or cysts were evident in the subretinal space between the EZ and RPE layers. Disruption of the RPE was indicated as present when the most posterior hyperreflective band of the retina was discontinuous. Serous pigment epithelial detachment was indicated as present with evidence of a hyporeflective area underneath the detached RPE. SHRM was indicated as present when hyperreflective foci were evident in the subretinal space between the EZ and RPE layers.

Dataset	Statistics	Label Type	Label Names
PRIME Clinical	29000+ Images 40 Patients 40 Unique Eyes	Clinical	BCVA, CST, DRSS, Eye ID,Patient ID, Diabetes Type, BMI, Age, Race, Gender Years with Diabetes, HbA1c, Leakage Index, Injection Arm
PRIME Biomarker	3900+ Images 40 Patients 40 Unique Eyes	Clinical	BCVA, CST, DRSS, Eye ID,Patient ID, Diabetes Type, BMI, Age, Race, Gender Years with Diabetes, HbA1c, Leakage Index, Injection Arm
		Biomarker	16 Biomarkers (DME, IRF, IRHRF, etc.)
TREX-DME Clinical	38000+ Images 47 Patients 56 Unique Eyes	Clinical	BCVA, Snellen Score, CST, Eye ID, Patient ID
TREX-DME Biomarker	5300+ Images 47 Patients 56 Unique Eyes	Clinical	BCVA, Snellen Score, CST, Eye ID, Patient ID
		Biomarker	16 Biomarkers (DME, IRF, IRHRF, etc.)
TREX-DME + PRIME Biomarker	9200+ Images 87 Patients 96 Unique Eyes	Clinical	BCVA, CST, Eye ID, Patient ID
		Biomarker	16 Biomarkers (DME, IRF, IRHRF, etc.)
TREX-DME + PRIME Clinical	67000+ Images 87 Patients 96 Unique Eyes	Clinical	BCVA, CST, Eye ID, Patient ID

Table 8: Summary of clinical and biomarker data present within each individual study.

B.5.2 Clinical Labels and their Generation

Full Clinical Labels The clinical labels obtained from the PRIME trials include BCVA, DRSS, CST, eye ID, patient ID, diabetes type, BMI, age, race, gender, HbA1c, leakage index, years with diabetes, and injection arm. The clinical labels from the TREX-DME trials include BCVA, Snellen score, CST, Eye ID, and Patient ID. Since OLIVES is a combination of the two, we use only the common labels from both trials as our clinical labels in our experiments. These common labels include BCVA, CST, Patient ID and Eye ID which are listed in Table 1. However, we provide access to all available labels as described in Appendix A.4.

The Early Treatment Diabetic Retinopathy Study (ETDRS) diabetic retinopathy severity scale (DRSS) has 13 levels describing DR severity and change over time based on color fundus photograph grading. The scale starts at level 10 and ends at level 90 with irregular scale numbering. Nonproliferative diabetic retinopathy (NPDR) DRSS levels on the scale are below 61 and proliferative diabetic retinopathy (PDR) levels are 61 and above. Diabetes type refers to the patient's diagnosis of either type one or type two diabetes mellitus. HbA1c is the measurement of glycated hemoglobin, commonly referred to as blood sugar, which serves as an indicator for diabetes diagnosis or diabetic control. Leakage index refers to the panretinal leakage index used in the PRIME trial in which areas of leakage, regions of hyperfluorescence in fluorescein angiography images, were divided by areas of interest, region of total analyzable retinal area, and converted to a percentage. Injection arm refers to either the DRSS-guided (1) cohort or the PLI-guided (2) cohort in the PRIME trial. Snellen score is the visual acuity testing procedure commonly used in ophthalmic clinical settings. The first number indicates the distance in feet that the letter chart was read, in U.S., this number is commonly 20, followed by a number indicating the distance a person with "normal" vision (20/20) would have to be to read something the person tested could read at 20 feet. Thus, a larger denominator would indicate poorer vision.

Other self-explanatory demographic information including body mass index (BMI), age, race, and gender are provided. We caution the users regarding the societal impact of using these labels since the underlying PRIME trial did not study the causality of these labels.

ML Centric Clinical Labels We describe BCVA and CST in this section. ETDRS best-corrected visual acuity (BCVA) is a visual function assessment performed by certified examiners where a standard vision chart is placed 4-meters away from the patient. The patient is instructed to read the chart from left to right from top to bottom until the subject completes 6 rows of letters or the subject is unable to read any more letters. The examiner marks how many letters were correctly identified by the patient. Central subfield thickness (CST) is the average macular thickness in the central 1-mm radius of the ETDRS grid. CST was obtained from the automated macular topographic information in the Heidelberg Eye Explorer OCT software.

The remaining clinical labels of Patient ID and Eye ID are self-explanatory and collected on clinical visits.

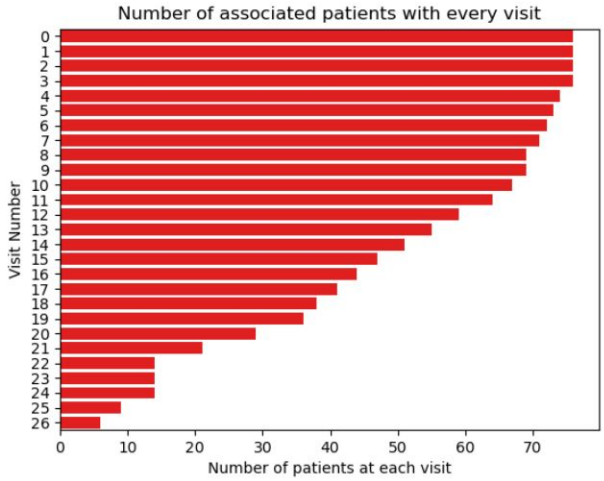

Figure 8: Number of patients at every visit within one of the training sets used for the treatment predicition analysis.

Figure 9: 1) A plot of average number of visits by patients that were an improvement or deterioration from previous week. Red bars indicate the standard deviation across all patients. 2) Plot of average change in BCVA with respect to the first week. 1) and 2) Polar representation plots with respect to the effective treatment of individual patients. The angle of rotation is $\theta = 360$ / (number of visits for patient). The angle rotates by θ counterclockwise if the collected BCVA on the current visit is better than that of the previous visit and clockwise by θ if it's worse.

B.5.3 Time Series Labels

The labels generated for the time-series experiments were based on changes in week to week BCVA values. For an individual visit, the treatment label was set to 1 if the following visit resulted in an improvement in BCVA and a 0 if the result wasn't an improvement. The goal is to predict whether the next visit would result in an improvement based on the associated modality (Fundus or 3D OCT Volume). Fig. 5 provides statistics regarding visit-wise changes of BCVA within the dataset. Further analysis of the dataset requires the number of patients treated on each visit which is provided in Fig. 8. As is apparent, the number of patients keep decreasing across visits. This can be for a variety of reasons all of which are discussed in the clinical trail publications at (2) and (3; 4; 5; 6). These numbers provide further context to the changes in Fig. 5. Presumably, as the treatment continues, it is the challenging patients who return for treatment and who qualify for injections. Their visit-wise average BCVA change skews the cohort in the negative direction in Fig. 5.

This change in treatment improvements can be understood through the polar plots of Figure 9. The polar representation plots show the effective treatment with respect to an individual patient. This

works by a vector beginning at the 0 degree point and rotating by an angle θ that is 360/(number of patient visits). After each turn counterclockwise, the hue of the associated color becomes darker by a fixed degree with a darker green hue indicating a higher degree of improvement and a darker red hue indicating a higher degree of worsening in the clockwise direction. This is shown in plots 1 and 2 in Fig. 9. Plot 1 shows that the patient had 4 rotations as indicated by the hue of green becoming darker by four degrees. This indicates that out of the total number of visits, this patient experienced 4 more visits with improvements, rather than deterioration. The converse is true for Plot 2.

C Additional Results

C.1 Multi-Modal Integration Between OCT and Biomarker/Clinical Labels

Table 9: Benchmark results for DR/DME detection showing precision and recall.

Experiments	Model	Precision		Recall	
		DR	DME	DR	DME
OCT	R-18	0.747	0.670	0.608	0.794
Clinical	MLP	0.753	0.756	0.758	0.751
Biomarker	MLP	0.703	0.870	0.826	0.771
OCT + Clinical	R-18 + MLP	0.888	0.765	0.566	0.952
OCT + Biomarker	R-18 + MLP	0.885	0.778	0.742	0.904

Experimental Details The greyscale B-scans are rescaled to 128×128 and normalized with $\mu = 0.482$ and $\sigma = 0.037$ as the baseline data for DR/DME detection. For OCT, we utilize Resnet-18 (R-18) ([37]) along with Adam optimizer and a learning rate of $1.5e-4$. There are 20 eyes in the test set; 10 having DR and the remaining DME. The validation set has 5 eyes with DR and the other 5 exhibiting DME. The train set is composed of the remaining 66 eyes, 26 of which have DR while 40 have DME. Therefore, we utilize $6,468$ images in the training set, $1,960$ images in our test set and 980 images in the validation set. For supervised learning with clinical labels, we train a shallow Multi Layer Perceptron (MLP) with two linear layers and Relu activation between. Biomarker features are normalized to zero mean and unit standard deviation. For supervised learning with biomarkers, we train a shallow MLP with four linear layers and LeakyRelu activation between. Biomarker features are normalized to zero mean and unit standard deviation. For multi-modal learning with OCT and clinical labels/biomarkes, we use the same train, test and validation split as the baseline OCT model and the clinical labels/biomarkers associated with each B-scan.

Optimization via Guided Loss Each modality is input to its independent model. At the output of the MLP biomarkers/clinical label model are logits $\phi^{MLP}(x_i)$, while the output logits of the Resnet OCT model are $\phi^{Resnet}(x_i)$. During optimization, learned features from one modality (biomarkers or clinical labels) are used to optimize the learning of the other (OCT features). The guided loss, $\mathcal{L}_{Guided}$, is one component of the overall loss function $\mathcal{L}$. Guided loss is the mean square error between MLP logits and Resnet logits. At every epoch, we minimize the disparity between these logits until the stopping criteria for training is met. The other two components, $\mathcal{L}_{Resnet}$ and $\mathcal{L}_{MLP}$ are binary cross entropy losses computed between the ground truth labels and logits from each model respectively. Collectively the three terms allow a joint optimization of both models and a transfer of knowledge from MLP model to Resnet model.

$$\mathcal{L} = \mathcal{L}_{Resnet} + \mathcal{L}_{MLP} + \mathcal{L}_{Guided} \tag{1}$$

$$\mathcal{L}_{Guided} = \mathbf{1}\left[\hat{y}^{MLP} = y\right] \frac{1}{2} \left|\left|\phi^{Resnet}(x_i) - \phi^{MLP}(x_i)\right|\right|_2^2 \tag{2}$$

Medical Perspective of Benchmark Results The results show that biomarkers as features are more effective at discriminating between the disease classes in both uni- and multi-modal training scenarios.

This makes sense from a medical perspective because biomarker features have direct correlation to the presence of DR/DME. Also, biomarker vectors assigned to any OCT slice are specific features that visually manifest themselves within that OCT slice. This means that biomarker features are fine-grained signs of diagnostic patterns indicative of disease. Clinical labels on the other hand are more coarse. Some clinical labels, like CST, represent characteristics of an OCT volume as a whole rather than any specific slice within a volume. Other clinical labels, like BCVA, are not derived from OCT and represent an evaluation of the eye as a whole. CST and BCVA are clinical parameters that are not indicative of a specific retinal disease diagnosis, but are instead representations of retinal anatomy or visual function, respectively. These two features are used in the context of monitoring retinal disease progression; thus, it is unsurprising that within a machine learning framework they yield sub-optimal performance in discriminating between disease classes.

C.2 Biomarker Interpretation with Contrastive Learning

A number of variations of contrastive learning exist in literature. The authors in (66) use the term contrastive to design visual explanations. They then extend these explanations to perform contrastive reasoning in inferential framework in (32).

This clinically aware supervised contrastive loss can be represented by:

$$L_{supcon_{clinical}} = \sum_{i \in I} \frac{-1}{|C(i)|} \sum_{c \in C(i)} log \frac{exp(z_i \cdot z_c / \tau)}{\sum_{a \in A(i)} exp(z_i \cdot z_a / \tau)} \tag{3}$$

where i is the index for the image of interest x_i. All positives c for image x_i are obtained from the set $C(i)$ and all positive and negative instances a are obtained from the set $A(i)$. Every element c of $C(i)$ represents all other images in the batch with the same clinical label c as the image of interest x_i. Additionally, z_i is the embedding for the image of interest, z_c represents the embedding for the clinical positives, and z_a represents the embeddings for all positive and negative instances in the set $A(i)$. Embeddings are obtained after passing the representations from an encoder network $f(.)$ through a projection head $G(.)$ that we set to be a multi-layer perceptron network. τ is a temperature scaling parameter that is set to .07 for all experiments. For example, a loss represented as L_{BCVA} indicates a supervised contrastive loss where BCVA is utilized as the clinical label of interest and all positives are chosen based on having the same BCVA value as the target image.

Training In this study, we leverage knowledge learnt from training on the large set of clinical labels to improve performance in classifying the smaller set of biomarkers. To test this setup, we take 76 eyes from the `OLIVES` Dataset to form a training set and take the remaining set of 20 eyes to form a test set. From this set of 20 eyes, we form an individual balanced test set for each biomarker by sampling 500 OCT scans with the biomarker present and 500 OCT scans with the biomarker absent. We train for 25 epochs and utilize a stochastic gradient descent optimizer with a learning rate of $1e-3$ and momentum of 0.9. The applied augmentations are random resize crops of size of 224, random horizontal flips, random color jitter, and data normalization to the mean and standard deviation of the respective dataset with a batch size set at 64. We use Intraretinal Hyperreflective Foci (IRHRF), Partially Attached Vitreous Face (PAVF), Fully Attached Vitreous Face (FAVF), Intraretinal Fluid (IRF), and Diffuse Retinal Thickening or Diabetic Macular Edema (DRT/ME) as the biomarkers in the study.

Medical Interpretation of Benchmark Results Another aspect of the results is how well the used clinical labels correspond with the biomarker classification performance. In all cases, the results act as validation to the hypothesis that taking advantage of correlations that exist with certain clinical labels is beneficial for biomarker detection of individual OCT scans. However, from a medical perspective, certain outcomes would intuitively be more likely. For example, the severity of IRF and DME tend to be correlated with CST due to higher levels of fluid corresponding to changes in CST. Therefore, it makes sense that the best performance for IRF and DME is associated with using CST values as the clinical label for the loss. Additionally, it can be observed in Fig. 4 that because BCVA and CST have different distributions of values, there is a different number of associated eyes and images for each respective clinical value. Effectively, this means that there is varying diversity with respect to any individual clinical label, which explains the varying performance depending on which clinical label is used. The Eye ID works due to images from the same eye having many features in common

that serve to identify a good positive set for the loss. However, from a medical perspective, the Eye ID alone does not confer any additional medical insight.

Method	Biomarkers					
	IRF	DME	IRHRF	FAVF	PAVF	AUROC
PCL (28)	**76.50**% ± .513	80.11% ± .335	59.1% ± 1.03	76.30% ± .378	51.40% ± .556	.767 ± .0017
SimCLR (29)	75.13% ± .529	80.61% ± .837	59.03% ± 2.54	75.43% ± .378	52.69% ± 2.68	.754 ± .0017
Moco v2 (30)	76.00% ± .305	82.24% ± 1.38	59.6% ± .702	75.00% ± .608	52.69% ± .472	.770 ± .0035
Eye ID	72.63% ± .264	80.2% ± .384	58% ± 2.56	74.93% ± 1.36	**65.56**% ± .200	.767 ± .0005
CST	75.53% ± .608	**83.06**% ± .213	**64.3**% ± 2.57	76.13% ± .264	62.16% ± 1.47	.790 ± .0006
BCVA	74.03% ± .351	80.27% ± .853	58.8% ± 1.82	**77.63**% ± .305	58.06% ± 1.27	.776 ± .0017

Table 10: We show the performance of supervised contrastive training on the OLIVES dataset. In this table we explicitly show the standard deviation for the average across three runs for both accuracy and AUROC.

C.3 Time-series Treatment Analysis

Experimental Procedure for Predicting Successive Treatment Effect To perform this experiment, we generate treatment effect labels. For every OCT volume or fundus image, we assign a label 1 if the following visit resulted in an increase in BCVA and a label 0 if the next week resulted in a decrease in BCVA. We then train models to perform this binary classification task of next visit improvement or deterioration. Each architecture is trained for 25 epochs with a SGD optimizer, learning rate of .0001, momentum of .9, and a batch size of 10. We use a Resnet-18 (37), ResNet-50 (37), DenseNet-121 (38), EfficientNet (39), and Vision Transformer (40) (using a patch size of 32, 16 transformer blocks, 16 heads in multi-attention layer). EfficientNet on OCT performs with the best results from Table 4. It can also be observed that the vision transformer model did not significantly improve over the traditionally CNN models. It is possible that the attention mechanism of the transformer needs further training and refinement to learn patches of importance within a Fundus image. This is especially true within the context of medical data as small fine-grained locations are oftentimes the most important and difficult to identify. From the overall results, it is clear that the main bottleneck to good performance is overfitting of the model towards a single class, which, in this case, was the treatment effect. This makes sense as these volumes tend to have more readily distinguishable features due to a more severe variation of the disease. Also, the best performance in Table 4 came when using OCT Volumes as well as the smaller ResNet-18 and EfficientNet models which may be due to having more data than in the Fundus case as well as less prone to overfitting due to a smaller size.

Medical Perspective of Predicting Successive Treatment Effect In Table 4, it is observed that the model is able to predict whether the next visit will experience an improvement or worsening of BCVA on the following visit with the associated performance seen in this table. From a medical perspective, indicators that predict whether treatment will be successful or not is not so clear simply from imaging data. This is reflected by performance measures for accuracy that are barely better than random chance. Part of the challenge is that responses to the treatment could potentially be due to factors independent from the imaging data. For example, lifestyle choices on the part of the patient could have a corresponding impact on how well the treatment is able to perform. Additionally, patients do not receive treatments equally due to the specific nature of an individual's condition, which limits predictability. In order to improve upon this benchmark, future studies should investigate the effect of utilizing more powerful time-series models as well as multi-modal fusion of fundus, OCT, clinical, and biomarker data.

Predicting Final Ocular State We perform a similar analysis with the biomarkers available from patient's first and final visit. In this analysis we explore the predictive power of the initial biomarkers at forecasting the biomarkers at the final visit. $3,234$ biomarkers were used as input features in the training set while 980 and 490 biomarkers were used for test and validation set respectively. A shallow MLP of two linear layers and a Relu activation was the model used in this analysis. Intersection over union served as the metric to evaluate the quality of the predicted final visit biomarkers relative to the ground truth. Figure 10, shown in Appendix C.3, shows the overall performance of the model on the test set. We see that initial biomarkers serve as good features for prediction of final ocular state.

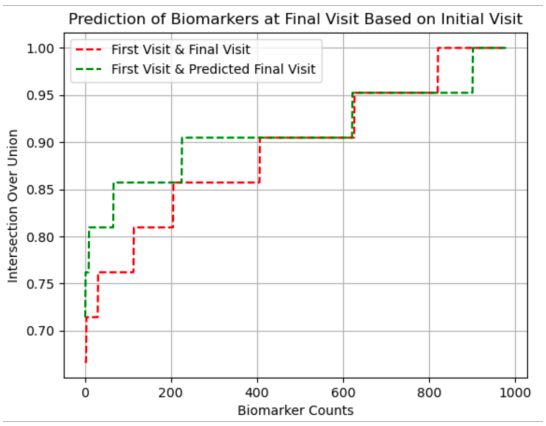

Figure 10: Benchmark Prediction of final visit biomarkers from initial visit biomarkers

Medical Perspective of Predicting Final Ocular State This axes of Fig. 10, show the number of biomarker vectors in the test set having varying intersection over union values. The red curve is the intersection over union (IOU) between the biomarkers at the first and final visits. This is the reference to show how final visit biomarkers changed relative to initial biomarkers. From the red curve, we see that of the 980 samples in the test set, only a few, approximately 200, have an IOU of 1. This means these biomarkers remained constant between first and final visits. A final visit biomarker being the same as the initial may not be an indicator of the patient's response to treatment. Rather, it indicates that no additional biomarkers manifested themselves at the final visit. The green curve represents the IOU between initial visit biomarkers and the predicted final visit biomarkers. Ideally, a complete overlap of red and green curves is desired. This would indicate that changes in biomarkers between first and final visits are being properly captured by the model. Complete overlap occurred for approximately 400 biomarkers whose IOU range from $0.90 - 0.95$. The time when the green curve exceeds the red indicates when there are larger intersection between first and predicted final visit biomarkers compared to the intersection between first and true final visit biomarkers. This means that the model predicted additional biomarkers within those biomarker vectors than what is actually present at the final visit. Conversely, there are a few cases when the green curve recedes the red and these are times when the model predicted fewer biomarkers within the biomarker vectors than the actual amount present at the final visit.

C.4 Other potential applications

The rich set of labels in the OLIVES dataset allows for utilizing the clinical labels and biomarkers in multiple ways. We demonstrate multi-modal fusion, medically-grounded contrastive learning, and time-series predictions in Section 4. In addition, Active Learning (67; 68) can utilize the clinical labels and biomarkers as indicators of disease states. The two paradigms of active learning - uncertainty and diversity - can be derived not through the model predictions, but from the auxiliary data in OLIVES. Similarly, biomarkers provide an annotated set of visual characteristics that show the manifestations of diseases within OCT scans. These biomarkers, along with the disease states and OCT scans, can be utilized for clinical reasoning. Other potential applications include domain difference analysis and adaptation which is described in Sections. C.5 and C.6 respectively.

C.5 Domain Shift

Domain shift based on PRIME and TREX trials Both the PRIME and TREX DME clinical trials are conducted using the same imaging equipment, in the same clinic. Hence, the domain difference w.r.t. PRIME and TREX is more due to the different disease manifestations they study and treat rather than imaging. We perform the biomarker detection experiments as detailed in Section 4.2 using PRIME and TREX trials separately. Specifically, we showcase the performance of intra-trial vs inter-trial experiments. Intra-trial refers to within PRIME and within TREX experiments - train and test within respective trials. Inter-trial refers to training and testing on different trials. The results are shown in Table 11. The best results are obtained when training and testing on TREX. This is because

Table 11: Benchmark results for characterizing domain shifts for data arising from the PRIME or TREX DME clinical trials.

Training Set	Test Set	Multi-Label AUROC
Prime	TREX	.649 ± .024
TREX	Prime	.547 ± .013
Prime	Prime	.599 ± .042
TREX	TREX	.727 ± .011

Table 12: Benchmark results for characterizing domain shifts before and after treatments in terms of first and last patient visits.

Training Set	Test Set	Multi-Label AUROC
First Visit	Last Visit	.628 ± .023
Last Visit	First Visit	.678 ± .012
First Visit	First Visit	.712 ± .026
Last Visit	Last Visit	.546 ± .018

of a larger diversity in TREX training data due to larger clinical trial window of 3 years. The eyes in TREX have more severe conditions than those present in PRIME. For this reason, combining the two dataset allows for a more complete distribution in terms of the severity of the disease and creates a more complete study and better results as shown in Table 3. Interestingly, the inter-trial results when training on TREX and testing on PRIME is higher than intra-trial training and testing on PRIME validating the need for larger diversity.

Domain shift before and after treatment The temporal element of the treatments studied in this dataset organically creates a domain shift between first and last visit's data for the same patient. We test this in Table 12. The experimental setup is biomarker detection from Section 4.2. From the results in Table 12, it is clear that there is a domain difference between the first and last visits based on the inter vs intra-visit training and testing setups. However, the results of the last visit maybe skewed because of the effects of treatment that may cause improvement in some and deterioration/no change in others that might cause intra-visit variation in the last visit thereby leading to lower results. Hence, we further conduct domain adaptation experiments for this modality in Appendix C.6 in the coarser setting of DR/DME detection.

Figure 11: A comparison between domain adaptation experiments using uni-modal and multi-modal data.

C.6 Domain Adaptation before and after treatment

In this section, we study the domain adaptation between the first and last week's worth of data in terms of DR/DME detection. With that in mind, we train models on four different training sets and use a fixed test set to evaluate the transfer of knowledge between domains. The first training set consists of OCT collected solely from the patients initial visit to the clinic. The remaining training sets consist of OCT from the first visit and 10/20/30 % of the OCT from the last visit respectively. The test set consists of the remaining 70% of OCT from the final visit. There is no overlap between train and test sets. Training is repeated three times with different seeds and an average balanced accuracy and standard deviation is noted. To compare the effect of multiple modalities on domain adaptation, the same experiments are repeated using OCT and biomarkers.

In Fig 11, the `x-axis` represents the composition of the training set used and `y-axis` the balanced accuracy achieved on the test set. Shaded regions around each curve show standard errors. The red curve highlights performance of OCT while the green curve shows the performance of OCT with biomarkers. We see that using data from the first visit only, results in the lowest performance for both curves. This training set has the largest disparity in its feature space compared to the test set. Conversely, the training set that combines 30% of data from the final visit with the first visit achieved the highest balanced accuracy for both curves due to higher similarity existing between the source and target domains.

C.7 Incorporation of Other Datasets alongside OLIVES

Table 13: Performance of leveraging data from a healthy dataset for a novel contrastive learning task is indicated by the Kermany + OLIVES row. Multi-Label is the average AUC from the multi-label classification task.

Severity Label Training Results (Accuracy / F1-Score)						
Method	IRF	DME	IRHRF	FAVF	PAVF	Multi-Label
SimCLR (29)	75.13% / .715	80.61% / .772	59.03% /.675	75.43% / .761	52.69% / .249	.754
PCL (28)	**76.50**% / .717	80.11% / .761	59.1% /.683	**76.30**% / .773	51.40% / .165	.767
Moco v2 (30)	76.00% / .720	82.24% / .793	59.6% / .692	75.00% / **.784**	52.5% / .201	.769
Kermany + OLIVES	75.20% / .698	81.46% / .786	**66.83**% / .695	75.39% / .756	54.7% / .314	**.774**

One of the useful features of the OLIVES dataset is that other medical datasets can be used in conjunction to develop other novel tasks. We incorporate the large amount of readily available healthy images from the Kermany dataset (13) to train an auto-encoder which is later utilized to generate anomaly scores on the unlabeled data in the OLIVES dataset. Results from using this setup is shown in Table 13 and the proposed strategy is identified with Kermany + OLIVES. We then use this anomaly score similar to a clinical label, within the contrastive learning setup in Section 4.2. We observe that leveraging this information out-performs standard state of the art contrastive learning strategies, but doesn't out-perform the clinical contrastive learning with respect to multi-label AUROC we show in Table 3. This example demonstrates the adaptability of the OLIVES dataset through its potential to leverage information in other datasets to develop novel perspectives that didn't exist in the OLIVES dataset originally.

C.8 Computational Resources

All experiments were run on PCs with two NVIDIA GeForce GTX TITAN X 12 GB GPUs.

D Datasheets

D.1 PRIME and TREX DME Clinical trials

The PRIME (2) and TREX-DME (3; 4; 5; 6) clinical trials included at least 96 eyes with either center-involving diabetic macular edema (CI-DME, n = 56) or diabetic retinopathy without CI-DME (DR, n = 40) between December 2013 and April 2021. Each participant signed an informed consent form to participate in the clinical trial. Both trials were prospective, randomized clinical trials. Prospective trials refer to longitudinal studies that evaluate the outcome of a particular disease during treatment. In PRIME, 40 eyes with nonproliferative diabetic retinopathy (NPDR) or proliferative

diabetic retinopathy (PDR) without CI-DME received intravitreal aflibercept injections (IAI) monthly until the eyes achieved a diabetic retinopathy severity scale (DRSS) score improvement of ≥ 2 steps; at baseline, eyes were randomized 1:1 into two management strategies for DR: 1) DRSS-guided or 2) panretinal leakage index (PLI)-guided management. In TREX-DME, 150 eyes with CI-DME were randomized 1:2:2 into three cohorts for management with ranibizumab (0.3 mg): 1) monthly treatment or 2) treat and extend, or 3) treat and extend with angiography-guided macular laser photocoagulation. For each patient, general demographics, ocular disease state data (e.g., best corrected visual acuity (BCVA)), central subfield thickness measurements (CST), and detailed ocular imaging (e.g., spectral-domain optical coherence tomography (SD OCT), fundus photography, and fluorescein angiography) was obtained per the protocol in Section B.5.2. All SD-OCT images were obtained using the Heidelberg Spectralis HRA+OCT (Heidelberg Engineering, Heidelberg, Germany) with a volume-per-cube acquisition protocol (20 x 20, 49 lines, 768 A-scans per line) with 9-times image averaging.

D.2 Clinical Study Process Description

The Intravitreal Aflibercept as Indicated by Real-Time Objective Imaging to Achieve Diabetic Retinopathy Improvement (PRIME) study was a prospective, randomized, phase II clinical trial (ClinicalTrials.gov identifier, NCT03531294; IND138997). The purpose of the study was to assess the safety and efficacy of as-needed intravitreal aflibercept injections for eyes with diabetic retinopathy without center-involved diabetic macular edema via the guidance of real-time Diabetic Retinopathy Severity Scale (DRSS) level or panretinal leakage index (PLI) assessment. The DRSS level was determined by color fundus photography graded by a trained image analyst. PLI assessment was conducted by an automated ultrawidefield fluorescein angiography image analysis platform. Between May 2018 and March 2019, forty subjects were enrolled in PRIME given the following inclusion criteria: 18 years of age and older with type 1 or type 2 diabetes mellitus, a DRSS level of 47A to 71A as determined by the CRC (Cole Eye Institute, Cleveland Clinic, Cleveland, Ohio, USA), and Early Treatment Diabetic Retinopathy Study (ETDRS) best-corrected visual acuity (BCVA) of 20/800 or better. The exclusion criteria consisted of CST greater than 320 μm in the study eye; central DME causing vision loss; vitreous hemorrhage; previous treatment of anti-vascular endothelial growth factor (VEGF) pharmacotherapies, corticosterids, dexamethasone, or fluocinolone acetonide in the study eye; and a history of vitrectomy or panretinal photocoagulation. Further details are available in (2).

The Treat and Extend Protocol in Patients with Diabetic Macular Edema (TREX-DME) study was a prospective, randomized, phase I/II, multicenter clinical trial (ClinicalTrials.gov identifier, NCT01934556). The purpose of the study was to compare the administration of intravitreal ranibizumab injections for eyes with center-involving diabetic macular edema on the basis of monthly dosing or a treat and extend algorithm with and without angiography-guided macular laser photocoagulation. Between November 2013 and April 2015, 150 eyes from 116 subjects were enrolled in TREX-DME given the following inclusion crtieria: type 1 or type 2 diabetes mellitus, center-involving DME, ETDRS BCVA between 79 and 24 letters (Snellen equivalent, 20/25-20/320). The exclusion criteria consisted of previous treatment of anti-VEGF pharmacotherapies, corticosteroids, or focal macular laser. Fifty-six out of the 150 study eyes were evalulated at Retina Consultants of Texas study sites and are included in this data set. Further details are available in (3; 4; 5; 6).

A summary of the two studies and the processes involved is provided as Summary-DR-DME-Studies.docx under the labels folder accessed through Appendix A.

D.3 ML Centric Label description

The ml centric labels directory within the label access provided in Appendix A consists of two CSV files. The first is the Biomarker-Clinical-Data-Images.csv. The labels for all 9408 images with biomarker alebls are provided in the csv file. Two screenshots of this CSV file from both the PRIME and TREX-DME trials are shown in Fig. 12 and 13 respectively. The first column provides a path within the image folder structure. The path includes the following:

1. Trial: Can refer to either PRIME or TREX-DME.
2. Arm: TREX-DME has an additional cohort-based subfolder within the trial: GILA, Monthly, and TREX that identify specific cohorts of patients based on treatment.

	A	B	C	D	E	F	G	H
1	Path (Trial/Arm/Folder/Visit/Eye/Image Name)	Scan (n/49)	Atrophy / thinning of retinal layers	Disruption of EZ	DRIL	IR hemorrhages	IR HRF	Partially attached vitreous face
2	/TREX DME/GILA/0201GOD/V1/OD/TREXJ_000000.tif	1	0	0	0	0	1	0.0
3	/TREX DME/GILA/0201GOD/V1/OD/TREXJ_000001.tif	2	0	0	0	0	1	0.0
4	/TREX DME/GILA/0201GOD/V1/OD/TREXJ_000002.tif	3	0	0	0	0	1	0.0
5	/TREX DME/GILA/0201GOD/V1/OD/TREXJ_000003.tif	4	0	0	0	0	1	0.0
6	/TREX DME/GILA/0201GOD/V1/OD/TREXJ_000004.tif	5	0	0	0	0	1	0.0
7	/TREX DME/GILA/0201GOD/V1/OD/TREXJ_000005.tif	6	0	0	0	0	1	0.0
8	/TREX DME/GILA/0201GOD/V1/OD/TREXJ_000006.tif	7	0	0	0	0	1	0.0
9	/TREX DME/GILA/0201GOD/V1/OD/TREXJ_000007.tif	8	0	0	0	0	1	0.0
10	/TREX DME/GILA/0201GOD/V1/OD/TREXJ_000008.tif	9	0	0	0	0	1	0.0

Figure 12: ML Centric Labels Datasheet within the OLIVES Dataset for PRIME trial

7020	/Prime_FULL/01-040/W104/OD/11.tif	12	0	0	0	0	1	1.0
7021	/Prime_FULL/01-040/W104/OD/12.tif	13	0	0	0	0	1	1.0
7022	/Prime_FULL/01-040/W104/OD/13.tif	14	0	0	0	0	1	1.0
7023	/Prime_FULL/01-040/W104/OD/14.tif	15	0	0	0	0	1	1.0
7024	/Prime_FULL/01-040/W104/OD/15.tif	16	0	0	0	0	0	1.0
7025	/Prime_FULL/01-040/W104/OD/16.tif	17	0	0	0	0	1	1.0
7026	/Prime_FULL/01-040/W104/OD/17.tif	18	0	0	0	0	0	1.0
7027	/Prime_FULL/01-040/W104/OD/18.tif	19	0	0	0	0	0	1.0
7028	/Prime_FULL/01-040/W104/OD/19.tif	20	0	0	0	0	0	1.0
7029	/Prime_FULL/01-040/W104/OD/20.tif	21	0	0	0	0	0	1.0
7030	/Prime_FULL/01-040/W104/OD/21.tif	22	0	0	0	0	0	1.0

Figure 13: ML Centric Labels Datasheet within the OLIVES Dataset for TREX-DME trial

3. Folder: Refers to the code that identifies each patient.
4. Visit: Refers to the visit that the current images and labels refer to. Note that in both the clinical studies, the biomarkers are retrospectively added to the first and last visits. Hence, in TREX-DME, the biomarkers are labeled at V1 and V22 for the first considered patient and so on.
5. Eye: The possible values are "OD" or "OS" and this serves to identify the right or left eye, respectively.
6. Image name: The name that is provided in the dataset directory on Zenodo.

Scan can be one of 49 slices that exists in a 3D volume which is obtained from the OCT machine for every patient for every visit. The next 16 columns refer to biomarkers the full list of which is present in Table 1 and whose generation process and abbreviations are described in Appendix B.5.1. 1 indicates their presence for the considered scan while 0 indicates their absence.

The last four columns refer to clinical labels - Eye ID, BCVA, CST, and Patient ID. The ranges of BCVA and CST are shown in Fig. 4 while their significance is expanded in Appendix B.5.2.

The second csv file under ml centric labels is Clinical-Data-Images.csv. This file holds the path name and only the four clinical labels for all $78,189$ scans.