TRACE: A Concept Bottleneck Model for Longitudinal 3D Glioblastoma Response Assessment

Alia Tarek[1,†], Hamsa Saber[1,†], Hamza Elghonemy[1,†], Youssef Afify[1,†], Tamer Basha[1], Omair Shahzad Bhatti[2], Abdulrahman M. Selim[2,3,*], Hasan Md. Tusfiqur Alam[2,*] and Daniel Sonntag[2,3]

[1]*Department of Biomedical and Healthcare Data Engineering, Faculty of Engineering, Cairo University, Giza, Egypt*
[2]*German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany*
[3]*University of Oldenburg, Oldenburg, Germany*

Abstract

Longitudinal glioblastoma response assessment requires comparing subtle tumor changes across MRI time points according to structured clinical criteria such as RANO. However, most deep learning approaches predict response labels directly from imaging features, which limits inspection and correction. Therefore, we introduce TRACE, an exploratory RANO 2.0-aligned concept bottleneck model for interpretable 4-class response classification evaluated on the LUMIERE dataset. TRACE processes paired baseline and follow-up multimodal MRI with a shared 3D vision encoder, predicts clinically meaningful root measurements from the tumor, computes downstream RANO-derived concepts deterministically, and adds scan interval and new-lesion information as passthrough concepts. This frames response assessment as structured concept reasoning rather than direct image-to-label prediction. Using 5-fold patient-wise cross-validation, TRACE achieves a 4-class macro F1 of 0.4769 ± 0.1229 and a binary progression-versus-non-progression macro F1 of 0.7085 ± 0.0935. It improves over a concept bottleneck baseline and remains within the range of published non-interpretable deep learning approaches. Ablations show that the expert RANO graph and intervention-consistency training are both important, and intervention experiments show that correcting concepts can improve downstream predictions. These findings support structured concept bottlenecks as a transparent direction for longitudinal GBM response assessment, while motivating larger protocol-aligned datasets for rare-class performance and external validation. Our codebase is publicly available online on GitHub.

Keywords

Concept Bottleneck Models (CBMs), Disease Progression, Longitudinal MRI, Glioblastoma (GBM), Response Assessment in Neuro-Oncology (RANO)

1. Introduction

Glioblastoma (GBM), a World Health Organization (WHO) grade 4 astrocytoma, is the most common and aggressive primary malignant brain tumor in adults, and accounts for about half of all gliomas [1]. Because disease progression is nearly universal, accurate detection of tumor recurrence on longitudinal Magnetic Resonance Imaging (MRI) is essential for timely treatment decisions. Clinicians commonly assess the treatment response using the Response Assessment in Neuro-Oncology (RANO) criteria. Although the updated RANO 2.0 criteria improve assessment by requiring confirmatory scans within 12 weeks after radiotherapy [2], conventional MRI alone has limited ability to distinguish treatment-related changes, such as pseudoprogression, from true tumor progression. This limitation highlights the need for more objective and automated decision-support methods for GBM treatment response assessment.

These challenges have motivated an increased interest in Artificial Intelligence (AI) methods for

EXPLIMED 2026 - Third Workshop on Explainable Artificial Intelligence for the medical domain - 15-17 August 2026, Bremen, Germany

*Corresponding authors.
†These authors contributed equally and are listed alphabetically.
tamer.basha@eng1.cu.edu.eg (T. Basha); omair_shahzad.bhatti@dfki.de (O. S. Bhatti); abdulrahman.mohamed@dfki.de (A. M. Selim); hasan.alam@dfki.de (H. Md. T. Alam); daniel.sonntag@dfki.de (D. Sonntag)
0009-0009-7824-4500 (A. Tarek); 0009-0005-9866-3375 (H. Saber); 0009-0007-5277-9545 (H. Elghonemy); 0009-0001-9486-0997 (Y. Afify); 0000-0003-4431-8646 (T. Basha); 0000-0001-7983-2384 (O. S. Bhatti); 0000-0002-4984-6686 (A. M. Selim); 0000-0003-1479-7690 (H. Md. T. Alam); 0000-0002-8857-8709 (D. Sonntag)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

CEUR-WS.org/Vol-4339/paper_5.pdf

longitudinal MRI analysis. AI-based approaches have been applied across the full spectrum of GBM care, from initial diagnosis through treatment planning and longitudinal monitoring. Among these tasks, treatment response assessment is especially challenging because it requires detecting small changes over time rather than findings from a single scan; this is difficult because adjacent time points can look highly similar, which can cause standard models to miss subtle biomarkers of disease evolution [3]. Deep learning approaches have shown promising results; for example, a 3D DenseNet model achieved strong performance in distinguishing pseudoprogression from true progression [4]. However, most AI models for GBM still face critical challenges in clinical use, such as data heterogeneity, small sample sizes, and limited interpretability [5, 6, 7].

Interpretability is particularly important because treatment response is assessed according to explicit clinical criteria. Deep learning models often do not expose the reasoning behind their predictions, a limitation that has been widely discussed in oncology AI and in brain tumor MRI analysis [8, 9, 10]. A common response has been to apply post-hoc explainability techniques to black-box architectures [11], such as Gradient-weighted Class Activation Mapping (Grad-CAM) [12]. However, such explanations are not always reliable or clinically meaningful. The limitations of post-hoc explainability have motivated a shift toward inherently interpretable models. Concept Bottleneck Models (CBMs) [13, 14] offer an inherently interpretable alternative by predicting explicit intermediate concepts before the final label. This is useful in medical imaging because clinicians can inspect concept predictions and, when needed, correct them at test time so that corrections propagate to the output [15, 16]. Despite growing interest in CBMs across medical imaging tasks, e.g., using causal CBMs [17] for chest x-ray interpretation that integrate radiologists' knowledge and demonstrate improved calibration, to the best of our knowledge, no prior work has applied CBMs to longitudinal brain-tumor progression or explicitly encoded temporal and causal dependencies in a concept graph.

In this paper, we introduce TRACE (**T**emporal Causal **R**easoning with **A**cyclic **C**oncept **E**xplanation), a RANO-aligned temporal CBM for interpretable GBM treatment response classification. TRACE processes paired baseline and follow-up multimodal MRI with a shared-weight siamese 3D encoder, predicts clinically grounded tumor measurements as root concepts, and computes downstream RANO-derived concepts through deterministic nodes. This frames longitudinal response assessment as a structured concept reasoning problem rather than a direct image-to-label prediction task.

Our contributions are as follows:

- We propose TRACE, a RANO-aligned CBM architecture for longitudinal GBM response assessment. To our knowledge, this is the first CBM designed for RANO-based longitudinal GBM response classification. Unlike standard CBMs that treat concepts as a flat set of predictors, TRACE represents clinically defined dependencies as a directed acyclic graph over tumor measurements, percentage changes, threshold flags, passthrough metadata, and the final response label.
- We encode deterministic RANO reasoning inside the bottleneck. Root tumor measurements are estimated from paired MRI, while downstream RANO quantities are recomputed from their parent concepts. This keeps derived concepts mutually consistent, constrains the task head to human-interpretable bottleneck variables, and allows concept-level corrections to propagate through the response pathway.
- We evaluate TRACE on the publicly available LUMIERE dataset [18] using 5-fold patient-wise cross-validation. TRACE achieves a 4-class macro F1 of 0.4769 ± 0.1229, improves over a data-efficient CBM baseline, and remains within the range of published non-interpretable deep learning baselines. We further use ablations, concept-effect analysis, and intervention experiments to examine which parts of the structured bottleneck contribute to performance and where the model remains limited.

Overall, this work shows that explicitly encoding temporal and clinically specified dependencies in a CBM can support more interpretable GBM response assessment. We present TRACE as an exploratory step under the practical constraints of small and imbalanced clinical datasets, and as a design pattern that may extend to other longitudinal medical imaging tasks where decisions depend on explicit clinical

criteria, temporal change, and human-verifiable intermediate measurements. Our codebase is publicly available online on https://github.com/yusufafify/TRACE-Temporal-Causal-Reasoning-with-Acyclic-Concept-Explanation.

2. Related Work

Prior studies on automated GBM MRI analysis have mostly focused on tumor segmentation and volumetric assessment instead of direct treatment response classification. Tools such as BraTumIA [19] and HD-GLIO-AUTO [20, 21, 22] can automatically segment tumors and estimate quantitative imaging measures from multimodal MRI. However, these systems still rely on clinicians to interpret the outputs and apply the RANO criteria for response assessment [23].

Recent work has begun to predict RANO response categories directly from longitudinal MRI using multi-timepoint CNNs [24], radiomics [25], or hybrid deep learning and radiomics models [26]. These methods provide important performance references, but they predict labels directly and do not address the need for clinically interpretable reasoning under the RANO framework. The difficulty is that RANO assessment depends on explicit clinical evidence, including spatial localization of new lesions, measurement of enhancing tumor components, and comparison across imaging time points. Models that predict only the final response label do not make this reasoning visible, making their outputs difficult to interpret and verify in practice. This limits their usefulness in clinical settings, where decision-support systems are expected to provide reasoning that clinicians can understand and assess [9].

2.1. Concept Bottleneck Models for Structured and Interpretable Prediction

CBMs provide a natural way to connect image-based prediction with structured clinical reasoning. Instead of mapping an input directly to a label, CBMs first predict explicit intermediate concepts and then use these concepts to derive the final decision [13]. This makes them well-suited to RANO-based response assessment, where the final class depends on clinically meaningful intermediate criteria rather than on image appearance alone.

Beyond static prediction, recent work has begun to extend concept bottlenecks to time-dependent data. X-CHAR [27] models complex human activity recognition as sequences of human-understandable concepts, while MoTIF [28] uses per-concept temporal self-attention to model concept dynamics in video classification. These works show that temporal concept modeling is an emerging direction, but the studied settings remain far from longitudinal medical response assessment, where predictions must follow explicit clinical criteria across patient time points.

Recent work has also extended CBMs in directions relevant to medical imaging. Some studies use large language models (LLMs) or vision-language models (VLMs) to define or align concepts automatically, reducing the need for fully manual concept annotation [29]. Others have shown that CBMs can improve robustness and reduce reliance on spurious correlations compared with standard visual encoders [30]. In breast ultrasound, Bunnell et al. [31] showed that clinician intervention on predicted concepts can improve downstream classification, using BI-RADS features as the concept set [32].

At the same time, standard CBMs still have practical limitations. They often require concept annotations, which are expensive to obtain, especially in specialized domains such as neuro-oncology. They can also require larger datasets and may still learn undesirable concept-label associations if the concept layer is not designed carefully [33, 34]. Several extensions have been proposed to address these issues; for example, Label-Free CBMs reduce the burden of manual annotation by using multimodal foundation models to discover or align concepts automatically [35, 36]; Data-Efficient CBMs (DCBM) aim to improve concept learning when only limited training data are available [37]; and Causal CBMs (C^2BM) further model structured dependencies between concepts and predictions to reduce reliance on shortcut correlations [17].

Within this broader literature, two CBM variants are especially relevant to our setting. **DCBMs** are designed for data-sparse scenarios, making them a natural point of comparison when training data are limited. In addition, **C^2BM** [17] is particularly relevant because it combines concept-based

prediction with an explicit directed acyclic graph (DAG) over concepts and the task label, which constrains information flow to predefined dependencies and supports propagation of concept-level interventions through the reasoning chain. For longitudinal GBM response assessment, these two directions address complementary needs, i.e., data efficiency and structured clinical reasoning. We compare these variants to identify which strategy better balances predictive accuracy and clinical interpretability for longitudinal GBM response assessment.

2.1.1. Directed Acyclic Graphs for Structured Clinical Reasoning

RANO assessment follows an explicit clinical decision process in which measurable findings, lesion status, and temporal evidence are combined to determine the final response category. DAGs provide a natural way to represent these dependencies in an explicit and interpretable form. In clinical and epidemiological research, DAGs are used to encode expert knowledge, make assumed causal pathways explicit, and support adjustment-variable selection in observational studies [38, 39]. More recently, DAG-based reasoning has also been discussed in clinical prediction modeling, where the goal is to ensure that models reflect clinically meaningful structure rather than relying only on correlations in the data [40]. This perspective is relevant here because RANO is already organized around structured dependencies between intermediate findings and the final response label.

However, a DAG alone does not define how these intermediate findings should be predicted from imaging data. For that, a model, such as C^2BM, is needed which can both learn concepts from MRI and preserve the structured reasoning process that links those concepts to the final response category.

3. Methodology

For treatment response assessment, clinicians compare measurable tumor changes between longitudinal MRI scans across two time points, a baseline scan t_0 and a follow-up scan t_1. In accordance with the RANO 2.0 criteria [2], we use two complementary measurements of tumor size, which are the tumor volume V in cm^3, estimated from segmentation masks, and the Sum of Products of Diameters (SPD), which is a 2D clinical measurement in cm^2 defined as the product of the two largest perpendicular tumor diameters on a single imaging slice. Both quantities are measured at baseline and follow-up, resulting in V_{t_0}, V_{t_1}, SPD_{t_0}, and SPD_{t_1}. Additional clinical background on SPD is provided in Appendix A.1.

The main idea of our method is to replace black-box temporal reasoning with an explicit, RANO-aligned concept reasoning chain. Rather than mapping paired MRI scans directly to a response label, the model first predicts clinically meaningful root concepts, derives downstream RANO concepts through deterministic operations, and then predicts the final response class from this structured bottleneck. This results in an auditable pathway:

$$(x_{t_0}, x_{t_1}) \longrightarrow \underbrace{C_r}_{\text{root concepts}} \longrightarrow \underbrace{C_d}_{\text{derived RANO concepts}} \longrightarrow \hat{Y}, \tag{1}$$

where C_r contains image-derived tumor measurements, C_d contains deterministic RANO deltas and threshold flags, and $\hat{Y}$ is the predicted RANO response category.

This design is motivated by the fact that RANO response assessment is already concept-based in clinical practice. Clinicians do not reason directly from voxel intensities to a response label; instead, they compare tumor measurements across time, assess threshold changes, consider new lesions and interpret the result according to a structured protocol. Our model encodes this reasoning structure as a DAG, making the intermediate steps inspectable and correctable.

3.1. RANO 2.0 Response Criteria

RANO 2.0 defines four response categories: Complete Response (CR), Partial Response (PR), Stable Disease (SD), and Progressive Disease (PD) [2]. These classes are determined by measurable changes

Table 1
RANO 2.0 response categories and decision thresholds. A measurability floor is applied: $V_{t_0} \geq 0.5\,\mathrm{cm}^3$ for volume-based criteria and $\mathrm{SPD}_{t_0} \geq 0.01\,\mathrm{cm}^2$ for SPD-based criteria.

Response	Volume criterion	SPD criterion
Complete Response (CR)	No measurable tumor	No measurable tumor
Partial Response (PR)	$\Delta V_{\%} \leq -65\%$	$\Delta\mathrm{SPD}_{\%} \leq -50\%$
Stable Disease (SD)	$-65\% < \Delta V_{\%} < +40\%$	$-50\% < \Delta\mathrm{SPD}_{\%} < +25\%$
Progressive Disease (PD)	$\Delta V_{\%} \geq +40\%$	$\Delta\mathrm{SPD}_{\%} \geq +25\%$

between baseline and follow-up tumors. We compute percentage changes as

$$\Delta V_{\%} = \frac{V_{t_1} - V_{t_0}}{V_{t_0} + \varepsilon} \times 100, \qquad \Delta\mathrm{SPD}_{\%} = \frac{\mathrm{SPD}_{t_1} - \mathrm{SPD}_{t_0}}{\mathrm{SPD}_{t_0} + \varepsilon} \times 100, \tag{2}$$

where ε is a small constant used for numerical stability.

Table 1 summarizes the decision thresholds used in this work. A response category can be triggered by either the volume-based or SPD-based criterion. In addition, the appearance of a new lesion directly supports PD under RANO. Volume and SPD are treated as complementary rather than redundant measurements. Volume captures the 3D tumor burden through both enhancing and non-enhancing tumor volumes (i.e., complementary tumor regions visible across different MRI sequences), while SPD retains compatibility with established clinical response criteria and historical neuro-oncology trials. Including both measurements allows the model to follow the full clinical RANO reasoning chain rather than relying on a single proxy for disease extent.

3.2. Structured Concept Graph

Standard CBMs [13] typically learn a mapping from input x to concepts C, followed by a mapping from concepts C to target label Y:

$$x \to C \to Y. \tag{3}$$

However, RANO concepts are not statistically independent attributes. Percentage changes are mathematical functions of baseline and follow-up measurements, and threshold flags are deterministic functions of the percentage changes. Treating these variables as independent concepts can create mutually inconsistent concept states and can weaken the reliability of test-time intervention.

Therefore, we encode the RANO reasoning process as a DAG

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \tag{4}$$

where each node in $\mathcal{V}$ corresponds to a concept or the task label, and each edge in $\mathcal{E}$ represents a clinically defined dependency. The graph contains four types of nodes:

- **Root concepts** C_r: image-derived tumor measurements predicted from MRI and segmentation inputs.
- **Derived concepts** C_d: deterministic RANO quantities computed from parent concepts.
- **Passthrough concepts** M: metadata or precomputed clinical indicators injected directly into the graph.
- **Task node** Y: the final 4-class RANO response prediction.

The complete concept set is shown in Table 2. Unlike causal CBM settings where the concept graph must be learned or inferred from data [17], our graph is specified directly from the clinical RANO protocol. This is a key advantage of the task, i.e., the relevant dependencies are known a priori and can be encoded as architectural constraints.

Table 2
Concepts in the proposed RANO-aligned concept graph. Root concepts are predicted from image features. Derived concepts are computed deterministically from their parents. Passthrough concepts are injected directly as metadata or precomputed clinical indicators.

Concept	Symbol	Type	Derivation / Source
Baseline volume	V_{t_0}	Root	Predicted from $\mathbf{X}_{\text{enc}}$ via a ConceptBlock
Follow-up volume	V_{t_1}		
Baseline SPD	SPD_{t_0}		
Follow-up SPD	SPD_{t_1}		
Volume change	$\Delta V_{\%}$	Derived	$(V_{t_1} - V_{t_0})/(V_{t_0} + \varepsilon) \times 100$
SPD change	$\Delta\text{SPD}_{\%}$		$(\text{SPD}_{t_1} - \text{SPD}_{t_0})/(\text{SPD}_{t_0} + \varepsilon) \times 100$
Volume PD flag	$b_{\text{PD}}^{\text{vol}}$		$\mathbf{1}[\Delta V_{\%} \geq 40\% \wedge V_{t_0} \geq 0.5]$
Volume PR flag	$b_{\text{PR}}^{\text{vol}}$		$\mathbf{1}[\Delta V_{\%} \leq -65\% \wedge V_{t_0} \geq 0.5]$
SPD PD flag	$b_{\text{PD}}^{\text{SPD}}$		$\mathbf{1}[\Delta\text{SPD}_{\%} \geq 25\% \wedge \text{SPD}_{t_0} \geq 0.01]$
SPD PR flag	$b_{\text{PR}}^{\text{SPD}}$		$\mathbf{1}[\Delta\text{SPD}_{\%} \leq -50\% \wedge \text{SPD}_{t_0} \geq 0.01]$
Scan interval	Δt	Passthrough	Administrative metadata
New-lesion flag	b_{NL}		Precomputed lesion-topology indicator

3.3. Theoretical Properties of the Structured Bottleneck

The proposed architecture is designed to satisfy two important properties for clinical interpretability, i.e., no hidden image-feature leakage into the final classifier and deterministic propagation of concept corrections. These properties align with recent views of actionable interpretability, in which interpretable models should use variables with human-aligned semantics and structures that users can reason about [41].

Proposition 1: No-leakage bottleneck. Let X denote the paired MRI and segmentation inputs, C_r the predicted root concepts, C_d the deterministic derived concepts, M the passthrough metadata, and $\hat{Y}$ the final prediction. If the task classifier receives only (C_r, C_d, M) and no direct image embedding, then

$$\hat{Y} \perp X \mid (C_r, C_d, M). \tag{5}$$

In other words, once the bottleneck variables are fixed, the final prediction has no additional access to image information.

Interpretation. This property does not claim that the learned root concepts are always perfectly accurate. Rather, it states that the final classifier cannot use hidden image features outside the concept graph. This reduces the risk of information leakage, a known problem in CBMs where unintended information beyond the intended concepts can be exploited by the label predictor [42]. This architectural guarantee applies to raw image features bypassing the bottleneck, but the learned concept embeddings passed to the task head may carry residual information beyond the scalar concept value.

Proposition 2: Deterministic intervention propagation. Let each derived concept $c_j \in C_d$ be a deterministic function of its parents:

$$c_j = f_j(\text{Pa}(c_j)). \tag{6}$$

Then any intervention on a root concept $c_i \in C_r$, denoted $do(c_i = a)$, induces a unique downstream concept state:

$$C_d^{do(c_i=a)} = F(C_r^{do(c_i=a)}, M). \tag{7}$$

Therefore, correcting a root concept automatically updates all downstream RANO quantities that depend on it.

Interpretation. For example, if a clinician corrects V_{t_1}, then $\Delta V_{\%}$, $b_{\text{PD}}^{\text{vol}}$, and $b_{\text{PR}}^{\text{vol}}$ are recomputed consistently. This is not guaranteed in a flat CBM, where all concepts are treated as independent predictors. The proposition formalizes why our graph is suitable for test-time concept correction.

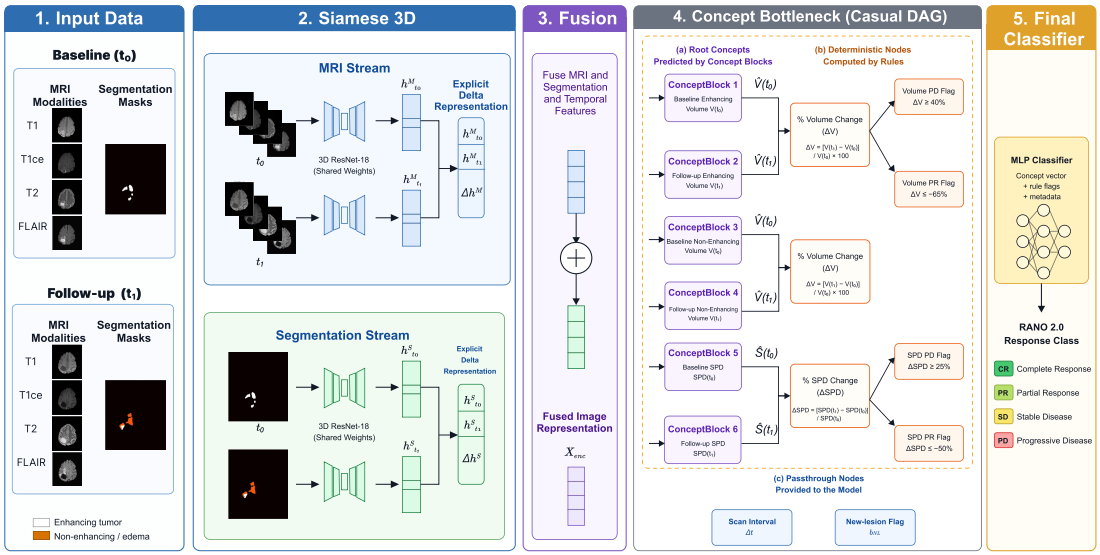


Figure 1: Overview of our proposed TRACE CBM. (1) Baseline and follow-up MRI with segmentation masks are provided as input; (2) shared siamese 3D encoders extract MRI- and segmentation-based features and explicit delta representations; (3) these features are fused into a joint representation; (4) a causal concept bottleneck combines predicted root concepts, rule-based deterministic nodes, and passthrough clinical metadata; (5) a final MLP predicts the RANO 2.0 response category (CR, PR, SD, or PD).

3.4. TRACE Architecture

Figure 1 summarizes the proposed architecture for TRACE. The model first encodes paired baseline and follow-up MRI volumes and segmentation masks into an image-only representation $\mathbf{X}_{\text{enc}}$. This representation is used to predict root concepts. All downstream RANO concepts are then either computed deterministically or injected as passthrough metadata. Finally, a task head predicts the 4-class RANO response.

3.4.1. Longitudinal Image Encoder

Image features are extracted using a siamese 3D CNN encoder. Two weight-sharing encoder instances process the baseline scan t_0, and follow-up scan t_1 separately. In our implementation, the encoder is a MedicalNet 3D ResNet-18 [43], pretrained on medical imaging tasks. The same framework can be instantiated with another 3D backbone.

We use two input streams; the first stream encodes the MRI volumes, where each scan is represented as a 4-channel tensor containing T_1, T_1^{ce}, T_2, and FLAIR, resampled to $128 \times 128 \times 128$ voxels; and the second stream encodes the segmentation masks as a separate multi-channel input. Since the first convolutional layer of MedicalNet is originally designed for single-channel input, we adapt it to multi-channel input by repeating the pretrained weights across channels and dividing by the number of channels to preserve activation scale. For each input volume, global average pooling over the final feature map produces a 512-dimensional feature vector:

$$\mathbf{h} = \text{AdaptiveAvgPool3D}(f_\theta(x)) \in \mathbb{R}^{512}. \tag{8}$$

For each stream, we represent the longitudinal pair using a temporal triplet:

$$\mathbf{T}(\mathbf{h}_{t_0}, \mathbf{h}_{t_1}) = \left[\mathbf{h}_{t_0} \parallel \mathbf{h}_{t_1} \parallel \mathbf{h}_{t_1} - \mathbf{h}_{t_0}\right] \in \mathbb{R}^{3\times 512}. \tag{9}$$

The MRI and segmentation triplets are concatenated and passed through a fusion layer g_ϕ:

$$\mathbf{X}_{\text{enc}} = g_\phi\left([\mathbf{T}^{\text{MRI}} \parallel \mathbf{T}^{\text{SEG}}]\right) \in \mathbb{R}^d, \tag{10}$$

where g_ϕ consists of a linear projection, layer normalization, and LeakyReLU activation. The representation $\mathbf{X}_{\text{enc}}$ is the only path through which image information enters the concept bottleneck. Non-image metadata, such as Δt and b_{NL}, bypass the encoder and enter the graph only through passthrough nodes.

3.4.2. Root Concept Prediction

The root concepts are the image-derived tumor measurements:

$$C_r = \left\{V_{t_0}, V_{t_1}, \text{SPD}_{t_0}, \text{SPD}_{t_1}\right\}. \tag{11}$$

Each root concept $c_k \in C_r$ is predicted from $\mathbf{X}_{\text{enc}}$ by a concept-specific MLP, called a ConceptBlock. The ConceptBlocks share the same input representation but have independent parameters, allowing each concept head to specialize in its assigned clinical measurement. For a root concept c_k, the ConceptBlock computes

$$\mathbf{e}_k = \text{MLP}_k(\mathbf{X}_{\text{enc}}) \in \mathbb{R}^{d_c}, \tag{12}$$

followed by a scalar prediction

$$\hat{c}_k = \mathbf{w}_k^\top \mathbf{e}_k + b_k. \tag{13}$$

The scalar prediction is trained against the z-scored ground-truth concept value. The embedding $\mathbf{e}_k$ is retained as the representation of the root concept for downstream prediction.

3.4.3. Graph Propagation

The graph is represented by an adjacency matrix

$$\mathbf{G} \in \{0, 1\}^{N \times N}, \tag{14}$$

where N is the number of concept nodes plus the task node, and $G_{ij} = 1$ denotes an edge $c_i \to c_j$. The graph is topologically sorted so that root concepts are computed first, followed by deterministic derived concepts, and finally the task node. At level 0, root concepts are predicted from $\mathbf{X}_{\text{enc}}$:

$$(\mathbf{e}_k, \hat{c}_k) = \text{ConceptBlock}_k(\mathbf{X}_{\text{enc}}), \qquad k \in C_r. \tag{15}$$

These are the only concept nodes that receive image features directly. Intermediate nodes are not learned propagators; they are either deterministic RANO nodes or passthrough nodes. This design reflects the structure of the clinical decision rule, i.e., once the root measurements and metadata are known, percentage changes and RANO threshold indicators are fully specified by fixed computations. Additional implementation details for the ConceptBlock formulation are provided in Appendix A.2.

3.4.4. Deterministic RANO Nodes

Deterministic nodes implement RANO quantities that are completely determined by their parents. These nodes have neither ConceptBlocks nor trainable parameters. Since root concepts are predicted in z-scored space, we first convert them back to the clinical scale before applying RANO computations. Let μ_k and σ_k denote the training-fold mean and standard deviation of concept c_k. The raw-scale prediction is

$$\tilde{c}_k = \text{expm1}(\hat{c}_k \sigma_k + \mu_k), \tag{16}$$

where the expm1 transformation is used when the concept labels were log-transformed during preprocessing. For example, the volume-based derived concepts are computed as

$$\widehat{\Delta V}_{\%} = 100 \cdot \text{clamp}\left(\frac{\tilde{V}_{t_1} - \tilde{V}_{t_0}}{\tilde{V}_{t_0} + \varepsilon}, -1, 5\right), \tag{17}$$

$$\hat{b}_{\text{PD}}^{\text{vol}} = \mathbf{1}\left[\widehat{\Delta V}_{\%} \geq 40 \ \wedge \ \tilde{V}_{t_0} \geq 0.5\right], \tag{18}$$

$$\hat{b}_{\mathrm{PR}}^{\mathrm{vol}} = \mathbf{1}\left[\widehat{\Delta V}_{\%} \le -65 \ \wedge \ \tilde{V}_{t_0} \ge 0.5\right]. \tag{19}$$

The SPD-derived concepts are computed analogously using the thresholds in Table 1. Deterministic nodes are excluded from the concept supervision loss because their values are fixed functions of their parents. Hard threshold indicators near the RANO boundaries could be replaced by a temperature-controlled sigmoid relaxation $\sigma((u-\tau)/T)$ to provide smoother gradients. In practice, the hard-threshold formulation was stable, so we use it as the default. Additional implementation details, including normalization, clipping, and the SPD-derived flags, are provided in Appendix A.3.

3.4.5. Passthrough Metadata Nodes

Two concepts are treated as passthrough nodes: the scan interval Δt and the new-lesion flag b_{NL}. These quantities are not predicted from image features; instead, their values are injected directly into the graph:

$$\hat{c}_k = c_k^*, \qquad k \in \{\Delta t, b_{\mathrm{NL}}\}. \tag{20}$$

This choice is clinically motivated because the scan interval is administrative metadata, while the new-lesion flag is computed from lesion topology across the two time points. Including these concepts as explicit passthrough nodes prevents the image encoder from being forced to represent non-image information, while still allowing the RANO graph to use clinically relevant context. Further details on the construction of the new-lesion flag and the handling of passthrough nodes are provided in Appendix A.4.

3.4.6. Task Prediction

The task node Y predicts the 4-class RANO response. It receives the outputs of its parent concepts, including root concept embeddings and scalar deterministic or passthrough values. We denote these inputs by $\mathbf{z}_{p_1}, \dots, \mathbf{z}_{p_m}$. The prediction is

$$\hat{\mathbf{y}} = \mathrm{softmax}\left(\mathrm{MLP}_Y\left([\mathbf{z}_{p_1} \parallel \cdots \parallel \mathbf{z}_{p_m}]\right)\right), \tag{21}$$

where $\hat{\mathbf{y}} \in \mathbb{R}^4$ is a probability distribution over $\{\mathrm{CR}, \mathrm{PR}, \mathrm{SD}, \mathrm{PD}\}$.

Although the task head is learned, it is constrained to operate only over the structured bottleneck. This makes its input inspectable and allows clinicians to apply corrections at the concept level.

3.5. Training Objective

The model is trained with a combination of task supervision, deep-supervision on the dominant PD decision, root-concept supervision, intervention-consistency regularization, and a small L_1 penalty on the first layer of every learnable concept propagator:

$$\mathscr{L}_{\mathrm{total}} = \mathscr{L}_{\mathrm{task}} + \alpha_{\mathrm{aux}}\mathscr{L}_{\mathrm{aux}} + \lambda_c\mathscr{L}_{\mathrm{concept}} + \beta_{\mathrm{cons}}\mathscr{L}_{\mathrm{IC}} + \gamma\left\|\theta_{\mathrm{prop}}^{(1)}\right\|_1. \tag{22}$$

In our final configuration $\lambda_c = 0.8$, $\alpha_{\mathrm{aux}} = 0.1$, $\beta_{\mathrm{cons}} = 0.5$, and $\gamma = 10^{-4}$. The L_1 term targets only the input layer of each propagator so that the bottleneck cannot route image information into the task head through a few high-magnitude weights; it acts as a shortcut suppressor rather than a generic regularizer.

Task loss. The primary task loss is a soft macro F1 objective combined with focal weighting [44], which directly optimizes the imbalanced evaluation metric:

$$\mathscr{L}_{\mathrm{task}} = 1 - \frac{1}{4}\sum_{j=1}^{4}\frac{2\sum_i(1-\hat{y}_{ij})^{\gamma_f}\, y_{ij}\,\hat{y}_{ij}}{\sum_i\left[(1-\hat{y}_{ij})^{\gamma_f}\, y_{ij} + \hat{y}_{ij}\right] + \varepsilon}, \tag{23}$$

where γ_f is the focal exponent applied to confident predictions and ε stabilizes the ratio for classes with few positives in a batch. The class-weighted cross-entropy $-\sum_j w_j y_j \log \hat{y}_j$ with inverse-frequency weights w_j is retained only as an ablation baseline.

Auxiliary PD head. The four RANO outcomes collapse into a clinically dominant PD-vs-rest decision. We add a deep-supervision auxiliary head that reads the same task logits and predicts the binary PD probability $\hat{y}_{\text{PD}}^{\text{bin}} = \hat{y}_{\text{PD}}$, supervised by class-weighted binary cross-entropy:

$$\mathscr{L}_{\text{aux}} = -\big[w_+ \, y_{\text{PD}}^{\text{bin}} \log \hat{y}_{\text{PD}}^{\text{bin}} + w_- \, (1 - y_{\text{PD}}^{\text{bin}}) \log(1 - \hat{y}_{\text{PD}}^{\text{bin}})\big], \tag{24}$$

with w_+, w_- inverse-frequency weights. This term provides a stable supervision signal for the most frequent clinical decision and indirectly stabilizes the rest of the bottleneck.

Concept loss. Concept supervision is applied only to learnable root concepts:

$$\mathscr{L}_{\text{concept}} = \frac{1}{|C_r|} \sum_{c_k \in C_r} \text{MSE}(\hat{c}_k, c_k^*). \tag{25}$$

Derived and passthrough nodes are excluded because their values are fixed functions of their parents or are injected directly. Concept supervision prevents the bottleneck from collapsing into arbitrary latent codes and keeps each ConceptBlock aligned with its assigned clinical measurement.

Encoder freezing. The 3D encoder is fine-tuned jointly with the concept and task heads during the first phase of training, and frozen at epoch 50 for the remainder of training. This stabilizes the concept-to-task mapping by preventing late-stage drift of the shared image representation.

3.6. Intervention-Consistency Training

A known failure mode of CBMs is that the task head adapts to systematic errors in the predicted concept distribution; when those concepts are replaced by correct values at test time, the input becomes distribution-shifted, and the final prediction degrades [45]. To prevent this train-test mismatch, we train the task head jointly under predicted and intervened concept states, and explicitly penalize disagreement between the two forward passes. Ablation results (In the upcoming Section 5.2) show that it is the largest single contributor to performance. During training, each root concept prediction is randomly replaced by its ground-truth value with probability p_{int}:

$$\tilde{c}_k = m_k c_k^* + (1 - m_k)\hat{c}_k, \qquad m_k \sim \text{Bernoulli}(p_{\text{int}}). \tag{26}$$

The deterministic downstream RANO concepts are then recomputed from the mixed root state, $\tilde{C}_d = F(\tilde{C}_r, M)$, and the intervened prediction is $\tilde{\mathbf{y}} = g_Y(\tilde{C}_r, \tilde{C}_d, M)$. We penalize disagreement between the fully predicted and the intervened predictions with the Kullback–Leibler (KL) divergence term

$$\mathscr{L}_{\text{IC}} = D_{\text{KL}}\left(\hat{\mathbf{y}} \,\|\, \tilde{\mathbf{y}}\right), \tag{27}$$

where $\hat{\mathbf{y}} = g_Y(\hat{C}_r, \hat{C}_d, M)$. We anneal the intervention probability linearly over the first 15 epochs,

$$p_{\text{int}}(e) = \min\!\left(p_{\max},\ p_{\min} + (p_{\max} - p_{\min}) \cdot \frac{e}{E_{\text{anneal}}}\right), \qquad p_{\min} = 0.3,\ p_{\max} = 0.8,\ E_{\text{anneal}} = 15, \tag{28}$$

so that the task head first learns from largely predicted concepts and then increasingly relies on corrected ones. This schedule, together with the KL term, makes the model robust to clinically plausible concept corrections at test time, which is the practical setting in which a clinician can audit or override an intermediate RANO measurement. We also considered monotonicity constraints on the task head, but treat them as an optional extension rather than part of the final model; details are provided in Appendix A.5.

3.7. Adapting to Longitudinal RANO Assessment

Our model builds on the causal CBM from Felice et al. [17], but adapts the framework to longitudinal GBM response assessment. We extended the original CBM from single-image prediction to paired baseline and follow-up multimodal MRI. A siamese 3D encoder processes both time points, together with their segmentation masks, and represents the longitudinal change through feature-level differences. This allows the model to predict clinically meaningful root concepts from volumetric temporal inputs rather than from isolated 2D images.

The graph is also adapted to the RANO setting. Rather than learning the concept structure from data, we manually specify the DAG from the RANO decision process described in Section 3.1. Image information enters the bottleneck only through the root concepts, while downstream RANO concepts, such as percentage changes and threshold flags, are computed deterministically from their parents. This prevents derived concepts from carrying hidden image embeddings to the task head and keeps the reasoning pathway auditable. We also use intervention-consistency training, with corrected concept values shown to the task head during training, so that concept-level corrections can propagate more reliably to the final response prediction. Overall, the method converts longitudinal RANO classification into a structured concept reasoning problem designed for small and imbalanced neuro-oncology datasets.

4. Experiments

4.1. Dataset

For our experiments, we selected the LUMIERE dataset [18] because it provides scan-pair-level RANO response labels assigned through formal response assessment, covering the four response categories CR, PR, SD, and PD. This is important for our setting because Trace is designed around the RANO decision process. The model uses image-derived measurements, deterministic threshold flags, and passthrough clinical metadata, so the training labels must match the same response assessment framework. We also considered using other datasets, for example [46, 47], but did not use them for training because they do not provide equivalent RANO annotations at the scan-pair level. Reconstructing RANO labels from segmentation-derived volume change introduced label ambiguity rather than a like-for-like supervision signal. Further details are provided in Appendix A.6.

LUMIERE contains data from 91 patients. We ended up with 599 scan sessions after filtering for complete modality sets and valid longitudinal follow-ups. Baseline and follow-up pairs are constructed from patients with at least two valid consecutive time points, resulting in 361 longitudinal pairs, with a mean of 4.8 visits per patient and a median follow-up of 53 weeks. Each session contains four spatially aligned, skull-stripped 3D MRI sequences (T1, T1ce, T2, FLAIR) resampled to 1×1×1 mm and an automatically generated tumor segmentation from HD-GLIO-AUTO [20] with three labels: background, non-enhancing tumor, and contrast-enhancing tumor. The segments are used to compute the volumetric concepts described in Sections 3.1 and 3.2. Figure 2 shows an example segmentation case.

4.1.1. Preprocessing

We applied a four-step preprocessing pipeline to improve spatial and intensity consistency across patients and longitudinal scans. First, all MRI sequences were registered to the MNI152 1 mm template to support voxel-wise correspondence across subjects. Second, intensity inhomogeneity was corrected using N4 bias field correction [48]. Third, histogram matching was applied to reduce inter-scan intensity variation. Finally, each volume was z-score normalized. Segmentation masks were transformed to the same template space using the MRI-to-template transformation and resampled with nearest-neighbor interpolation to preserve discrete labels. The pipeline follows the approach in [26], with segmentation handling adapted to the concept-based framework. Additional details are provided in Appendix A.7.

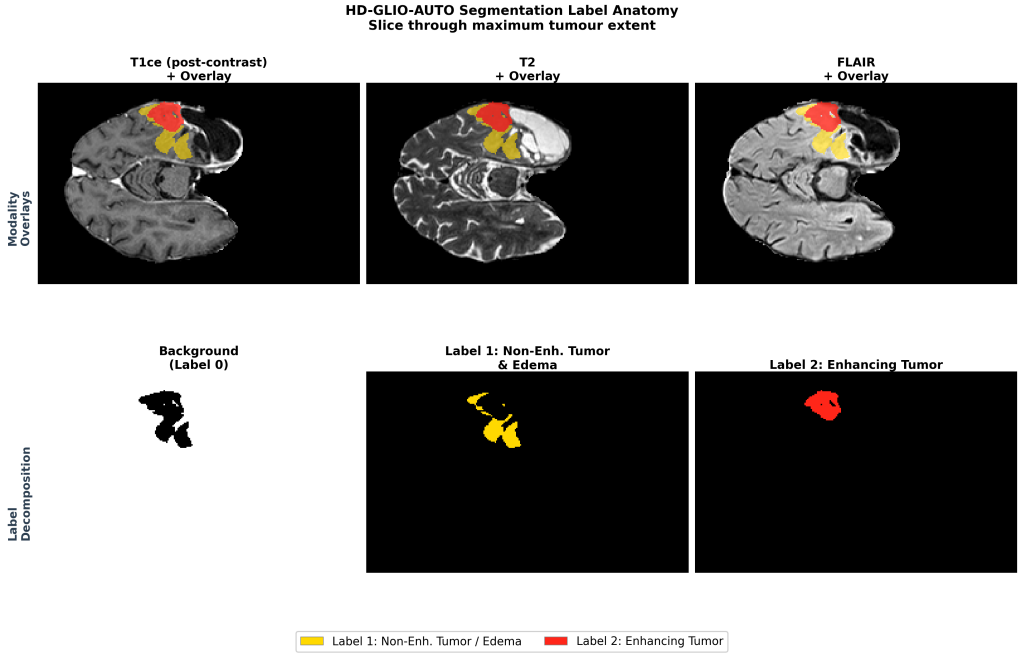


Figure 2: Representative segmentation case from LUMIERE. Top row: three MRI modalities with the HD-GLIO-AUTO overlay. Bottom row: non-enhancing (yellow) and enhancing (red) tumor labels in isolation. These regions are used to compute volumetric concepts for TRACE.

4.1.2. Label Distribution

The RANO label distribution is strongly imbalanced, as shown in Figure 3, with PD accounting for 63.7% of the 361 pairs. This imbalance affects both training and evaluation, increasing the uncertainty of per-class estimates and making plain accuracy less informative. Therefore, we use macro F1 as the primary metric and address class imbalance during training, as described in Section 4.3. The small number of longitudinal samples also motivates the use of a pretrained 3D encoder and concept-level supervision rather than training a large 3D classifier from scratch.

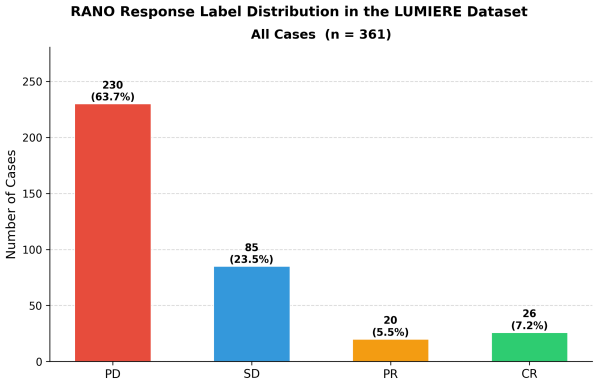


Figure 3: RANO label distribution across all 361 baseline–follow-up pairs in LUMIERE. PD accounts for 63.7% of samples, leading to a strongly imbalanced 4-class classification task.

4.2. Baseline Models

We compare TRACE against two baseline groups.

Black-Box Deep Learning Baselines. First, against three published black-box deep learning baselines evaluated on LUMIERE [24, 26, 25]. For Matoso et al. [24], we use the replicated values reported

by Amato et al. [25], as they follow the same data splits as our experiments. For Tikhonov et al. [26], we report the published values, which were consistent with our reproduction. These baselines provide useful performance references, but they do not encode the RANO decision process or support the same concept-level inspection and correction as TRACE.

Interpretable Baselines. Second, against interpretable baselines, including our implementation of DCBM [37] and two rule-based RANO oracle references. DCBM tests whether automatically discovered concepts from frozen 3D medical embeddings can support longitudinal RANO response classification without explicit concept supervision; we adapt DCBM to paired MRI scans by extracting patch-level embeddings at baseline and follow-up, learning sparse concept activations, and classifying longitudinal changes in those activations; implementation details are provided in Appendix A.8. The RANO oracle references apply deterministic threshold rules to either ground-truth concepts (GT concept) or predicted concepts (predicted concepts), isolating the behaviour of threshold-only decision logic and its sensitivity to concept-estimation noise.

4.3. Experimental Setup

We evaluate TRACE on both 4-class RANO response classification and binary progression-versus-non-progression classification, where PD is compared against all non-PD responses.

Cross-validation. We used 5-fold Stratified Group K-Fold cross-validation. Stratification is applied at the scan-pair level to preserve the RANO class distribution across folds, while grouping ensures that scans from the same patient do not appear in both the training and validation splits.

Backbone and inputs. The TRACE image encoder is a 3D MedicalNet ResNet-18 [43], with the four MRI modalities and three segmentation channels provided at 128×128×128 resolution, in bf16 mixed precision. The encoder is fine-tuned jointly with the concept and task heads during the first phase of training and frozen at epoch 50 to stabilize the concept-to-task mapping.

Loss and intervention training. We train with a soft macro F1 loss combined with focal weighting, so that optimization is better aligned with the primary metric under class imbalance. We also add a deep-supervision auxiliary binary head on the PD probability of the 4-class output, with weight $\alpha_{\text{aux}} = 0.1$, to provide a stable PD-vs-rest supervision signal. Intervention-aware training [17] is used with intervention probability p_{int} annealed from 0.3 to 0.8 over the first 15 epochs. An intervention-consistency KL term between the predicted-concept and intervened-concept forward passes is added with weight $\beta_{\text{cons}} = 0.5$. We apply an L_1 penalty of 10^{-4} to the first layer of the task head to discourage reliance on a small number of high-magnitude bottleneck inputs.

Optimization. All variants are trained with AdamW [49], learning rate 10^{-3}, L_2 weight decay 0.01, batch size 16, and a maximum of 100 epochs. A `ReduceLROnPlateau` scheduler monitors validation macro F1, and early stopping is applied after 20 epochs without improvement. The concept loss weight is $\lambda_c = 0.8$, and the checkpoint with the best validation macro F1 is retained. Full hyperparameters are listed in Appendix A.9.

Metrics. We used macro F1 over the four RANO classes as the primary metric. We also report weighted accuracy, per-class F1, and binary PD-vs-rest macro F1 to assess performance beyond the aggregate 4-class score. For interpretability, we report CaCE-TV [50], Concept-Alignment Score, and an Ordered Intervention Score; these metrics measure concept influence, concept prediction quality, and the effect of ordered concept interventions, respectively.

4.4. Ablation Studies

We used ablation studies to test which parts of TRACE contribute to performance. The ablations focus on two aspects of the model, i.e., the training objective and the expert RANO graph. All variants are evaluated on fold 4, which is the strongest individual fold and provides a consistent reference for comparing effect direction. Each ablation starts from the full configuration and disables one component while keeping the remaining hyperparameters fixed. Additional ordered-versus-random intervention policy ablation results are provided in Appendix A.10.

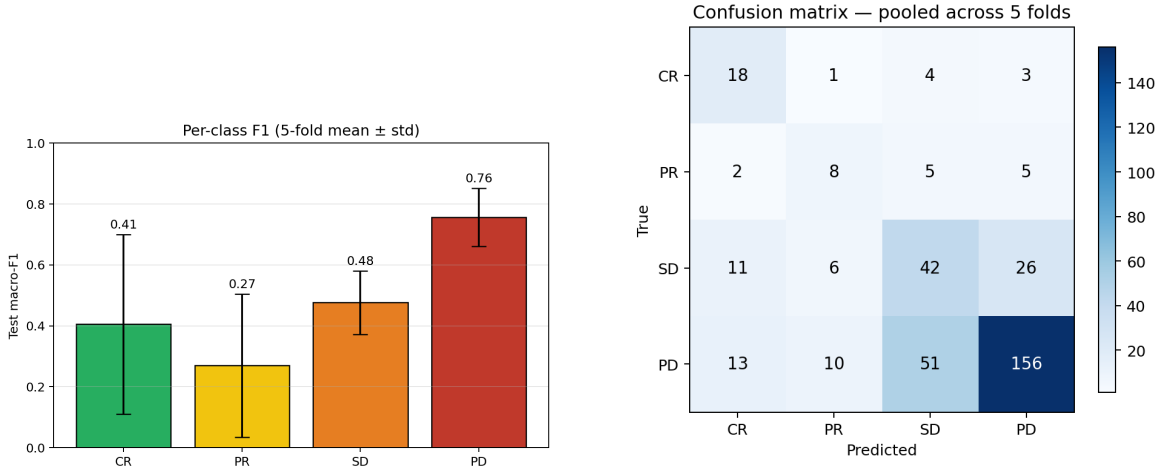
- **A0 (full model):** uses the expert RANO DAG, soft macro F1 with focal weighting, the auxiliary PD head, intervention-consistency KL, the deterministic derivation chain, and L_1 regularization on the first layer of the task head.
- **A1 (−consistency KL):** sets $\beta_{\text{cons}} = 0$. This tests whether aligning the predicted-concept and intervened-concept forward passes is needed to maintain a stable bottleneck under concept correction.
- **A3 (−auxiliary PD head):** sets $\alpha_{\text{aux}} = 0$. This tests whether the PD-vs-rest deep-supervision signal contributes beyond the 4-class task loss.
- **A4 (cross-entropy task loss):** replaces the soft macro F1 with focal weighting by plain cross-entropy. This tests whether the macro F1-aligned task loss improves performance under the strong class imbalance in LUMIERE.
- **B1 (bipartite graph):** replaces the expert RANO DAG with a flat bipartite CBM. In this setting, the deterministic derivation chain and passthrough nodes are disabled because the bipartite graph does not encode the parent-child dependencies required for closed-form RANO computation. This tests whether the clinical dependency structure adds value beyond exposing the same concepts to the classifier.
- **B2 (−deterministic chain):** keeps the expert DAG but disables the deterministic derivation chain, forcing derived concepts to be predicted by learnable graph modules. This isolates the value of deterministic RANO computation from the value of the graph structure itself.

5. Results and Discussion

We evaluate TRACE on 4-class RANO treatment-response classification using 5-fold patient-wise cross-validation on the LUMIERE dataset [18]. Across folds, TRACE achieves a 4-class macro F1 of 0.4769 ± 0.1229 and a weighted accuracy of 0.5447 ± 0.1318. On the clinically important PD-vs-rest binary decision, macro F1 reaches 0.7085 ± 0.0935, and weighted accuracy reaches 0.7329 ± 0.1104. The main source of variation is the small number of CR and PR cases in each fold; the detailed fold-level results are provided in Appendix A.11. Figure 4a shows that PD achieves the highest and most consistent F1 across folds, whereas PR has the largest fold-to-fold variation, reflecting the small number of PR cases in several validation folds. The pooled confusion matrix in Figure 4b shows that most errors occur between clinically adjacent response categories, especially SD versus PD and CR/PR versus SD. This is consistent with the RANO task itself, where class boundaries depend on continuous tumor-burden measurements and threshold-based changes that can be sensitive to segmentation and measurement noise. The full per-class table is provided in Appendix A.12.

5.1. Baseline Comparison.

Table 3 compares TRACE with published black-box baselines [24, 25, 26] and the concept-based DCBM baseline [37]. We also report a *rule-based RANO oracle*, which applies the textbook RANO decision tree directly to the ground-truth concept vector for each validation pair. The oracle assigns PD when a new lesion is present, $\Delta V_{\text{enh}} \geq 40\%$, or $\Delta\text{SPD} \geq 25\%$, with analogous threshold rules for PR and CR; this provides a threshold-only reference with perfect access to the concepts that TRACE must estimate from MRI, but without a learned task head.



(a) Per-class F1 across folds.

(b) Pooled five-fold confusion matrix.

Figure 4: Per-class performance and pooled errors on LUMIERE.

Despite using ground-truth concepts, the Rule-based RANO oracle (GT concept) reaches only 0.346 ± 0.038 4-class macro F1 and 0.681 ± 0.058 binary macro F1, which is 13 percentage points below TRACE on the 4-class task. It also recovers none of the 26 ground-truth CR cases and misses 41% of PD cases across the five folds. This suggests that the LUMIERE labels are not fully determined by segmentation-derived thresholds alone, likely because expert RANO assessment can also reflect non-measurable disease, steroid status, clinical deterioration, and radiologist judgment. Using the same threshold rules, the Rule-based RANO oracle (predicted concepts) further reduces macro F1 to 0.238 ± 0.041, showing that hard-threshold decisions are sensitive to concept estimation noise.

TRACE also improves over DCBM, increasing macro F1 from 0.254 to 0.4769 and weighted accuracy from 0.313 to 0.5447. The strongest black-box baselines retain a macro F1 advantage [25, 26], but they rely on direct label-discriminative image or radiomic features and do not constrain prediction through an auditable RANO-aligned concept pathway. The DenseNet-264 result [24] further shows that end-to-end classifier capacity alone is not enough to resolve the class imbalance in this dataset. Overall, TRACE offers a different trade-off as it improves over concept-based alternatives while preserving concept-level inspection and correction.

Table 3
RANO classification on LUMIERE reported as mean ± std. Values are rounded to three decimal places. The results reported by Matoso et al. [24] are based on the replicated values from Amato et al. [25]. The accuracy type varies across methods due to differences in how published results are reported; therefore, direct numerical comparisons should be treated with caution.

Model	Macro F1	Accuracy Score		Interpretable
Matoso et al. [24] (+ Radiomic Features)	0.505 ± 0.098	0.470 ± 0.091	Balanced acc.	×
Amato et al. [25]	0.616 ± 0.048	0.322 ± 0.055	Balanced acc.	×
Tikhonov et al. [26]	0.500 ± 0.080	0.720 ± 0.050	Accuracy	×
DCBM Baseline	0.254 ± 0.040	0.313 ± 0.070	Balanced acc.	
Rule-based RANO oracle (GT concepts)	0.346 ± 0.038	0.601 ± 0.037	Accuracy	
Rule-based RANO oracle (predicted concepts)	0.238 ± 0.041	0.645 ± 0.044	Accuracy	
TRACE (Our Approach)	**0.477 ± 0.123**	**0.545 ± 0.132**	**Weighted acc.**	

5.2. Ablation Results

We use the fold 4 ablations in Table 4 to examine which parts of TRACE contribute most to performance. Each variant starts from the full model A0 and disables one component while keeping the remaining hyperparameters fixed, as explained in Section 4.4. The ablations focus on two parts of the design, which are the training objective and the expert RANO graph.

Table 4
Component ablation on fold 4. A0 is the full model; each row disables one component. Negative Δ indicates the drop relative to A0.

ID	Variant	4-class F1	Binary F1	Δ F1 vs A0
A0	Full model	0.6963	0.8572	–
A1	−intervention-consistency KL	0.3538	0.6000	−0.342
A3	−auxiliary PD head ($\alpha_{\text{aux}} = 0$)	0.6078	0.7565	−0.089
A4	Cross-entropy task loss	0.6635	0.7159	−0.033
B1	Bipartite graph	0.3549	0.5921	−0.342
B2	Expert DAG without deterministic derivation	0.3214	0.5774	−0.375

Core bottleneck components. The largest drops occur when removing intervention-consistency training or disrupting the expert RANO graph. Removing the intervention-consistency KL term (A1) reduces 4-class macro F1 from 0.6963 to 0.3538 and binary macro F1 from 0.8572 to 0.60; this supports the role of the consistency term in reducing the mismatch between predicted-concept inputs and intervened-concept inputs, which is important for stable test-time concept correction. Replacing the expert RANO DAG with a flat bipartite graph (B1) produces a similar reduction, while keeping the expert DAG but disabling deterministic RANO derivation (B2) gives the largest drop, reducing 4-class macro F1 to 0.3214.

The graph ablations clarify the role of the structured bottleneck. B1 shows that exposing the concepts to the classifier is not sufficient when the clinical dependency structure is removed. B2 shows that the expert graph is also not sufficient if derived RANO quantities are learned rather than computed deterministically from their parent concepts. In other words, the main benefit comes from combining the graph structure with deterministic RANO computation, because root measurements are estimated from MRI, while percentage changes and threshold flags are recomputed by fixed rules before being passed to the task head.

Auxiliary supervision and loss design. The remaining ablations show that the auxiliary and loss terms also contribute, but with smaller effects on the 4-class task. Removing the auxiliary PD head (A3) reduces 4-class macro F1 by 8.9 percentage points and binary macro F1 by 10.1 percentage points, suggesting that the PD-vs-rest supervision signal helps stabilize the clinically dominant decision boundary. Replacing the soft macro F1 with focal weighting by plain cross-entropy (A4) has a smaller effect on 4-class macro F1 (−3.3 percentage points), but a larger effect on the binary PD-vs-rest task (−14.1 percentage points). This suggests that the macro F1-aligned loss is useful under the strong class imbalance in LUMIERE, especially for the PD decision.

Relation to the rule-based oracle. These ablations also help interpret the rule-based RANO oracle in Table 3. The oracle tests whether fixed thresholding of ground-truth concepts is sufficient, while B1 and B2 test what happens when the learned model loses either the expert dependency structure or the deterministic RANO derivation chain. Together, these results suggest that TRACE benefits from both parts of the design, where deterministic RANO computation provides a clinically constrained reasoning path, while the learned task head can account for aspects of the expert labels that are not fully captured by segmentation-derived thresholds alone.

5.3. Causal Effect of Concepts (CaCE) and Concept Alignment

We assess whether TRACE uses the bottleneck in a clinically meaningful way through Causal Effect of Concepts total variation (CaCE-TV) [50, 17]. For each concept, CaCE-TV measures the change in

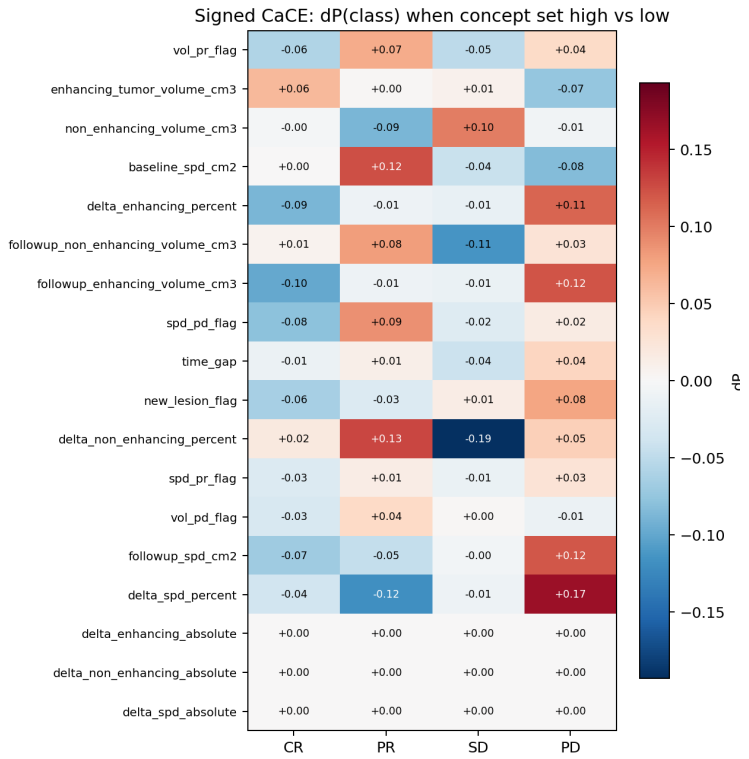


Figure 5: Signed CaCE per concept and per RANO class. Longitudinal-change concepts shift probability mass in directions consistent with RANO response categories.

the 4-class output distribution when the concept is intervened from a low-value to a high-value state. Table 5 reports the top concepts by CaCE-TV, and Figure 5 shows the signed per-class effects.

Table 5
Top concepts by CaCE-TV. Higher values indicate a stronger effect on the predicted RANO distribution under intervention.

Rank	Concept	CaCE	Std
1	delta_non_enhancing_percent	0.195	0.131
2	delta_spd_percent	0.170	0.053
3	baseline_spd_cm2	0.143	0.048
4	followup_spd_cm2	0.129	0.030
5	followup_enhancing_volume_cm3	0.127	0.015
6	delta_enhancing_percent	0.116	0.012
7	followup_non_enhancing_volume_cm3	0.115	0.084
8	spd_pd_flag	0.111	0.059
9	vol_pr_flag	0.111	0.070
10	non_enhancing_volume_cm3	0.108	0.053
11	new_lesion_flag	0.091	0.024
12	enhancing_tumor_volume_cm3	0.081	0.024
13	time_gap	0.053	0.032
14	vol_pd_flag	0.053	0.026
15	spd_pr_flag	0.049	0.014
16	delta_enhancing_absolute	0.000	0.000
17	delta_non_enhancing_absolute	0.000	0.000
18	delta_spd_absolute	0.000	0.000

The CaCE results show two useful patterns. First, concept influence is distributed across the bottleneck rather than concentrated in one variable. Ten of the eighteen concepts exceed a CaCE-TV value of 0.1,

with a mean CaCE-TV of 0.110 across non-trivial concepts (excluding the three concepts with CaCE-TV of zero). Second, the highest-effect concepts are mainly RANO-relevant change and size variables, including percentage change in non-enhancing tumor, percentage change in SPD, follow-up enhancing volume, baseline SPD, and percentage change in enhancing tumor. Their signed effects in Figure 5 generally move probability mass in directions expected from the RANO criteria.

The main exception is `time_gap`, which ranks thirteenth by CaCE-TV (0.069) but remains noteworthy as the highest-ranked non-imaging metadata concept. This suggests that scan timing carries dataset-level signal in LUMIERE. We keep `time_gap` as a passthrough concept because scan interval is clinically available metadata. Notably, three absolute-change concepts (`delta_enhancing_absolute`, `delta_non_enhancing_absolute`, `delta_spd_absolute`) yield a CaCE-TV of exactly zero, indicating that the bottleneck has learned to rely on percentage-change and SPD-derived representations rather than raw absolute differences.

Concept alignment. We also evaluate concept alignment between predicted and target concepts. The full Concept-Alignment Score table is reported in Appendix A.15. In summary, most RANO threshold flags are recovered with high binary accuracy, while `spd_pd_flag` is less reliable. Follow-up tumor measurements are also predicted more accurately than baseline measurements, suggesting that the baseline branch of the siamese encoder is a useful target for future improvement.

5.4. Intervention Behavior

We probe the bottleneck causally by replacing predicted concept values with ground-truth values and measuring the change in 4-class macro-F1. Figure 6 reports the Ordered Intervention Score (OIS), defined as the average macro-F1 over $k = 1, \dots, 8$ oracle concept replacements, computed per fold and aggregated across the five LUMIERE folds. The *ordered* policy ranks each concept by its measured single-concept contribution in that fold; the *random* policy reveals random subsets of size k, averaged over 12 random subsets per k.

Across folds, $\mathrm{OIS}_{\mathrm{ord}} = 0.576 \pm 0.085$ versus $\mathrm{OIS}_{\mathrm{rnd}} = 0.392 \pm 0.066$, gives a per-fold gap of $+0.184 \pm 0.048$ macro-F1. The ordered curve in Figure 6 (a) rises sharply over the first $\sim$ five concepts and then saturates, indicating that a small set of high-impact concepts carries most of the post-intervention performance; this set overlaps with the RANO drivers identified by CaCE in Section 5.3. The random curve, in contrast, drops *below* the no-intervention baseline once more than two concepts are corrected, because partial corrections of correlated concepts propagated through the deterministic RANO chain can introduce inconsistencies between root measurements, percentage deltas, and threshold flags.

The per-fold view in Figure 6 (b) shows that the ordering benefit is consistent across folds, where folds with a low no-intervention baseline (Fold 1: 0.359, Fold 5: 0.388) gain the most from oracle correction (+14 and +26 percentage points respectively), while Fold 4 (i.e., already at macro-F1 = 0.696 at $k = 0$) saturates within two interventions and provides a fold-level ceiling that bounds the achievable improvement. Together, these two observations support the role of intervention-consistency training, in which the bottleneck responds to ground-truth concept correction and distinguishes between informative and uninformative concepts, reflecting the kind of selective concept refinement expected in real-world clinician annotation and review.

5.5. Patient-Level Walkthrough

Figure 7 shows a representative misclassified case (Patient-004, validation fold 1). In the default forward pass, TRACE predicts CR for a true PD case, with visible errors in the volumetric root concepts, including an under-estimate of `enhancing_tumor_volume_cm3`. We then correct two root concepts, `enhancing_tumor_volume_cm3` and `followup_non_enhancing_volume_cm3`, both highlighted in the figure. The deterministic RANO quantities are recomputed from these corrected values, including the percentage change concepts and threshold flags, and the prediction changes from CR to PD. This example illustrates the intended use of our proposed structured bottleneck, where concept-level corrections can

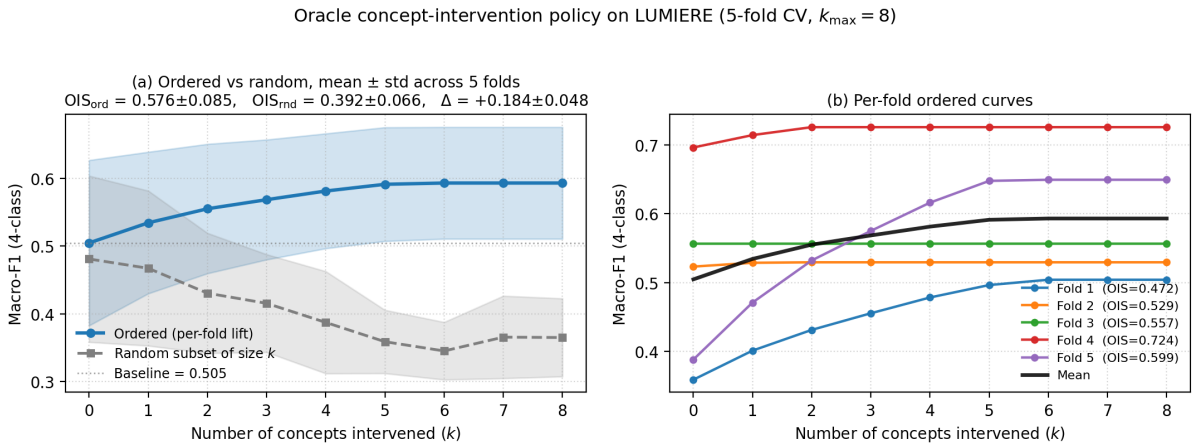


Figure 6: Oracle concept-intervention policy on LUMIERE (5-fold CV, k_{max}=8). **(a)** Mean 4-class macro-F1 across folds when the top-k concepts are replaced with ground-truth values, ranked by per-fold single-concept lift (blue, ordered) or chosen at random (grey, mean over 12 random subsets). Shaded bands are ±1 std across folds. The ordered policy improves macro-F1 monotonically and saturates around k=5, while the random policy degrades the prediction once k exceeds two. **(b)** Per-fold ordered curves. Folds with a low baseline (F1, F5) gain the most from oracle correction, while fold 4 starts at 0.696 and saturates within two interventions. $OIS_{ord} = 0.576 \pm 0.085$ vs $OIS_{rnd} = 0.392 \pm 0.066$, gap $+0.184 \pm 0.048$.

be propagated through the RANO-aligned graph and can change the downstream response prediction without retraining.

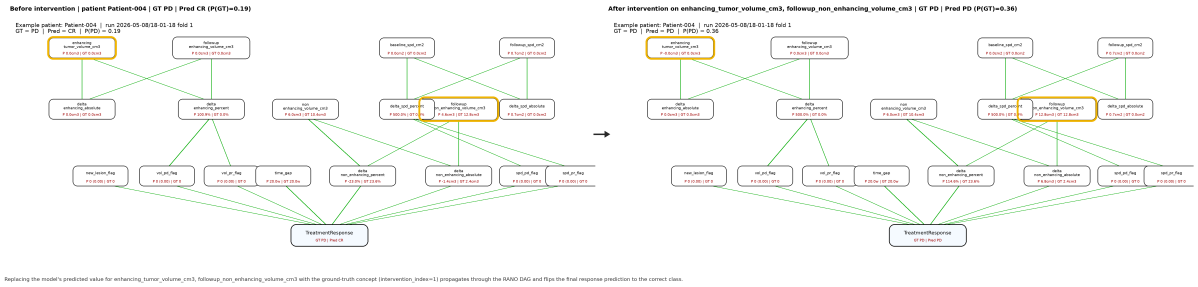


Figure 7: Concept-intervention walkthrough on a misclassified case (Patient-004, validation fold 1). **Left:** default forward pass; the model predicts CR while the ground-truth label is PD. **Right:** same input with intervention_index = 1 on `enhancing_tumor_volume_cm3` and `followup_non_enhancing_volume_cm3` (yellow outlines). The corrected concept values propagate deterministically along the RANO chain (Δ% →flags→response) and the prediction flips to the correct PD class. No retraining or gradient step is performed.

6. Conclusion, Limitations, and Future Work

Conclusion. This work introduced TRACE, an exploratory RANO-aligned concept bottleneck model for longitudinal GBM response assessment. This work utilized RANO-based response assessment as a structured concept reasoning problem rather than a direct image-to-label prediction task. TRACE estimates root tumor measurements from paired MRI, computes downstream RANO quantities deterministically, and allows concept-level corrections to propagate through the response pathway. On LUMIERE, TRACE improved over the concept-based baseline on 4-class RANO classification and performed strongly on the PD-vs-rest decision, while preserving an inspectable bottleneck of clinical concepts. The ablation and intervention results further suggest that this behavior depends on both the expert RANO graph and intervention-consistency training.

Limitations. These results should be interpreted with several limitations in mind. First, CR and PR remain difficult because they are represented by a few cases across several validation folds, increasing fold-to-fold variation in per-class F1. Second, `time_gap` has a strong effect on the output, suggesting that scan interval captures dataset-level signal in LUMIERE; this concept should therefore be examined carefully in external validation and may require sensitivity analysis or CaCE-based filtering. Third, concept alignment remains weaker for some baseline measurements and for the SPD-derived PD flag, suggesting that root concept estimation remains an issue. Finally, all experiments are based on a single dataset, i.e., LUMIERE, so broader validation on more protocol-aligned longitudinal RANO datasets is needed before drawing stronger conclusions about generalization.

Future work. Future work should prioritize validation on more datasets, stronger root concept estimation, and better handling of rare response classes. In particular, improving the estimation of baseline tumor measurements and SPD-derived progression flags may strengthen the reliability of the structured bottleneck. More broadly, the same design may extend to other longitudinal medical imaging tasks where decisions depend on explicit clinical criteria, temporal change, and human-verifiable intermediate measurements.

7. Acknowledgments

This work was partially funded by the Federal Ministry of Research, Technology and Space (BMFTR) under grant number 16IW23002 (No-IDLE), grant number 16IW24006 (NoIDLEChatGPT), and grant number 16IW26002 (Amenable); the Lower Saxony Ministry of Science and Culture (MWK) in the zukunft.niedersachsen program; and the Endowed Chair of AAI at the University of Oldenburg.

Declaration on Generative AI Usage

During the preparation of this manuscript, the authors used Claude, ChatGPT, and Gemini to assist with language-related tasks, such as grammar and spelling correction, as well as to improve clarity and writing style. After using these tools, the authors carefully reviewed and edited the content as needed and take full responsibility for the final content of the publication.

References

[1] J. P. Thakkar, T. A. Dolecek, C. Horbinski, Q. T. Ostrom, D. D. Lightner, J. S. Barnholtz-Sloan, J. L. Villano, Epidemiologic and molecular prognostic review of glioblastoma, Cancer Epidemiology, Biomarkers & Prevention 23 (2014) 1985–1996. doi:10.1158/1055-9965.EPI-14-0275.

[2] P. Y. Wen, M. van den Bent, G. Youssef, T. F. Cloughesy, B. M. Ellingson, M. Weller, E. Galanis, D. P. Barboriak, J. de Groot, M. R. Gilbert, R. Huang, A. B. Lassman, M. Mehta, A. M. Molinaro, M. Preusser, R. Rahman, L. K. Shankar, R. Stupp, J. E. Villanueva-Meyer, W. Wick, D. R. Macdonald, D. A. Reardon, M. A. Vogelbaum, S. M. Chang, RANO 2.0: Update to the Response Assessment in Neuro-Oncology Criteria for High- and Low-Grade Gliomas in Adults, Journal of Clinical Oncology 41 (2023) 5187–5199. doi:10.1200/JCO.23.01059.

[3] S. Zhang, Y. Sun, Y. Ao, X. Zhang, R. Yang, J. Xu, Z. Ai, H. Zhang, X. Yang, Y. Xu, K. Li, D. Chen, Glomia-pro: A generalizable longitudinal medical image analysis framework for disease progression prediction, 2025. arXiv:2507.12500.

[4] M. Moassefi, S. Faghani, G. M. Conte, R. O. Kowalchuk, S. Vahdati, D. J. Crompton, C. Perez-Vega, R. A. D. Cabreja, S. A. Vora, A. Quiñones-Hinojosa, I. F. Parney, D. M. Trifiletti, B. J. Erickson, A deep learning model for discriminating true progression from pseudoprogression in glioblastoma patients, Journal of Neuro-Oncology 159 (2022) 447–455. doi:10.1007/s11060-022-04080-x.

[5] S. Khalighi, K. Reddy, A. Midya, K. B. Pandav, A. Madabhushi, M. Abedalthagafi, Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment, npj Precision Oncology 8 (2024) 80. doi:10.1038/s41698-024-00575-0.

[6] A. Rončević, N. Koruga, A. S. Koruga, R. Rončević, Artificial intelligence in glioblastoma — transforming diagnosis and treatment, Chinese Neurosurgical Journal 11 (2025) 6. doi:10.1186/s41016-025-00399-2.

[7] D. J. Ghadimi, A. M. Vahdani, H. Karimi, P. Ebrahimi, M. Fathi, F. Moodi, A. Habibzadeh, F. Khodadadi Shoushtari, G. Valizadeh, H. Mobarak Salari, H. Saligheh Rad, Deep Learning-Based Techniques in Glioma Brain Tumor Segmentation Using Multi-Parametric MRI: A Review on Clinical Applications and Future Outlooks, Journal of Magnetic Resonance Imaging 61 (2024) 1094–1109. doi:10.1002/jmri.29543.

[8] M. Hagenbuchner, The black box problem of ai in oncology, Journal of Physics: Conference Series 1662 (2020) 012012. doi:10.1088/1742-6596/1662/1/012012.

[9] M. A. Gulum, C. M. Trombley, M. Kantardzic, A review of explainable deep learning cancer detection models in medical imaging, Applied Sciences 11 (2021) 4573. doi:10.3390/app11104573.

[10] H. Charaabi, H. Mzoughi, R. E. Hamdi, M. Njah, EXplainable Artificial Intelligence (XAI) for MRI Brain Tumor Diagnosis: A Survey, in: Proceedings of the International Conference on Cyberworlds, 2023. doi:10.1109/CW58918.2023.00033.

[11] K. Desai, P. K. Patel, A. Barve, Enhancing trust in ai-driven diagnostics: A review of brain tumor classification using cnns with a hybrid grad-cam and counterfactual xai framework, in: 2025 4th International Conference on Applied Artificial Intelligence and Computing (ICAAIC), IEEE, Salem, India, 2025, pp. 1592–1598. doi:10.1109/ICAAIC64647.2025.11330252.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.

[13] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, Concept bottleneck models, in: Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 5338–5348. URL: https://proceedings.mlr.press/v119/koh20a.html.

[14] H. M. T. Alam, D. Srivastav, M. A. Kadir, D. Sonntag, Towards interpretable radiology report generation via concept bottlenecks using a multi-agentic RAG, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part III, volume 15574 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 201–209. URL: https://doi.org/10.1007/978-3-031-88714-7_18. doi:10.1007/978-3-031-88714-7_18.

[15] S. Shin, Y. Jo, S. Ahn, N. Lee, A closer look at the intervention procedure of concept bottleneck models, in: Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022, 2022. URL: https://openreview.net/forum?id=PUspzfGsgY.

[16] H. M. T. Alam, D. Srivastav, A. Mohamed Selim, M. A. Kadir, M. M. H. Shuvo, D. Sonntag, Cbm-rag: Demonstrating enhanced interpretability in radiology report generation with multi-agent rag and concept bottleneck models, in: Companion Proceedings of the 17th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, EICS '25 Companion, Association for Computing Machinery, New York, NY, USA, 2025, p. 59–61. URL: https://doi.org/10.1145/3731406.3731970. doi:10.1145/3731406.3731970.

[17] G. D. Felice, A. C. Flores, F. D. Santis, S. Santini, J. Schneider, P. Barbiero, A. Termine, Causally reliable concept bottleneck models, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2026. URL: https://openreview.net/forum?id=UX143QGvb8.

[18] Y. Suter, U. Knecht, W. Valenzuela, M. Notter, E. Hewer, P. Schucht, R. Wiest, M. Reyes, The lumiere dataset: Longitudinal glioblastoma mri with expert rano evaluation, Scientific Data 9 (2022) 768. doi:10.1038/s41597-022-01881-7.

[19] N. Abu Khalaf, A. Desjardins, J. J. Vredenburgh, D. P. Barboriak, Repeatability of automated image

segmentation with BraTumIA in patients with recurrent glioblastoma, AJNR. American Journal of Neuroradiology 42 (2021) 1080–1086. doi:10.3174/ajnr.A7071.

[20] P. Kickingereder, F. Isensee, I. Tursunova, J. Petersen, U. Neuberger, D. Bonekamp, G. Brugnara, M. Schell, T. Kessler, M. Foltyn, et al., Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study, The Lancet Oncology 20 (2019) 728–740. doi:10.1016/S1470-2045(19)30098-1.

[21] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2020) 203–211. doi:10.1038/s41592-020-01008-z.

[22] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, M. Bendszus, K. H. Maier-Hein, P. Kickingereder, Automated brain extraction of multisequence MRI using artificial neural networks, Human Brain Mapping 40 (2019) 4952–4964. doi:10.1002/hbm.24750.

[23] Y. Suter, M. Notter, R. Meier, T. Loosli, P. Schucht, R. Wiest, M. Reyes, U. Knecht, Evaluating automated longitudinal tumor measurements for glioblastoma response assessment, Frontiers in Radiology 3 (2023). doi:10.3389/fradi.2023.1211859.

[24] A. Matoso, C. Passarinho, M. P. Loureiro, J. M. Moreira, P. Figueiredo, R. G. Nunes, Towards a deep learning approach for classifying treatment response in glioblastomas, 2025. arXiv:2504.18268.

[25] D. Amato, S. Calderaro, L. D. Reitano, G. Lo Bosco, R. Rizzo, F. Vella, Integrating deep learning and radiomic features for glioblastoma treatment response classification, in: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2025. doi:10.1109/BIBM66473.2025.11356532.

[26] D. Tikhonov, M. Scatolin, M. Banerjee, Q. Ji, A. Jaheen, M. Salem, A. Elsayed, H. Wang, S. Hashmi, M. Yaqub, Predicting Brain Tumor Response to Therapy using a Hybrid Deep Learning and Radiomics Approach, 2025. doi:10.48550/arXiv.2509.06511.

[27] J. V. Jeyakumar, A. Sarker, L. A. Garcia, M. Srivastava, X-CHAR: A Concept-based Explainable Complex Human Activity Recognition Model, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 7 (2023) 17:1–17:28. doi:10.1145/3580804.

[28] P. Knab, S. Marton, P. J. Schubert, D. Guggiana, C. Bartelt, Concepts in Motion: Temporal Concept Bottleneck Model for Interpretable Video Classification, 2026. doi:10.48550/arXiv.2509.20899, arXiv:2509.20899 [cs.CV] version: 3.

[29] F. Bai, Y. Du, T. Huang, M. Q.-H. Meng, B. Zhao, M3d: Advancing 3d medical image analysis with multi-modal large language models, 2024. doi:10.48550/arXiv.2404.00578. arXiv:2404.00578.

[30] A. Yan, Y. Wang, Y. Zhong, Z. He, P. Karypis, Z. Wang, C. Dong, A. Gentili, C.-N. Hsu, J. Shang, J. McAuley, Robust and interpretable medical image classifiers via concept bottleneck models, 2023. arXiv:2310.03182.

[31] A. Bunnell, Y. Glaser, D. Valdez, T. Wolfgruber, A. Altamirano, C. Zamora González, B. Y. Hernandez, P. Sadowski, J. A. Shepherd, Learning a Clinically-Relevant Concept Bottleneck for Lesion Detection in Breast Ultrasound, in: M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, J. A. Schnabel (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, Springer Nature Switzerland, Cham, 2024, pp. 650–659. doi:10.1007/978-3-031-72384-1_61.

[32] S. J. Magny, R. Shikhman, A. L. Keppke, Breast Imaging Reporting and Data System, 2023. URL: http://www.ncbi.nlm.nih.gov/books/NBK459169/.

[33] J. Kim, Z. Wang, Q. Qiu, Constructing Concept-Based Models to Mitigate Spurious Correlations with Minimal Human Effort, in: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXX, Springer-Verlag, Berlin, Heidelberg, 2024, pp. 137–153. doi:10.1007/978-3-031-72989-8_8.

[34] H. Fokkema, T. van Erven, S. Magliacane, Sample-efficient learning of concepts with theoretical guarantees: from data to concepts without interventions, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025. URL: https://openreview.net/forum?id=RCXF0UEmuE.

[35] Y. Wu, Y. Liu, Y. Yang, M. S. Yao, W. Yang, X. Shi, L. Yang, D. Li, Y. Liu, S. Yin, C. Lei, M. Zhang, J. C. Gee, X. Yang, W. Wei, S. Gu, A concept-based interpretable model for the diagnosis of choroid neoplasias using multimodal data, Nature Communications 16 (2025) 3504. doi:10.1038/s41467-025-58801-7.

[36] T. Oikarinen, S. Das, L. M. Nguyen, T.-W. Weng, Label-free concept bottleneck models, in: The Eleventh International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=FlCg47MNvBA.

[37] K. Prasse, P. Knab, S. Marton, C. Bartelt, M. Keuper, Dcbm: Data-efficient visual concept bottleneck models, in: Forty-second International Conference on Machine Learning, 2025. URL: https://openreview.net/forum?id=BdO4R6XxUH.

[38] I. Shrier, R. W. Platt, Reducing bias through directed acyclic graphs, BMC Medical Research Methodology 8 (2008) 70. doi:10.1186/1471-2288-8-70.

[39] D. Evans, B. Chaix, T. Lobbedez, C. Verger, A. Flahault, Combining directed acyclic graphs and the change-in-estimate procedure as a novel approach to adjustment-variable selection in epidemiology, BMC Medical Research Methodology 12 (2012) 156. doi:10.1186/1471-2288-12-156.

[40] M. Piccininni, S. Konigorski, J. L. Rohmann, T. Kurth, Directed acyclic graphs and causal thinking in clinical risk prediction modeling, BMC Medical Research Methodology 20 (2020) 179. doi:10.1186/s12874-020-01058-z.

[41] P. Barbiero, M. E. Zarlenga, F. Giannini, A. Termine, F. Bonchi, M. Jamnik, G. Marra, Actionable Interpretability Must Be Defined in Terms of Symmetries, 2026. doi:10.48550/arXiv.2601.12913, arXiv:2601.12913 [cs.AI].

[42] A. Sun, Y. Yuan, P. Ma, S. Wang, Eliminating information leakage in hard concept bottleneck models with supervised, hierarchical concept learning, 2024. arXiv:2402.05945.

[43] S. Chen, K. Ma, Y. Zheng, Med3D: Transfer Learning for 3D Medical Image Analysis, 2019. doi:10.48550/arXiv.1904.00625, arXiv:2601.12913 [cs.AI].

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (2020) 318–327. doi:10.1109/TPAMI.2018.2858826.

[45] M. Havasi, S. Parbhoo, F. Doshi-Velez, Addressing leakage in concept bottleneck models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022, pp. 23386–23397. URL: https://openreview.net/forum?id=tglniD_fn9.

[46] L. Gagnon, D. Gupta, U. Nguyen, M. Correia de Verdier, R. Saluja, G. Mastorakos, N. White, V. Goodwill, C. R. McDonald, T. Beaumont, C. Conlin, T. M. Seibert, J. Hattangadi-Gluth, S. Kesari, J. D. Schulte, D. Piccioni, K. M. Schmainda, N. Farid, A. M. Dale, J. D. Rudie, The University of California San Diego Post-Treatment Glioblastoma (UCSD-PTGBM) annotated multimodal MRI Dataset, Scientific Data 13 (2026) 183. doi:10.1038/s41597-025-06499-z.

[47] B. K. K. Fields, E. Calabrese, J. Mongan, S. Cha, C. P. Hess, L. P. Sugrue, S. M. Chang, T. L. Luks, J. E. Villanueva-Meyer, A. M. Rauschecker, J. D. Rudie, The university of california san francisco adult longitudinal post-treatment diffuse glioma mri dataset, Radiology: Artificial Intelligence 6 (2024) e230182. doi:10.1148/ryai.230182.

[48] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, J. C. Gee, N4itk: improved n3 bias correction, IEEE Transactions on Medical Imaging 29 (2010) 1310–1320. doi:10.1109/TMI.2010.2046908.

[49] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[50] Y. Goyal, A. Feder, U. Shalit, B. Kim, Explaining Classifiers with Causal Concept Effect (CaCE), 2020. doi:10.48550/arXiv.1907.07165, arXiv:1907.07165 [cs.LG].

A. Additional Implementation Details

A.1. Clinical Background on SPD

The Sum of Products of Diameters (SPD) is defined as the product of the two largest perpendicular tumor diameters measured on a single imaging slice. It is retained in RANO-based response assessment because the response thresholds are historically defined using this 2D measurement. We therefore include SPD alongside volume so that the model can represent both volumetric change and the SPD-based thresholds used in the RANO criteria.

A.2. ConceptBlock Details

The main model uses ConceptBlocks only for continuous image-derived root concepts. Each root concept c_k is predicted from the shared image representation $\mathbf{X}_{\text{enc}}$ using a concept-specific MLP:

$$\mathbf{e}_k = \text{MLP}_k(\mathbf{X}_{\text{enc}}) \in \mathbb{R}^{d_c}, \tag{29}$$

where d_c is the concept hidden size. The scalar concept prediction is then computed as

$$\hat{c}_k = \mathbf{w}_k^\top \mathbf{e}_k + b_k. \tag{30}$$

The prediction $\hat{c}_k$ is trained against the z-scored ground-truth concept value using mean-squared error. The embedding $\mathbf{e}_k$ is retained as the root concept representation passed to downstream prediction.

This continuous ConceptBlock setting is used for all image-derived root concepts in this work, namely V_{t_0}, V_{t_1}, SPD_{t_0}, and SPD_{t_1}. Derived and passthrough nodes do not use ConceptBlocks.

A.3. Deterministic Node Details

Deterministic nodes implement RANO-derived quantities that are fixed functions of their parent concepts. These nodes do not contain trainable parameters and are excluded from the concept supervision loss.

Because root concepts are predicted in z-scored space, deterministic computations are applied after converting predictions back to the raw clinical scale. Given a normalized prediction $\hat{c}_k$ for concept c_k, the raw-scale prediction is

$$\tilde{c}_k = \text{expm1}(\hat{c}_k \sigma_k + \mu_k), \tag{31}$$

where μ_k and σ_k are the training-fold mean and standard deviation of concept c_k. The expm1 operation is used when the corresponding concept labels were log-transformed during preprocessing.

The volume percentage change is computed as

$$\widehat{\Delta V}_{\%} = 100 \cdot \text{clamp}\left(\frac{\tilde{V}_{t_1} - \tilde{V}_{t_0}}{\tilde{V}_{t_0} + \varepsilon}, -1, 5 \right). \tag{32}$$

The small constant ε avoids division by zero. The clamp operation limits extreme ratios caused by very small baseline values and keeps the derived quantity within a stable numerical range.

The volume-based PD and PR threshold flags are then computed as

$$\hat{b}_{\text{PD}}^{\text{vol}} = \mathbf{1}\left[\widehat{\Delta V}_{\%} \geq 40 \ \wedge \ \tilde{V}_{t_0} \geq 0.5 \right], \tag{33}$$

$$\hat{b}_{\text{PR}}^{\text{vol}} = \mathbf{1}\left[\widehat{\Delta V}_{\%} \leq -65 \ \wedge \ \tilde{V}_{t_0} \geq 0.5 \right]. \tag{34}$$

The SPD percentage change is computed analogously:

$$\widehat{\Delta \text{SPD}}_{\%} = 100 \cdot \text{clamp}\left(\frac{\widetilde{\text{SPD}}_{t_1} - \widetilde{\text{SPD}}_{t_0}}{\widetilde{\text{SPD}}_{t_0} + \varepsilon}, -1, 5 \right). \tag{35}$$

The SPD-based PD and PR threshold flags are

$$\hat{b}_{\mathrm{PD}}^{\mathrm{SPD}} = \mathbf{1}\left[\widehat{\Delta\mathrm{SPD}}_{\%} \geq 25 \ \wedge \ \widetilde{\mathrm{SPD}}_{t_0} \geq 0.01\right], \tag{36}$$

$$\hat{b}_{\mathrm{PR}}^{\mathrm{SPD}} = \mathbf{1}\left[\widehat{\Delta\mathrm{SPD}}_{\%} \leq -50 \ \wedge \ \widetilde{\mathrm{SPD}}_{t_0} \geq 0.01\right]. \tag{37}$$

The measurability floors follow the thresholds in Table 1. If the baseline measurement is below the corresponding floor, the flag is not activated for that measurement type. This avoids unstable percentage changes from very small baseline values.

Hard threshold indicators near the RANO boundaries could be relaxed using a temperature-controlled sigmoid, $\sigma((u-\tau)/T)$, for smoother gradients. In practice, the hard-threshold formulation was stable, so we use it as the default.

A.4. Passthrough Node Details

The passthrough nodes represent concepts that are clinically relevant to the final RANO response but are not predicted from the image encoder. In this work, the passthrough nodes are the scan interval Δt and the new-lesion flag b_{NL}.

The scan interval Δt is the time difference between the baseline scan t_0 and the follow-up scan t_1. It is obtained from scan metadata and injected directly into the graph:

$$\hat{c}_{\Delta t} = c^*_{\Delta t}. \tag{38}$$

This avoids treating scan timing as an image-derived concept, while still allowing the task head to use temporal context when predicting the RANO response.

The new-lesion flag b_{NL} indicates whether a new lesion appears at follow-up. It is computed during preprocessing from the baseline and follow-up segmentation masks. In our implementation, connected components are identified in the follow-up mask and compared with the baseline mask. A component present at t_1 with no corresponding component at t_0 is treated as evidence for a new lesion:

$$b_{\mathrm{NL}} = \mathbf{1}\left[\exists\, L_{t_1}^{(m)} \text{ such that } L_{t_1}^{(m)} \cap L_{t_0} = \emptyset\right], \tag{39}$$

where $L_{t_1}^{(m)}$ denotes the m-th connected component at follow-up and L_{t_0} denotes the union of baseline lesion components.

The value is then injected directly into the DAG:

$$\hat{c}_{b_{\mathrm{NL}}} = c^*_{b_{\mathrm{NL}}}. \tag{40}$$

Under RANO 2.0, the appearance of a new lesion supports Progressive Disease. We therefore include b_{NL} as an explicit graph node rather than asking the image encoder to learn this decision implicitly.

Passthrough nodes do not have ConceptBlocks, do not have trainable parameters, and do not contribute to the concept supervision loss. They enter the graph only as scalar values supplied to downstream nodes. This keeps the distinction clear between image-derived root concepts, deterministic RANO-derived concepts, and metadata supplied outside the image encoder.

A.5. Optional Monotonicity Constraints

RANO reasoning implies monotonic relationships between some concepts and response classes: larger positive changes in tumor volume or SPD should not decrease the evidence for PD, and larger negative changes should not decrease the evidence for PR. These relationships can be enforced as monotonicity constraints on the task head (e.g., $\partial \mathrm{logit}_{\mathrm{PD}}/\partial \Delta V_{\%} \geq 0$ and $\partial \mathrm{logit}_{\mathrm{PR}}/\partial \Delta V_{\%} \leq 0$, with the analogous constraints on $\Delta\mathrm{SPD}_{\%}$), implemented via constrained linear layers or monotonic networks. We treat this as an optional extension rather than a required component of the main model because hard monotonic constraints can be too restrictive on noisy clinical measurements, and the intervention-consistency objective already enforces the dominant directional behavior empirically.

A.6. External Dataset Considerations

We examined whether external longitudinal glioma datasets could be used to supplement LUMIERE or provide additional validation data. In particular, we considered UCSD-PTGBM [46] and UCSF-ALPTDG [47]. These datasets are useful resources for longitudinal post-treatment glioma MRI analysis, but they do not provide scan-pair-level expert RANO response labels aligned with the four categories used in this work, namely CR, PR, SD, and PD.

This matters because TRACE is designed around RANO-based response assessment. The model predicts the final RANO label and structures its intermediate concepts around RANO measurements, threshold flags, and passthrough clinical metadata. Labels based on other clinical endpoints, dataset-specific categories, survival information, or post hoc thresholds are therefore not interchangeable with expert RANO response labels.

We tested whether approximate RANO labels could be reconstructed for the external datasets using segmentation-derived longitudinal volume changes. This was not suitable because RANO assessment is not determined by volume change alone. It also depends on measurable disease requirements, SPD-based thresholds, the appearance of new lesions, and assessment context. Reconstructed labels were especially ambiguous near decision boundaries between PR, SD, and PD.

For these reasons, we did not merge the external datasets with LUMIERE for model training. Doing so would have increased the number of scans, but it would also have mixed different label definitions. We therefore restrict the main experiments to LUMIERE, where the supervision target matches the RANO-based task. External validation remains an important future direction, but it requires datasets with protocol-aligned scan-pair-level RANO annotations.

A.7. Preprocessing Details

All MRI sequences were spatially normalized before model training. The four modalities were registered to the MNI152 1 mm template space. N4 bias field correction [48] was applied to reduce low-frequency intensity inhomogeneity. Histogram matching was then used to reduce intensity variation across scans, followed by z-score normalization. Segmentation masks were transformed using the same image-to-template transformation estimated from the corresponding MRI. Nearest-neighbor interpolation was used to preserve discrete label values. This ensured that the MRI volumes and tumor segmentation masks remained spatially aligned after registration. The final model input was resampled to 128×128×128 voxels. Figure 8 shows a summary of the 4-step preprocessing pipeline.

A.8. Data-Efficient CBM Baseline Details

We adapted the Data-Efficient CBM baseline [37] to the longitudinal 3D MRI setting. For each scan, tissue-density-filtered 3D patches of size 32×64×64 voxels were extracted across the four MRI modalities. Tissue-density filtering was used to reduce background-dominated patches and focus the baseline on image regions more likely to contain anatomical or tumor-related information.

Each patch was encoded using a frozen M3D-LaMed-Phi-3 encoder [29], producing a 3072-dimensional embedding. These embeddings were passed to a sparse autoencoder with 128 concepts and sparsity weight $\lambda = 0.001$. The sparse autoencoder was trained separately within each cross-validation fold to avoid leakage between training and validation splits.

For each baseline–follow-up pair, longitudinal change was represented by subtracting the baseline concept activations from the follow-up concept activations. The resulting difference representation was used as input to an L_2-regularized logistic regression classifier for 4-class RANO response prediction. This setup provides a concept-based comparison in which concepts are discovered from pretrained embeddings rather than defined from RANO measurements and clinical rules.

A.9. Training Hyperparameters

Table 6 summarizes the main training hyperparameters used for TRACE variants.

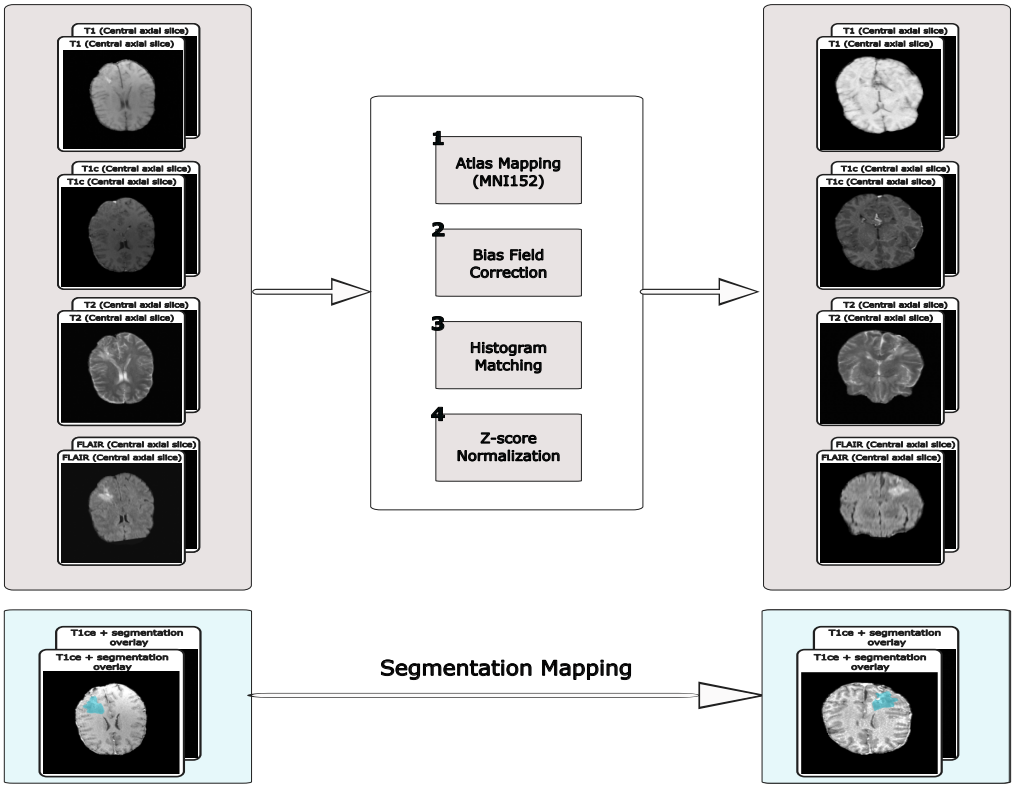


Figure 8: Preprocessing pipeline applied to the raw MRI data. The four MRI modalities undergo atlas registration to MNI152 space, N4 bias field correction, histogram matching, and z-score normalization. The HD-GLIO-AUTO segmentation mask is mapped to the same template space using the transformation estimated from the corresponding MRI, with nearest-neighbor interpolation used to preserve discrete label values.

Table 6
Training hyperparameters for TRACE variants.

Hyperparameter	**Value**
Optimizer	AdamW
Learning rate	10^{-3}
L_2 weight decay	0.01
Batch size	16
Input dimensions	128×128×128 voxels
Max epochs	100
LR scheduler	ReduceLROnPlateau factor 0.5, patience 10
Early stopping patience	20 epochs, validation macro F1
Gradient clipping	$\|\nabla\|_2 \leq 1.0$
Dropout	$p = 0.3$
Task loss	Soft macro F1 with focal weighting
Auxiliary loss weight α_{aux}	0.1
Intervention consistency weight β_{cons}	0.5
Concept loss weight λ_{c}	0.8
Intervention probability p_{int}	annealed from 0.3 to 0.8
Annealing period	first 15 epochs
L_1 penalty	10^{-4} on the first layer of the task head
Encoder freezing	epoch 50

A.10. Ordered vs Random Intervention Policy Ablation

To test whether concept ranking is meaningful for interventions, we compared an ordered policy (top-k concepts ranked by per-concept intervention lift) against a random-k policy. Table 7 reports the 4-class macro F1 results from the run in `paper_figures/intervention_policy_ablation_k6_r6.csv`. The ordered policy consistently outperforms random interventions for $k \geq 2$, and the gap increases with larger k, indicating that the ranking captures actionable concept importance.

Table 7
Ordered vs random intervention policy ablation (4-class macro F1, mean ± std).

k	Ordered policy	Random policy	Gap (Ordered - Random)
0	0.481 ± 0.123	0.481 ± 0.123	0.000 ± 0.000
1	0.481 ± 0.123	0.482 ± 0.119	−0.001 ± 0.010
2	0.481 ± 0.123	0.437 ± 0.096	0.044 ± 0.027
3	0.481 ± 0.123	0.411 ± 0.071	0.070 ± 0.065
4	0.481 ± 0.123	0.411 ± 0.067	0.070 ± 0.077
5	0.481 ± 0.123	0.399 ± 0.070	0.082 ± 0.073
6	0.481 ± 0.123	0.392 ± 0.074	0.089 ± 0.088

A.11. Fold-Level Metrics

Table 8 reports the per-fold metrics for TRACE. These values support the aggregate results in Section 5. Variation across folds is mainly driven by the minority CR and PR classes, while the PD-vs-rest binary task is more stable.

Table 8
Per-fold and aggregate metrics for TRACE.

Fold	4-class macro F1	4-class weighted acc	Binary macro F1	Binary weighted acc
F1	0.359	0.402	0.616	0.621
F2	0.523	0.464	0.656	0.662
F3	0.418	0.616	0.778	0.841
F4	0.696	0.768	0.857	0.891
F5	0.388	0.473	0.635	0.651
Mean ± Std	0.4769 ± 0.1229	0.5447 ± 0.1318	0.7085 ± 0.0935	0.7329 ± 0.1104

A.12. Per-Class Metrics

Table 9 reports the per-class F1 values across folds. The table complements Figure 4a in the main text.

Table 9
Per-class F1 across folds. The mean column aggregates all five folds; the displayed fold columns show folds where each class has non-trivial support.

Task	Class	Mean ± Std	F1	F2	F3	F4
4-class	CR	0.405 ± 0.295	0.000	0.400	0.222	0.875
4-class	PR	0.270 ± 0.235	0.316	0.571	0.000	0.462
4-class	SD	0.477 ± 0.104	0.489	0.359	0.606	0.571
4-class	PD	0.756 ± 0.095	0.632	0.763	0.842	0.877
Binary	non-PD	0.661 ± 0.103	0.600	0.549	0.714	0.837
Binary	PD	0.756 ± 0.095	0.632	0.763	0.842	0.877

A.13. Rule-Based RANO Oracle

The rule-based RANO oracle applies fixed RANO threshold rules directly to the ground-truth concept vector for each validation pair. A case is assigned to PD if a new lesion is present, if the enhancing-volume change satisfies the PD threshold, or if the SPD change satisfies the PD threshold. PR and CR are assigned using the corresponding RANO threshold rules, with SD used when no progression or response rule is activated.

This reference is included to test whether fixed thresholding of the available concept values is sufficient for the LUMIERE response labels. It reaches 0.346 ± 0.038 4-class macro F1 and 0.681 ± 0.058 binary macro F1. The result is below TRACE, indicating that the expert labels in LUMIERE are not fully recovered by segmentation-derived threshold rules alone. This is plausible because clinical response assessment can also reflect non-measurable disease, assessment context, and other clinical factors not represented in the segmentation-derived concepts.

A.14. Additional Concept Causal Effects Details

Figure 9 summarizes the overall magnitude of concept influence by ranking concepts according to their CaCE-TV scores, independent of the direction of their class-specific effects.

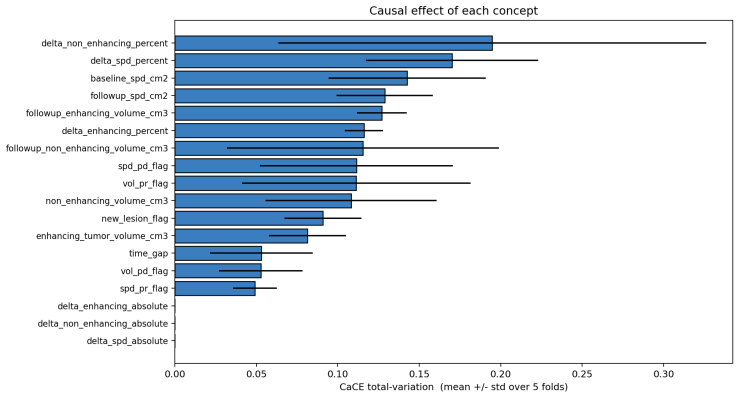


Figure 9: Top concepts by CaCE-TV. Concept influence is distributed across multiple RANO-related variables rather than concentrated in a single concept.

A.15. Concept-Alignment Details

Table 10 reports concept-alignment results across the five folds. Binary accuracy is used for flag concepts, and z-scored MAE is used for continuous concepts, where lower values indicate better alignment. Volume-based threshold flags are recovered with high binary accuracy (`vol_pd_flag`: 0.990 ± 0.015; `vol_pr_flag`: 0.968 ± 0.029), suggesting that the bottleneck reliably captures the volumetric progression and response signals. However, `spd_pd_flag` achieves only 0.399 ± 0.085 binary accuracy, which indicates that the SPD-based progression pathway is not reliably predicted, and we assume that this suggests the model relies more on volume-based concepts than the RANO graph implies.

Among continuous concepts, follow-up measurements are consistently predicted more accurately than their baseline counterparts, e.g., `followup_enhancing_volume_cm3` achieves a z-scored MAE of 0.295 ± 0.107 compared to 0.640 ± 0.505 for `enhancing_tumor_volume_cm3`, which may partly reflect the segmentation mask dependency discussed in the main limitations, since the encoder receives masks as input while supervision targets are derived from those same masks. SPD-based continuous concepts (`delta_spd_percent`: 0.840 ± 0.072; `baseline_spd_cm2`: 0.856 ± 0.308) show the weakest alignment overall, pointing to SPD estimation from volumetric 3D features as a key target for future improvement. We plan to investigate these assumptions further in future work using additional datasets.

Table 10
Concept-alignment summary across five folds. Binary accuracy is reported for flag concepts; z-scored MAE is reported for continuous concepts.

Concept	**Score ± Std**	**Metric**
new_lesion_flag	1.000 ± 0.000	binary accuracy
vol_pd_flag	0.990 ± 0.015	binary accuracy
vol_pr_flag	0.968 ± 0.029	binary accuracy
spd_pr_flag	0.864 ± 0.040	binary accuracy
spd_pd_flag	0.399 ± 0.085	binary accuracy
followup_enhancing_volume_cm3	0.295 ± 0.107	z-score MAE
followup_spd_cm2	0.434 ± 0.070	z-score MAE
followup_non_enhancing_volume_cm3	0.552 ± 0.179	z-score MAE
enhancing_tumor_volume_cm3	0.640 ± 0.505	z-score MAE
delta_non_enhancing_percent	0.688 ± 0.217	z-score MAE
delta_enhancing_percent	0.745 ± 0.060	z-score MAE
delta_spd_percent	0.840 ± 0.072	z-score MAE
baseline_spd_cm2	0.856 ± 0.308	z-score MAE
non_enhancing_volume_cm3	0.895 ± 0.112	z-score MAE