TIME-AWARE DIFFERENTIABLE SUBGROUP DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Subgroup Discovery (SD) seeks interpretable subsets exhibiting exceptional relationships with a target variable, but conventional SD largely treats observations as static and exceptionalness as a scalar. This is inadequate when subgroup differences emerge only during particular periods or follow distinct temporal trajectories. We introduce Time-Aware Differentiable Subgroup Discovery (TDSD), which formulates longitudinal discovery as a unified *who–when–how* problem: identify an interpretable subgroup, localize when its discrepancy is strongest, and characterize how that discrepancy evolves. TDSD uses differentiable soft predicates and feature-selection gates for subgroup generation and differentiable temporal boundaries for window discovery. It estimates a subgroup-specific treatment–control discrepancy trajectory and optimizes a window-aware temporal quality measure, while also supporting full-trajectory behavior discovery. Synthetic experiments evaluate subgroup, temporal-window, and temporal-behavior recovery, followed by a MIMIC-IV ICU treatment-response study. TDSD recovers informative temporal structure and substantially improves temporal localization over increasingly time-aware alternatives, while experiments also expose a remaining gap between soft optimization and exact hard-rule recovery.

1 INTRODUCTION

Subgroup discovery (SD) Atzmueller (2015); Herrera et al. (2011) identifies interpretable population segments exhibiting exceptional relationships or behaviors with respect to a target variable. Unlike global predictive modeling, it seeks local exceptionalness, making it useful for precision medicine, treatment-effect heterogeneity, anomaly analysis, and scientific discovery Du et al. (2020); Du (2020); Proença et al. (2022). In these settings, however, observations are often longitudinal: outcomes evolve over time, and subgroup differences may appear only during particular periods Daly et al. (2022); McArdle (2009); Hedeker & Gibbons (2006).

A scalar subgroup score can hide this structure. Two subgroups may have similar average treatment-control differences even though one exhibits an immediate, rapidly decaying response and the other a delayed response Duivesteijn et al. (2012; 2010). The resulting discovery problem is therefore not only *which* subgroup is exceptional, but also *when* and *how* its exceptionalness manifests. A useful longitudinal discovery system should consequently return an interpretable rule, a temporal localization, and the associated trajectory rather than a single scalar score Cao et al. (2023); Wu et al. (2018); Shi et al. (2022).

We introduce **Time-Aware Differentiable Subgroup Discovery (TDSD)**, which addresses these three questions jointly: (i) **who** is exceptional, (ii) **when** the discrepancy is concentrated, and (iii) **how** it evolves over time. Let X_i denote baseline attributes, $A_i \in \{0, 1\}$ a treatment/intervention indicator, and $Y_i(t)$ a longitudinal outcome. For a subgroup S, we define

$$m_a(t \mid S) = E[Y(t) \mid A = a, S(X) = 1], \qquad \Delta_S(t) = m_1(t \mid S) - m_0(t \mid S). \tag{1}$$

Under suitable causal identification assumptions, $\Delta_S(t)$ can be interpreted as a time-varying subgroup treatment effect; otherwise it is a treatment–control outcome discrepancy. TDSD retains this function rather than collapsing it to a scalar.

The central technical challenge is that both subgroup rules and temporal windows are discrete. We therefore replace hard predicates with differentiable soft predicates and feature gates, and replace a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

hard interval with a smooth temporal membership function Xu et al. (2024). The resulting subgroup and window can be optimized jointly with gradient descent. A second objective retains the complete discrepancy trajectory, allowing the framework to characterize transient, delayed, increasing, decreasing, and oscillatory behavior.

Our contributions are: (1) a time-aware formulation of subgroup discovery as a who–when–how problem; (2) a differentiable, interpretable subgroup parameterization; (3) a differentiable temporal-window quality measure enabling joint subgroup and temporal localization; (4) a complementary trajectory-based objective for temporal behavior; and (5) controlled synthetic and MIMIC-IV experiments that separately test these capabilities and the necessity of explicit temporal localization.

2 Related Work

Subgroup discovery and exceptional model mining. Classical subgroup discovery searches for concise descriptions associated with unusual target distributions Lemmerich et al. (2016); Kaytoue et al. (2017). Exceptional Model Mining extends this idea from target distributions to exceptional relationships or model parameters, allowing subgroups to be defined by differences in regression, correlation, classification, or probabilistic models Leman et al. (2008); Du et al. (2025); Duivesteijn et al. (2016); Duivesteijn & Thaele (2014). TDSD follows this local-exceptionalness perspective but makes time a first-class component of the discovered pattern.

Differentiable rule learning. Rule-based discovery is naturally combinatorial because feature selection, threshold selection, and logical conjunction are discrete Xu & Vreeken (2026); Paulus et al. (2026); Xu et al. (2026). Differentiable rule-learning methods relax predicates into continuous functions so that gradient optimization can learn interpretable conditions Wilms et al. (2026); Yang et al. (2026); Atzmueller et al. (2025). TDSD adopts this principle while additionally learning temporal localization and temporal discrepancy.

Longitudinal and treatment-effect analysis. Longitudinal and functional-data methods model trajectories, while heterogeneous treatment-effect methods estimate how treatment effects vary across covariates and, in longitudinal settings, over time Salmaso et al. (2026); Wang et al. (2026); Wei et al. (2020); Suk & Park (2026). These approaches generally assume groups, covariate partitions, or treatment-effect functions are specified or estimated rather than discovered as concise logical subgroups Lipkovich et al. (2023); Thomas & Bornkamp (2017); Lipkovich et al. (2024); Zhang et al. (2018). TDSD is complementary: it searches for interpretable regions of covariate space whose treatment–control trajectory is exceptional.

Position of TDSD. TDSD lies at the intersection of subgroup discovery, exceptional model mining, differentiable rule learning, and temporal/causal modeling. Its distinguishing object is the triple $(S, W, \Delta_S(t))$: an interpretable subgroup, an optional exceptional temporal window, and its temporal discrepancy trajectory.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{D} = \{(X_i, A_i, Y_i(t))\}_{i=1}^{N}$ denote a longitudinal dataset containing N individuals. For each individual i, $X_i \in \mathbb{R}^p$ denotes a vector of baseline or contextual attributes, $A_i \in \{0, 1\}$ denotes a binary treatment or intervention indicator, and $Y_i(t), \quad t \in \mathcal{T}$ denotes the observed outcome trajectory. We assume that an exceptional subgroup can be represented by an interpretable conjunction of conditions over X. For example, $S^*(X) = (X_1 > c_1) \wedge (X_3 < c_3) \wedge (X_6 > c_6)$. The subgroup S^* is unknown and must be discovered from the data. For controlled synthetic experiments, we use the following data-generating process: $Y_i(t) = \mu(t) + A_i S_i^* \tau(t) + \epsilon_i(t)$, where $\mu(t)$ represents the common temporal baseline, S_i^* indicates membership in the ground-truth exceptional subgroup, $\tau(t)$ represents the subgroup-specific temporal treatment effect, and $\epsilon_i(t)$ represents observation noise. This formulation deliberately allows the treatment effect to vary over time. Consequently, the exceptional behavior cannot necessarily be identified by a single average treatment effect.

3.2 SUBGROUP-SPECIFIC TEMPORAL DISCREPANCY

Given a candidate subgroup S, we define the treatment and control outcome trajectories within the subgroup as $m_1(t \mid S) = E[Y(t) \mid A = 1, S(X) = 1]$, and $m_0(t \mid S) = E[Y(t) \mid A = 0, S(X) = 1]$. The corresponding subgroup-specific temporal discrepancy is $\Delta_S(t) = m_1(t \mid S) - m_0(t \mid S)$ which is a function rather than a scalar. Under appropriate causal identification assumptions, such as consistency, treatment exchangeability, and positivity, $\Delta_S(t)$ can be interpreted as a subgroup-specific time-varying treatment effect. Without these assumptions, it should be interpreted more generally as a temporal treatment-control outcome discrepancy.

This distinction is important because TDSD is primarily a discovery framework, rather than a causal identification framework. The method can therefore be used both for treatment-effect discovery and for more general temporal outcome-discrepancy discovery. The key advantage of retaining the entire function $\Delta_S(t)$ is that two subgroups with similar average effects can nevertheless exhibit very different temporal behaviors. For example, $\Delta_{S_1}(t)$ may correspond to an immediate but rapidly decaying response, whereas $\Delta_{S_2}(t)$ may correspond to a delayed response that becomes stronger later. A static quality measure may regard these two subgroups as similar, even though their temporal interpretations are fundamentally different. Therefore, TDSD treats $\Delta_S(t), \quad t \in \mathcal{T}$ as the fundamental object of exceptionalness.

3.3 DIFFERENTIABLE SUBGROUP GENERATION

3.3.1 SOFT PREDICATE REPRESENTATION

Traditional subgroup discovery represents a condition using a hard logical predicate, such as $X_j > c_j$. Such a predicate contains discrete decisions concerning both feature selection and threshold selection, making direct optimization combinatorial. TDSD replaces this hard predicate with a differentiable soft predicate: $p_j(X_j) = \sigma\left(\frac{d_j(X_j - \theta_j)}{\lambda_x}\right)$,

where θ_j is a learnable threshold; d_j controls the direction of the predicate; $\lambda_x > 0$ is a temperature parameter controlling the smoothness; $\sigma(\cdot)$ is the sigmoid function. The direction parameter can be constrained through $d_j = \tanh(\tilde{d}_j)$, so that the model can learn whether the subgroup condition favors larger or smaller values of X_j.

For example, when $d_j > 0$, the predicate approximately represents $X_j > \theta_j$, whereas $d_j < 0$ represents approximately $X_j < \theta_j$. As λ_x decreases, the sigmoid transition becomes sharper and the soft predicate approaches a hard logical condition.

3.3.2 FEATURE-SELECTION GATES

Not every feature should participate in the subgroup description. We therefore introduce a learnable feature-selection gate $g_j = \sigma(a_j)$, where $a_j \in \mathbb{R}$ is a learnable parameter and $g_j \in (0, 1)$. A large g_j indicates that feature X_j contributes strongly to the subgroup description, whereas a small g_j effectively disables the corresponding predicate. This mechanism allows the model to simultaneously learn which features are relevant and where the subgroup boundaries should be located.

3.3.3 DIFFERENTIABLE LOGICAL CONJUNCTION

A subgroup is conventionally represented as a conjunction: $S(X) = \bigwedge_{j \in \mathcal{J}} p_j(X)$. Directly replacing the conjunction with multiplication, $\prod_j p_j(X)$, can result in vanishing gradients when many predicates are involved. We therefore construct a normalized logarithmic aggregation: $S_\theta(X) = \exp\left[\frac{\sum_{j=1}^{p} g_j \log(p_j(X)+\epsilon)}{\sum_{j=1}^{p} g_j + \epsilon}\right]$ where ϵ is a small numerical constant.

The resulting $S_\theta(X) \in (0, 1)$ represents soft subgroup membership. This formulation preserves the semantics of a logical conjunction while making subgroup membership differentiable with respect to $\theta = \{\theta_j, d_j, a_j\}_{j=1}^{p}$. As the temperature decreases, the learned subgroup can be converted into an interpretable rule consisting of feature conditions and thresholds.For example, a learned soft subgroup may eventually be expressed as $X_1 > 0.82 \wedge X_3 < -0.14 \wedge X_6 > 0.51$.

Thus, although optimization occurs in a continuous parameter space, the final discovery remains interpretable as a conventional subgroup rule.

3.4 Differentiable Estimation of Temporal Discrepancy

The soft subgroup membership $S_\theta(X_i)$ is used as a differentiable weight when estimating the treatment and control trajectories. For treatment observations, $m_1(t \mid S_\theta) = \frac{\sum_{i=1}^N S_\theta(X_i) A_i Y_i(t)}{\sum_{i=1}^N S_\theta(X_i) A_i + \epsilon}$, while for control observations, $m_0(t \mid S_\theta) = \frac{\sum_{i=1}^N S_\theta(X_i)(1-A_i) Y_i(t)}{\sum_{i=1}^N S_\theta(X_i)(1-A_i) + \epsilon}$.

Therefore, $\Delta_\theta(t) = m_1(t \mid S_\theta) - m_0(t \mid S_\theta)$ provides a differentiable estimate of the temporal discrepancy. Importantly, gradients can propagate from the temporal discrepancy back through the trajectory estimates to the subgroup parameters: $\Delta_\theta(t) \to S_\theta(X) \to \{\theta_j, d_j, g_j\}$. This allows the subgroup itself to be optimized according to its temporal exceptionalness.

3.5 Differentiable Temporal-Window Discovery

The first temporal objective is to determine **when** the subgroup is most exceptional. Let $W = [t_s, t_e]$ denote an unknown temporal window. A conventional hard temporal indicator is $I_W(t) = I(t_s \leq t \leq t_e)$. However, the indicator is not differentiable with respect to t_s and t_e. We therefore replace it with a soft temporal membership function: $w_W(t) = \sigma\left(\frac{t-t_s}{\lambda_t}\right)\sigma\left(\frac{t_e-t}{\lambda_t}\right)$ where λ_t controls the smoothness of the temporal boundaries. The function $w_W(t)$ approaches one inside the temporal window and zero outside it, while maintaining non-zero gradients around the boundaries.

3.6 Proposed Temporal Quality Measure

The central methodological contribution of TDSD is a quality measure that explicitly incorporates temporal exceptionalness. For a candidate subgroup S and temporal window W, we define

$$Q_{\text{TDSD}}(S, W) = \frac{\sum_{t \in \mathcal{T}} w_W(t) \, |\Delta_S(t)|}{\sum_{t \in \mathcal{T}} w_W(t) + \epsilon}, \tag{2}$$

where $\Delta_S(t) = m_1(t \mid S) - m_0(t \mid S)$.

This measure has three important properties. First, it is temporal. Unlike a static mean difference, it evaluates the discrepancy at individual time points. Second, it is localized. The temporal membership $w_W(t)$ emphasizes the region in which the discrepancy is strongest, allowing the model to discover the temporal window rather than requiring it to be specified beforehand. Third, it is differentiable. Both subgroup membership and temporal membership are continuous, allowing gradients to flow through $S_\theta(X), \quad t_s, \quad t_e$. Consequently, $\nabla_\theta Q_{\text{TDSD}}, \quad \nabla_{t_s} Q_{\text{TDSD}}, \quad \nabla_{t_e} Q_{\text{TDSD}}$ can be computed directly.

For discrete longitudinal observations, the continuous integral formulation $\frac{\int_W |\Delta_S(t)| dt}{|W|}$ is naturally approximated by the weighted sum above. The use of the absolute discrepancy, $|\Delta_S(t)|$, allows TDSD to discover both positive and negative exceptional effects. Thus, the method is not restricted to subgroups that benefit from treatment; it can also identify subgroups with unusually negative treatment-control differences.

3.7 Temporal-Behavior Quality

Temporal-window discovery answers when, but it does not fully answer **how** the discrepancy evolves. To capture the complete temporal behavior, we define a trajectory-based quality measure: $Q_{\text{behavior}}(S) = \left[\frac{1}{|\mathcal{T}|}\sum_{t \in \mathcal{T}} \Delta_S(t)^2\right]^{1/2}$ which corresponds to the L_2 magnitude of the discrepancy trajectory. Unlike a signed average, $\frac{1}{T}\sum_t \Delta_S(t)$, the L_2 formulation does not allow positive and negative temporal effects to cancel each other. This is particularly important for oscillatory or non-monotonic temporal behavior. A trajectory may have a mean close to zero while exhibiting large and systematic temporal deviations. The trajectory itself, $\hat{\Delta}_S(t)$, is retained as the final representation

of subgroup behavior. Therefore, TDSD can distinguish patterns such as transient responses; delayed responses; increasing effects; decreasing effects; oscillatory effects. This formulation directly corresponds to the temporal-behavior recovery experiment in which these five types of ground-truth trajectories are considered.

3.8 JOINT OPTIMIZATION

The final TDSD optimization problem combines subgroup discovery and temporal localization. For temporal-window discovery, we formulate $\max_{\theta,t_s,t_e} \quad Q_{\mathrm{TDSD}}\left(S_\theta, W\right) - \lambda_S \Omega_S(S_\theta) - \lambda_W \Omega_W(W)$ where S_θ is the differentiable subgroup; $W = [t_s, t_e]$ is the differentiable temporal window; Q_{TDSD} is the proposed temporal quality measure; Ω_S controls subgroup complexity; Ω_W controls temporal-window complexity; λ_S, λ_W control the corresponding regularization strengths. A natural subgroup complexity penalty is $\Omega_S = \frac{1}{p}\sum_{j=1}^{p} g_j$, which encourages the discovery of concise rules. A temporal-window regularizer can be used to discourage unnecessarily large or irregular windows, for example through a penalty on window length, $\Omega_W(W) = \frac{t_e - t_s}{|\mathcal{T}|}$.

For temporal-behavior discovery, the objective can instead be written as $\max_\theta \quad Q_{\mathrm{behavior}}(S_\theta) - \lambda_S \Omega_S - \lambda_{\mathrm{smooth}} \Omega_{\mathrm{smooth}}$ where the trajectory smoothness regularizer is $\Omega_{\mathrm{smooth}} = \sum_t \left[\hat{\Delta}_S(t+1) - \hat{\Delta}_S(t)\right]^2$. This term discourages arbitrary point-wise fluctuations caused by noise while preserving meaningful temporal structure.

The resulting optimization can be solved using gradient-based methods such as Adam: $\Theta = \{\theta, t_s, t_e\}$, and $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}$. The complete discovery process therefore optimizes subgroup conditions and temporal localization within a unified framework.

Gradient flow and optimization behavior. The optimization is fully differentiable because the temporal discrepancy is expressed through weighted treatment and control trajectory estimates, and the weights are themselves functions of the subgroup parameters. Consequently, a change in a threshold or feature gate changes subgroup membership, which changes both treatment and control trajectories and ultimately changes the temporal quality. The same pathway applies to the temporal boundaries through $w_W(t)$. This coupling is important: the window is not selected after subgroup discovery, but is learned from the discrepancy induced by the current subgroup.

The subgroup and window penalties also serve different purposes. Ω_S discourages rules that activate many predicates, while Ω_W discourages unnecessarily broad temporal intervals. The denominator in Q_{TDSD} prevents a window from receiving a larger score merely because it contains more time points. Together, these terms favor compact population descriptions and temporally concentrated exceptionalness. In practice, the learned soft membership is additionally inspected for sufficient treatment and control support before the final rule is extracted; this separates numerical optimization from the requirement that the discovered contrast be estimable from both groups.

Support and boundary constraints. The temporal objective is meaningful only when both treatment and control trajectories are estimable within the discovered subgroup. We therefore treat subgroup support as a constraint on the discovery problem rather than relying solely on the numerical quality score. Similarly, the smooth window provides nonzero gradients around its boundaries while the window-length penalty discourages degenerate solutions that cover the entire observation horizon. These constraints make the learned temporal interval a structural summary of exceptionalness rather than merely an optimization artifact.

3.9 FROM SOFT SUBGROUPS TO INTERPRETABLE RULES

Although optimization uses soft membership, the final output is intended to remain interpretable. After optimization, the learned gates g_j are used to identify active features. For each active feature, the learned threshold θ_j and direction d_j determine the corresponding logical condition. For example, $g_1, g_3, g_6 \gg 0$, while the remaining gates are close to zero, may yield $S(X) = (X_1 > 0.82) \wedge (X_3 < -0.15) \wedge (X_6 > 0.50)$.

Similarly, the learned temporal parameters provide $\hat{W} = [\hat{t}_s, \hat{t}_e]$, which directly answers the when question. Finally, the estimated trajectory $\hat{\Delta}_S(t)$ answers the how question. Thus, the output of

TDSD is not merely a numerical score, but a structured discovery: $\left(\hat{S}, \hat{W}, \hat{\Delta}_S(t)\right)$ corresponding to Who + When + How.

4 EXPERIMENTS

We evaluate three dimensions of discovery and the contribution of temporal localization. Experiments 1–3 use controlled synthetic data; Experiment 4 compares increasingly time-aware quality measures; Experiment 5 evaluates a MIMIC-IV ICU treatment-response task Pollard et al. (2026).

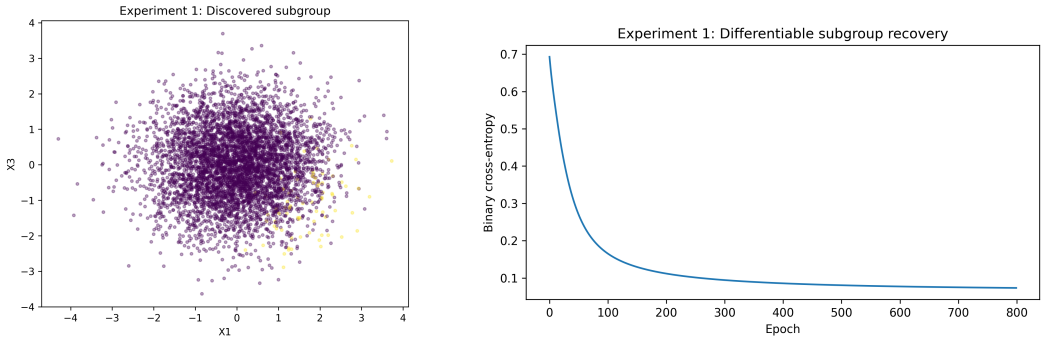
4.1 EXPERIMENTAL SETUP AND EVALUATION

The synthetic data are designed so that subgroup membership, temporal localization, and temporal behavior can be evaluated against known ground truth. Subgroup recovery is measured by precision, recall, F1, and Jaccard. Temporal-window recovery is evaluated using temporal intersection-over-union (IoU) and start/end boundary error. Temporal behavior is evaluated using trajectory correlation, RMSE, MAE, normalized error, peak-time error, and direction accuracy. These metrics intentionally separate three failure modes: selecting the wrong individuals, locating the wrong time period, and estimating the wrong trajectory shape.

For the clinical experiment, we use MAP trajectories over 0–24 hours and vasopressor exposure as treatment. The baseline variables include age, sex, SOFA, lactate, creatinine, and baseline MAP. The clinical analysis is observational and is therefore interpreted as pattern discovery rather than proof of a causal treatment effect.

4.2 EXPERIMENT 1: SYNTHETIC SUBGROUP RECOVERY

We generate 5,000 individuals with 10 covariates and 20 longitudinal observations. The ground-truth subgroup is $S^* = \{X_1 > 0.8 \wedge X_3 < 0 \wedge X_6 > 0.5\}$. TDSD obtains precision 0.750, recall 0.318, F1 0.447, and Jaccard 0.288; the discovered subgroup contains 76 instances versus 179 ground-truth instances. The largest learned feature weights correspond to X_1, X_6, and X_3 (1.631, 1.476, and -1.052), correctly identifying the dimensions used to define the planted subgroup.



(a) Discovered subgroup.

(b) Training curve.

Figure 1: Experiment 1. TDSD localizes the informative subgroup dimensions and the optimization converges smoothly. Exact hard-rule recovery is nevertheless incomplete, as reflected by recall.

The result demonstrates that the differentiable parameterization can identify the relevant feature dimensions, but it also reveals a limitation: the learned soft subgroup is more conservative than the planted rule. This limitation is important because a high-quality temporal trajectory does not automatically imply exact symbolic subgroup recovery.

4.3 Experiment 2: Temporal-Window Recovery

We next isolate temporal localization using 6,000 individuals, 10 covariates, and 30 time points. The true exceptional interval is $[12, 20]$. TDSD learns $[13.56, 19.05]$, giving temporal IoU 0.667, start boundary error 1.56, and end boundary error 0.95.

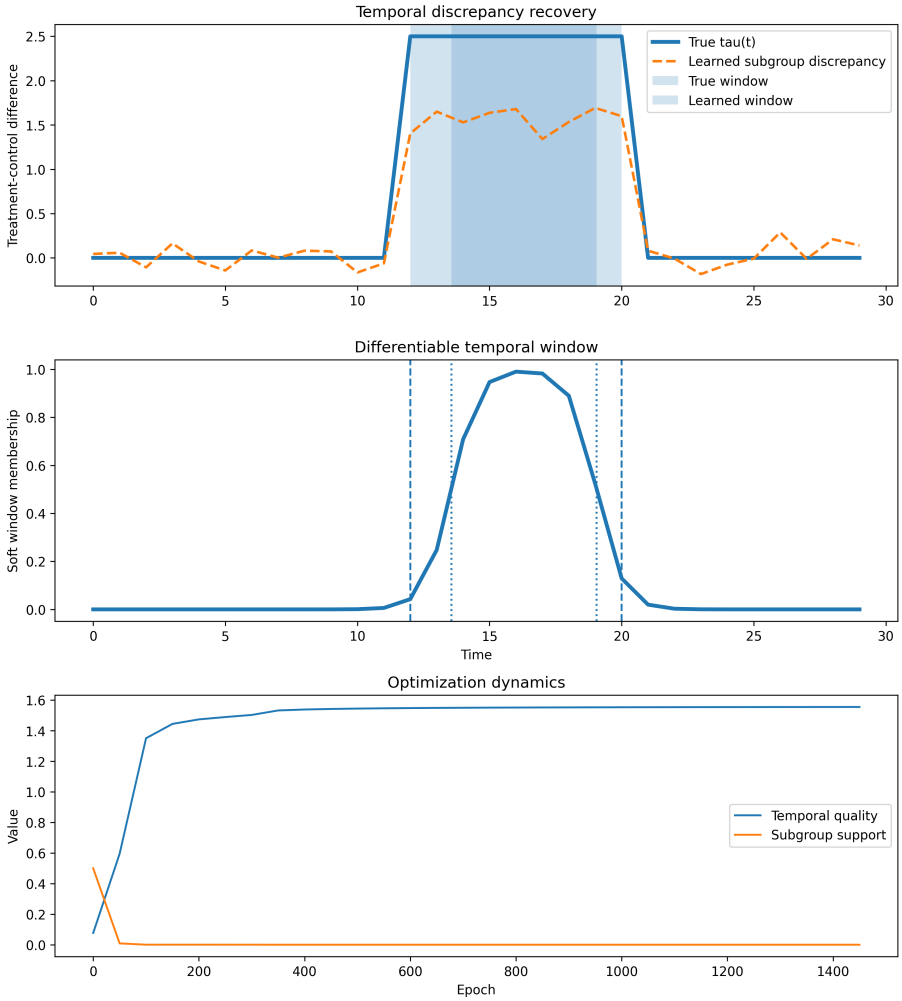


Figure 2: Experiment 2. Ground-truth and learned temporal discrepancy, soft temporal membership, and optimization dynamics. The learned interval moves toward the planted exceptional period.

The learned temporal membership rises and falls smoothly around the target interval, showing that temporal boundaries can be optimized rather than enumerated. After thresholding, hard subgroup F1/Jaccard are zero in this experiment, again separating temporal localization from exact rule extraction. The result supports the use of temporal IoU and boundary error as independent evaluation targets.

4.4 Experiment 3: Temporal-Behavior Recovery

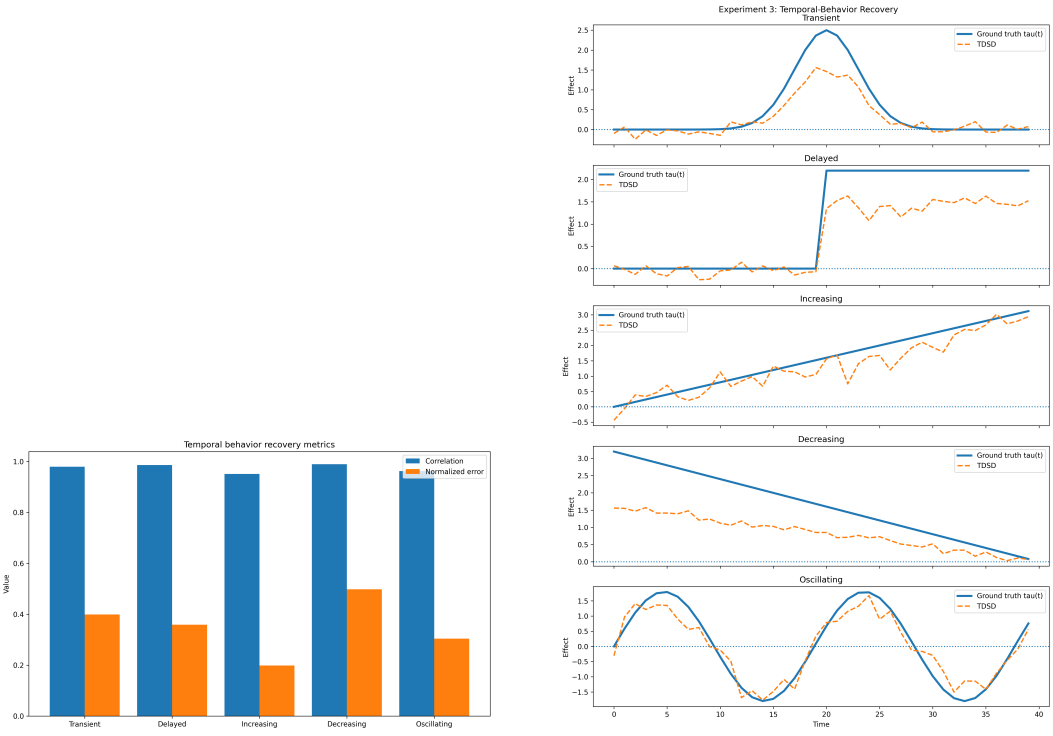
To test whether the method can characterize trajectories beyond a single window, we consider transient, delayed, increasing, decreasing, and oscillating temporal behaviors. Table 1 summarizes the complete metrics.

Trajectory correlation ranges from 0.951 to 0.990, indicating strong recovery of overall temporal shape. The delayed pattern reaches direction accuracy 1.000, while the increasing pattern has the smallest normalized error (0.199). In contrast, the decreasing pattern has high correlation (0.990)

Table 1: Temporal-behavior recovery.

Behavior	RMSE	MAE	Corr.	Norm. Err.	Peak Err.	Dir. Acc.
Transient	.363	.234	.980	.399	1	.536
Delayed	.558	.431	.987	.359	2	1.000
Increasing	.360	.290	.951	.199	3	.641
Decreasing	.937	.811	.990	.498	3	.615
Oscillating	.378	.321	.963	.304	19	.769

but relatively large RMSE (0.937), showing that correlation alone can conceal amplitude error. The oscillating pattern has correlation 0.963 but peak-time error 19, illustrating a complementary limitation in precise extremum localization.



(a) Metrics.

(b) Recovered trajectories.

Figure 3: Experiment 3. TDSD recovers diverse temporal behavior, while absolute magnitude and exact peak localization remain harder than trajectory shape.

These results motivate keeping the behavior objective separate from the window objective. A single interval is well matched to transient localization but can be a poor representation for broad, monotonic, or oscillatory behavior. TDSD therefore treats *when* and *how* as complementary outputs rather than forcing every pattern into a single window.

4.5 Experiment 4: Quality-Measure Ablation

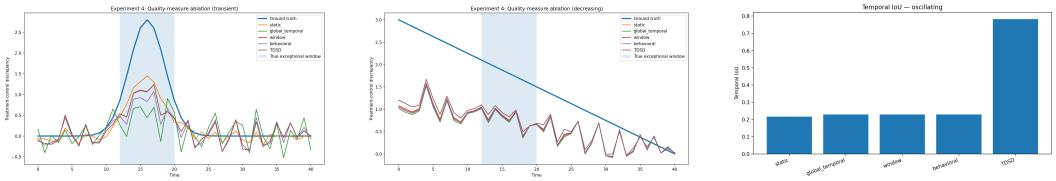
We compare Static, Global Temporal, Window, Behavioral, and TDSD objectives. Table 2 reports the aggregate results. The variants progressively add temporal aggregation, explicit window localization, or trajectory modeling; TDSD jointly uses subgroup and temporal-window parameters.

TDSD achieves the highest temporal IoU (0.455), trajectory correlation (0.865), and subgroup F1 (0.442), while reducing mean start/end boundary errors to 4.63/5.68. The strongest gains are tempo-

Table 2: Aggregate quality-measure ablation. Higher is better for F1, Jaccard, IoU, and correlation; lower is better for boundary errors and RMSE.

Variant	F1	Jac.	IoU	Start	End	Corr.	RMSE
Static	.382	.245	.284	6.80	12.40	.723	.998
Global Temporal	.422	.268	.225	6.20	15.40	.774	1.003
Window	.434	.277	.236	6.40	12.80	.827	.966
Behavioral	.435	.278	.286	6.40	12.40	.835	.967
TDSD	**.442**	**.284**	**.455**	**4.63**	**5.68**	**.865**	**.932**

ral rather than purely subgroup-based. This is consistent with the intended contribution: the method is designed to discover localized exceptionalness, not merely to improve a static rule score.



(a) Transient. (b) Decreasing. (c) Oscillating IoU.

Figure 4: Representative Experiment 4 results. Explicit temporal-window optimization is most beneficial when exceptionalness is localized or temporally structured.

The pattern-specific results further clarify the comparison. For transient and decreasing trajectories, TDSD concentrates the learned discrepancy more closely on the exceptional region. For the oscillating case, the alternatives remain near temporal IoU 0.22–0.23, whereas TDSD reaches approximately 0.783. Conversely, delayed and increasing trajectories are difficult to localize with a single window because their structure is broad or monotonic; these cases are better represented by the full trajectory objective. Thus, temporal localization is useful precisely when the phenomenon is temporally concentrated, while behavior modeling provides complementary coverage.

4.6 Experiment 5: MIMIC-IV ICU Treatment-Response Discovery

We use MAP over 0–24 hours and vasopressor exposure as treatment. Static SD, ATE-SD, and Temporal SD return the same rule $\text{baseline_MAP} \geq 91 \wedge \text{age} \leq 49 \wedge \text{SOFA} \geq 7 \wedge \text{creatinine} \geq 1.10$, covering 53 records (30 treated and 23 control), with an overall discrepancy of 16.003 and a late-window discrepancy of 21.814 over $[21, 23]$. TDSD instead returns the simpler rule $\text{lactate} \geq 2.10$, covering 6,753 records (1,755 treated and 4,998 control), with discrepancy 0.947 overall and 0.576 over $[4.52, 14.06]$. The reported permutation p-value is 0.0459, with bootstrap confidence-interval width 1.694.

Interpretation. The methods select different populations, so these quantities are not estimates of a common causal estimand. The comparison instead illustrates how the discovery objective changes the subgroup size, discrepancy magnitude, and temporal region being reported.

5 Discussion and Conclusion

TDSD formulates longitudinal subgroup discovery as the structured object $(S, W, \Delta_S(t))$. The experiments support the who–when–how decomposition: subgroup learning identifies informative covariate dimensions, temporal localization finds exceptional periods, and trajectory modeling captures behavior beyond a single window. Exact hard-rule recovery and single-window representation remain limitations, motivating better soft-to-hard extraction, adaptive temporal representations, and uncertainty-aware discovery. The MIMIC-IV analysis is exploratory rather than causal validation because treatment-time alignment and identification assumptions are not established.

REFERENCES

Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.

Martin Atzmueller, Carolina Centeio Jorge, Cláudio Rebelo de Sá, Behzad M Heravi, Jenny L Gibson, and Rosaldo JF Rossetti. Mining exceptional social behavior on attributed interaction networks. *Machine Learning*, 114(11):243, 2025.

Chengzhi Cao, Chao Yang, Ruimao Zhang, and Shuang Li. Discovering intrinsic spatial-temporal logic rules to explain human actions. *Advances in Neural Information Processing Systems*, 36: 67948–67959, 2023.

Michael Daly, Angelina R Sutin, and Eric Robinson. Longitudinal changes in mental health and the covid-19 pandemic: evidence from the uk household longitudinal study. *Psychological medicine*, 52(13):2549–2558, 2022.

Xin Du. The uncertainty in exceptional model mining. 2020.

Xin Du, Yulong Pei, Wouter Duivesteijn, and Mykola Pechenizkiy. Fairness in network representation by latent structural heterogeneity in observational data. In *AAAI*, volume 34, pp. 3809–3816, 2020.

Xin Du, Sikun Yang, Wouter Duivesteijn, and Mykola Pechenizkiy. Conformalized exceptional model mining: Telling where your model performs (not) well. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 528–544. Springer, 2025.

Wouter Duivesteijn and Julia Thaele. Understanding where your classifier does (not) work–the scape model class for emm. In *ICDM*, pp. 809–814. IEEE, 2014.

Wouter Duivesteijn, Arno Knobbe, Ad Feelders, and Matthijs van Leeuwen. Subgroup discovery meets bayesian networks–an exceptional model mining approach. In *ICDM*, pp. 158–167. IEEE, 2010.

Wouter Duivesteijn, Ad Feelders, and Arno Knobbe. Different slopes for different folks: mining for exceptional regression models with cook's distance. In *KDD*, 2012.

Wouter Duivesteijn, Ad J Feelders, and Arno Knobbe. Exceptional model mining: Supervised descriptive local pattern mining with complex target concepts. *DMKD*, 2016.

Donald Hedeker and Robert D Gibbons. *Longitudinal data analysis*. John Wiley & Sons, 2006.

Franciso Herrera, Cristóbal José Carmona, Pedro González, and María José Del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, 29 (3):495–525, 2011.

Mehdi Kaytoue, Marc Plantevit, Albrecht Zimmermann, Anes Bendimerad, and Céline Robardet. Exceptional contextual subgraph mining. *Machine Learning*, 106:1171–1211, 2017.

Dennis Leman, Ad Feelders, and Arno Knobbe. Exceptional model mining. In *ECMLPKDD*, pp. 1–16. Springer, 2008.

Florian Lemmerich, Martin Becker, Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. Mining subgroups with exceptional transition behavior. In *KDD*, pp. 965–974, 2016.

Ilya Lipkovich, David Svensson, Bohdana Ratitch, and Alex Dmitrienko. Overview of modern approaches for identifying and evaluating heterogeneous treatment effects from clinical data. *Clinical Trials*, 20(4):380–393, 2023.

Ilya Lipkovich, David Svensson, Bohdana Ratitch, and Alex Dmitrienko. Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data. *Statistics in Medicine*, 43(22):4388–4436, 2024.

John J McArdle. Latent variable modeling of differences and changes with longitudinal data. *Annual review of psychology*, 60(1):577–605, 2009.

Luis NJ Paulus, Jonas Fischer, and Jilles Vreeken. Differentiably discovering sets of rules. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3931–3942, 2026.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, August 2026. ISSN 3005-0693. doi: 10.1038/s44360-026-00096-z. URL https://doi.org/10.1038/s44360-026-00096-z.

Hugo M Proença, Peter Grünwald, Thomas Bäck, and Matthijs van Leeuwen. Robust subgroup discovery: Discovering subgroup lists using mdl. *DMKD*, 36(5):1885–1970, 2022.

Filippo Salmaso, Lorenzo Testa, and Francesca Chiaromonte. A doubly robust machine learning approach for disentangling treatment effect heterogeneity with functional outcomes. *arXiv preprint arXiv:2602.11118*, 2026.

Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Fang Zheng, Nanning Zheng, and Gang Hua. Social interpretable tree for pedestrian trajectory prediction. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2235–2243, 2022.

Youmi Suk and Chan Park. Causal mediation and functional outcome analysis with process data. *Psychometrika*, 91(2):424–444, 2026.

Marius Thomas and Björn Bornkamp. Comparing approaches to treatment effect estimation for subgroups in clinical trials. *Statistics in Biopharmaceutical Research*, 9(2):160–171, 2017.

Weize Wang, Zoran Bursac, and Nan Hu. Joint modeling of longitudinal and survival data in public health and biomedical research: A systematic review. *International Journal of Environmental Research and Public Health*, 23(4):492, 2026.

Yishu Wei, Lei Liu, Xiaogang Su, Lihui Zhao, and Hongmei Jiang. Precision medicine: Subgroup identification in longitudinal trajectories. *Statistical methods in medical research*, 29(9):2603–2616, 2020.

Matthias Wilms, Sascha Xu, and Jilles Vreeken. Explainable mixture models through differentiable rule learning. In *International Conference on Learning Representations*, volume 2026, pp. 140165–140185, 2026.

Xian Wu, Guanbin Li, Qingxing Cao, Qingge Ji, and Liang Lin. Interpretable video captioning via trajectory structured localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6829–6837. IEEE, 2018.

Sascha Xu and Jilles Vreeken. Differential subgroup discovery: Characterizing where two populations differ, and why. *arXiv preprint arXiv:2604.27741*, 2026.

Sascha Xu, Nils Philipp Walter, Janis Kalofolias, and Jilles Vreeken. Learning exceptional subgroups by end-to-end maximizing kl-divergence. *arXiv preprint arXiv:2402.12930*, 2024.

Sascha Xu, Nils Philipp Walter, and Jilles Vreeken. Neural rule lists: Learning discretizations, rules, and order in one go. *Advances in Neural Information Processing Systems*, 38:148617–148645, 2026.

Lincen Yang, Siri L van der Meijden, Sesmu M Arbous, and Matthijs van Leeuwen. Interpretable machine learning for identifying icu readmission risk in subgroups with probabilistic rules. *Journal of the American Medical Informatics Association*, 33(3):690–699, 2026.

Zhongheng Zhang, Heidi Seibold, Mario V Vettore, Woo-Jung Song, and Vieille François. Subgroup identification in clinical trials: an overview of available methods and their implementations with r. *Annals of translational medicine*, 6(7):122, 2018.

A Detailed Problem Setting and Preliminary Material

Let $\mathcal{D} = \{(X_i, A_i, Y_i(t))\}_{i=1}^N$ denote a longitudinal dataset. An unknown subgroup S^* is defined by a conjunction of interpretable conditions. For synthetic experiments, $Y_i(t) = \mu(t) + A_i S_i^* \tau(t) + \epsilon_i(t)$, where $\mu(t)$ is a common baseline, S_i^* indicates subgroup membership, $\tau(t)$ is the subgroup-specific temporal effect, and $\epsilon_i(t)$ is noise.

For a candidate subgroup S, $m_1(t \mid S)$ and $m_0(t \mid S)$ are treatment and control trajectories and $\Delta_S(t) = m_1(t \mid S) - m_0(t \mid S)$. Under consistency, exchangeability, and positivity, this can be interpreted causally; otherwise it is an outcome discrepancy. Window quality is $\int_W |\Delta_S(t)|dt/|W|$, and behavior quality is the trajectory L_2 magnitude. Evaluation uses subgroup F1/Jaccard, temporal IoU and boundary error, trajectory correlation/RMSE/MAE, peak-time error, and direction accuracy.

B Extended Methodology

B.1 Methodological Overview and Design Rationale

TDSD replaces the discrete components of rule-based subgroup discovery with differentiable representations: soft predicates for feature conditions, feature-selection gates for rule sparsity, and a smooth temporal membership function for interval boundaries. These components are optimized jointly through the temporal discrepancy objective. The resulting framework is designed to answer three complementary questions: who is exceptional, when the discrepancy is concentrated, and how it evolves over time.

B.2 Why the Temporal Quality Measure Is Necessary

The static formulation collapses temporal discrepancy into one scalar and cannot distinguish different temporal arrangements with similar averages. A global temporal formulation retains all time points but does not identify when exceptionalness occurs. A window formulation adds localization, while the behavior formulation retains the full trajectory but does not explicitly identify a temporal region. The proposed TDSD objective combines subgroup learning and temporal localization and is therefore evaluated against these progressively time-aware alternatives in the ablation study.

B.3 Methodological Contributions

The methodological contributions are the joint who–when–how formulation, differentiable interpretable subgroup generation, temporal discrepancy estimation, differentiable temporal-window discovery, joint optimization of subgroup and temporal parameters, and complementary temporal-behavior discovery. The final output is the structured object $(\hat{S}, \hat{W}, \hat{\Delta}_S(t))$, rather than a single scalar quality score.

The differentiable predicate is $p_j(X_j) = \sigma(d_j(X_j - \theta_j)/\lambda_x)$ with $d_j = \tanh(\tilde{d}_j)$ and gate $g_j = \sigma(a_j)$. The normalized logarithmic conjunction is

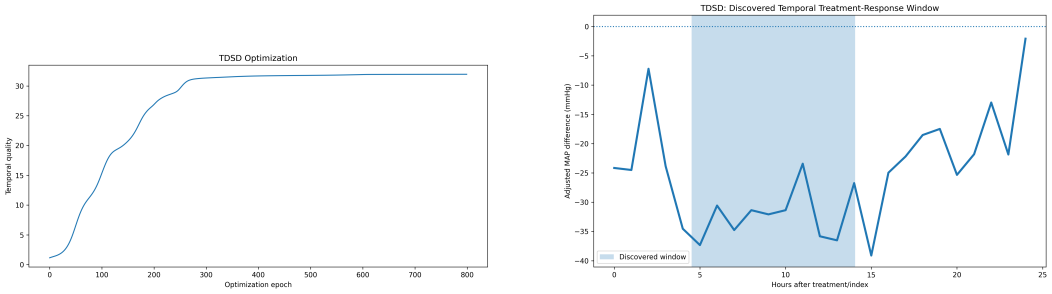
$$S_\theta(X) = \exp\left[\frac{\sum_j g_j \log(p_j(X) + \epsilon)}{\sum_j g_j + \epsilon}\right].$$

Weighted trajectory estimates are

$$m_a(t \mid S_\theta) = \frac{\sum_i S_\theta(X_i) \mathbb{1}[A_i = a] Y_i(t)}{\sum_i S_\theta(X_i) \mathbb{1}[A_i = a] + \epsilon}.$$

The soft temporal window is $w_W(t) = \sigma((t - t_s)/\lambda_t)\sigma((t_e - t)/\lambda_t)$ and Eq. 2 is optimized jointly with subgroup and window complexity penalties. For behavior discovery, Q_{beh} is optimized with a smoothness term $\sum_t [\Delta(t+1) - \Delta(t)]^2$. The final structured discovery is $(\hat{S}, \hat{W}, \hat{\Delta}_S(t))$.

Soft-to-hard extraction. For a final interpretable rule, feature gates and predicate probabilities are thresholded. Features with low gates are removed, while retained features use their learned directions and thresholds. This procedure is deliberately evaluated separately from the soft objective because thresholding can change subgroup support and therefore alter the estimated discrepancy trajectory.



(a) Optimization trajectory.

(b) Discovered temporal window.

Figure 5: Supplementary diagnostics for temporal-window optimization.

C Additional Experimental Details and Results

C.1 Experiment 1: Synthetic Subgroup Recovery

The synthetic setting contains 5,000 individuals, 10 covariates, and 20 observations. The ground-truth subgroup is $X_1 > 0.8 \wedge X_3 < 0 \wedge X_6 > 0.5$. TDSD obtains precision 0.750, recall 0.318, F1 0.447, and Jaccard 0.288; the discovered subgroup contains 76 instances versus 179 ground-truth instances. The three largest absolute feature weights correspond to X_1, X_6, and X_3 (1.631, 1.476, and -1.052).

C.2 Experiment 2: Temporal-Window Recovery

The true interval is $[12, 20]$ and TDSD learns $[13.56, 19.05]$, with IoU 0.667, start error 1.56, and end error 0.95. The learned temporal membership rises and falls smoothly around the target interval. Hard subgroup F1/Jaccard are zero after thresholding.

C.3 Experiment 3: Temporal-Behavior Recovery

Table 3: Temporal-behavior recovery metrics.

Behavior	RMSE	MAE	Corr.	Norm. Err.	Peak Err.	Dir. Acc.
Transient	.363	.234	.980	.399	1	.536
Delayed	.558	.431	.987	.359	2	1.000
Increasing	.360	.290	.951	.199	3	.641
Decreasing	.937	.811	.990	.498	3	.615
Oscillating	.378	.321	.963	.304	19	.769

C.4 Experiment 4: Quality-Measure Ablation

Table 4: Full aggregate ablation results.

Variant	F1	Jac.	IoU	Start	End	Corr.	RMSE
Static	.382	.245	.284	6.80	12.40	.723	.998
Global Temporal	.422	.268	.225	6.20	15.40	.774	1.003
Window	.434	.277	.236	6.40	12.80	.827	.966
Behavioral	.435	.278	.286	6.40	12.40	.835	.967
TDSD	.442	.284	.455	4.63	5.68	.865	.932

The pattern-specific results show the largest localization gains for transient, decreasing, and oscillating behaviors. For delayed and increasing patterns, the single-window formulation is less ap-

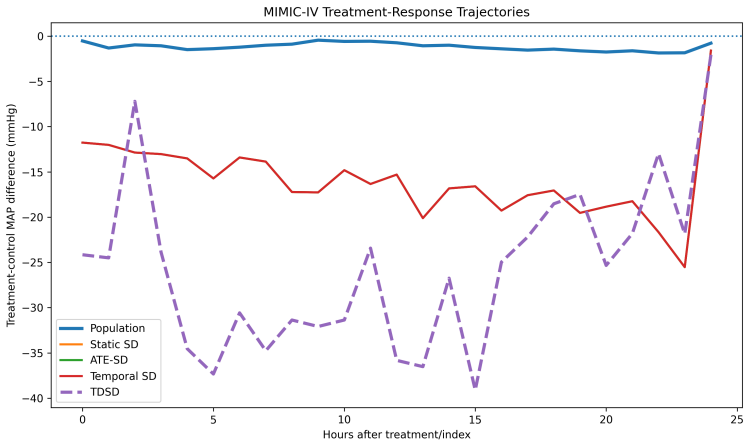


Figure 6: Supplementary MIMIC-IV MAP trajectories for the TDSD-discovered subgroup.

propriate because the exceptional change is broad or monotonic. This explains why high trajectory recovery can coexist with weak window IoU.

C.5 Experiment 5: MIMIC-IV ICU Treatment-Response Discovery

The clinical experiment uses MAP over 0–24 hours and vasopressor exposure as treatment. Static SD, ATE-SD, and Temporal SD return the same rule $\text{baseline_MAP} \geq 91 \wedge \text{age} \leq 49 \wedge \text{SOFA} \geq 7 \wedge \text{creatinine} \geq 1.10$, covering 53 records (30 treated, 23 control), with overall effect 16.003 and window effect 21.814 over $[21, 23]$. TDSD returns $\text{lactate} \geq 2.10$, covering 6,753 records (1,755 treated, 4,998 control), with overall effect 0.947 and window effect 0.576 over $[4.52, 14.06]$. The reported permutation p-value is 0.0459 and bootstrap confidence-interval width is 1.694. These estimates correspond to different discovered populations and are not directly comparable as a common causal estimand.

D Reproducibility Notes

Synthetic experiments are defined by the ground-truth subgroup and temporal behaviors reported above. The MIMIC-IV experiment uses the 0–24 hour MAP trajectory and the baseline variables described in the main text. Because MIMIC-IV is a restricted-access clinical dataset, the raw patient-level data are not redistributed with this submission.

E Additional Discussion of Interpretation and Validation

Soft versus symbolic discovery. The differentiable model is a search mechanism rather than the final symbolic rule. Soft predicates and gates provide gradients for threshold and feature selection, while post-hoc thresholding converts the result into an interpretable conjunction. The experiments therefore evaluate symbolic recovery separately from temporal recovery. This distinction is important when a conservative soft subgroup has high precision but incomplete coverage.

Complementarity of temporal objectives. Window quality and behavior quality summarize different aspects of the same discrepancy trajectory. A transient response can be summarized effectively by a localized interval, whereas delayed, monotonic, or oscillatory responses may require the full trajectory. The ablation results support retaining both outputs instead of forcing every temporal pattern into one representation.

Clinical validation. For observational clinical applications, future validation should examine treatment-time alignment, covariate balance, alternative baseline definitions, bootstrap stability, and

sensitivity to subgroup thresholds. These checks do not establish causality by themselves, but they can distinguish reproducible temporal structure from artifacts of a particular parameterization or sample realization.

F Additional Limitations and Future Directions

The current implementation uses a single soft interval for explicit temporal localization. This is advantageous for interpretability but cannot represent disconnected exceptional periods without extending the window parameterization. The soft-to-hard extraction procedure can also reduce support substantially, which explains the gap between continuous optimization and hard-rule F1. Finally, the observational clinical experiment does not by itself establish causal treatment effects. Future extensions can address these issues through multi-window discovery, uncertainty-aware subgroup scoring, causal adjustment, and more principled symbolic extraction.