DISCERN: IDENTIFIABLE DISEASE EVIDENCE FOR GENERALIZABLE MEDICAL MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal clinical models are trained on cohorts that differ in whom they enroll and how they measure. Enrollment that ties background factors such as age to the disease state leaks into every modality, so models learn shortcuts that fail on new cohorts. We model medical multimodal data as a latent structural causal model in which a latent disease state drives, and in a second regime rewires, a graph of latent factors; each modality and the recorded endpoint are partial views of these factors, and cohorts differ only in selection and measurement. We show that the disease state need not be observed and the full model need not be recovered: in common Gaussian coordinates, states can be aligned across cohorts, enrollment-induced changes identify a minimal selection statistic, and conditioning on it yields disease evidence that is identifiable without state labels and invariant to enrollment shifts within the identified family. When the disease state rewires the factors, this evidence can be invisible to each modality and to late fusion of their posteriors, yet recoverable from the modalities jointly. DISCERN instantiates the theory in a jointly trained encoder, latent-state model and outcome head. Under reversed synthetic enrollment, conditioning raises state AUROC from 0.043 to 0.883 and from 0.265 to 0.825, while random conditioning does not; on cross-database mortality prediction between eICU and MIMIC-IV, DISCERN improves AUROC over matched empirical risk minimization by 0.036 and 0.035.

1 INTRODUCTION

Medical multimodal learning integrates imaging, laboratory values, vital signs, cognitive scores, demographics, and genetics to support clinical prediction (Acosta et al., 2022; Kline et al., 2022; Huang et al., 2020). These observations provide complementary, partial views of a patient's underlying condition. Fusion architectures, cross-modal pre-training, and missing-modality handling have improved predictive performance (Hayat et al., 2022; Soenksen et al., 2022; Liang et al., 2024; Elsharief et al., 2025), and theory explains when multimodal representations help (Huang et al., 2021; Lu, 2023; Dufumier et al., 2025). Yet combining more information does not establish which of that information remains reliable when the model moves to a different cohort.

The difficulty is that clinical observations reflect both the patient and the process through which the patient enters a dataset. Cohorts differ in *measurement*: scanners, assays, and labeling protocols change how a condition is recorded. They also differ in *selection*: enrollment changes disease prevalence and the association between disease and background factors such as age. Because background factors also shape physiology, selection can change the distribution of each modality within a disease state even when the underlying physiological mechanisms remain fixed. Consider Alzheimer's disease assessment using structural MRI and cognitive scores. Both hippocampal volume and memory performance vary with age even among cognitively normal individuals (Jack et al., 2015). If a source cohort enrolls older patients with Alzheimer's disease and younger controls, both modalities can reinforce an age-related shortcut that becomes unreliable in a target cohort with a different age–disease association. Cross-modal agreement therefore need not indicate stable disease evidence. This concern is consistent with documented shortcut learning and failures of external validation in imaging and clinical prediction (Zech et al., 2018; DeGrave et al., 2021; Wong et al., 2021; Finlayson et al., 2021; Allyn et al., 2026; Tranchellini et al., 2026).

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.

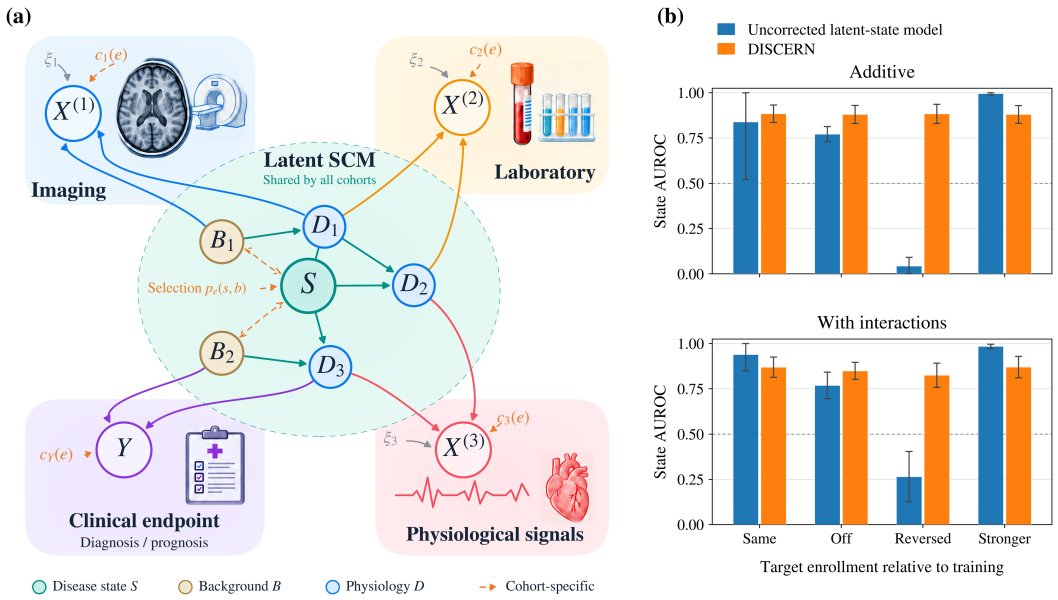


Figure 1: **A generative view of medical multimodal data and enrollment-induced failure.** (a) Disease state S and background factors B generate physiology D through a latent structural causal model shared across cohorts. Each modality observes an overlapping subset of latent coordinates with private variation ξ_m; the endpoint Y is another partial view. Orange indicates cohort-dependent selection $p_e(s,b)$ and measurement $c_m(e)$, $c_Y(e)$. (b) On the generator's latent coordinates, target cohorts preserve, remove, reverse, or strengthen the source enrollment pattern. The uncorrected latent-state model follows the enrollment and collapses under reversal, whereas the conditional evidence of DISCERN remains stable across all four settings (Table 1).

A further challenge is that the disease state itself is not directly observed. The recorded diagnosis or outcome is an imperfect, protocol-dependent view that can also depend on physiology beyond the state. This distinction complicates existing approaches to generalization. Domain-generalization methods seek stable predictors, risks, or feature statistics across training environments (Arjovsky et al., 2019; Krueger et al., 2021; Sagawa et al., 2020; Sun & Saenko, 2016), but their success depends on specifying an appropriate notion of invariance (Rosenfeld et al., 2021; Kamath et al., 2021); suppressing variation across cohorts can also remove disease information when prevalence and case mix change. Nuisance-based and causal methods make stability assumptions explicit (Puli et al., 2022; Wang et al., 2022; Peters et al., 2016; Londschien et al., 2026), but typically formulate prediction around an observed target or anchor. Causal representation learning establishes conditions for identifying latent variables (Khemakhem et al., 2020; Kivva et al., 2022; Yao et al., 2024; Sun et al., 2025), yet latent-variable identification alone does not determine which evidence transfers under changing enrollment (Appendix F discusses related work in detail). These limitations motivate our central question: *Which evidence about an unobserved disease state is identifiable from differently assembled cohorts and remains valid when enrollment changes?*

We address this question through a generative model of medical multimodal data (Figure 1a). A latent disease state and background factors jointly generate physiology through mechanisms shared across cohorts. Modalities observe overlapping subsets of these latent quantities, while the recorded endpoint provides an additional partial view. Cohorts vary in selection and measurement; our identification analysis isolates selection by holding acquisition fixed. Figure 1b illustrates why this distinction matters: even on the generator's latent coordinates, an uncorrected state model can exploit the source enrollment pattern and fail when that pattern reverses. Recovering latent coordinates is therefore insufficient unless the evidence extracted from them is also separated from enrollment.

Our answer is *conditional disease evidence*: the state log-likelihood ratio after subtracting the contribution carried by a *selection statistic*. This statistic captures how enrollment changes the state-conditional distributions across cohorts. Conditioning on it removes this contribution while retain-

ing the information needed to interpret physiology in context. In the Alzheimer's example, the aim is to interpret MRI findings and cognitive scores relative to a patient's age while excluding the enrollment-dependent age–disease association from the evidence. In common Gaussian coordinates with state-specific covariances shared across cohorts, we give conditions that align unobserved states up to one global permutation. Once aligned, their natural-parameter changes identify a minimal linear selection statistic and, in turn, the conditional evidence and a reference posterior. The same evidence transfers to target cohorts whose selection shifts lie within the identified family.

Estimating this evidence from multimodal observations also requires preserving the information on which it depends. Disease evidence may reside in interactions invisible to individual-modality posteriors or in latent factors observed by only a subset of modalities. Our analysis therefore motivates joint modeling and representations that retain shared, partially shared, and private information. These principles lead to DISCERN, which combines an information-lattice encoder with an aligned latent-state model and a conditional likelihood-ratio correction. Because the recorded endpoint can depend on continuous physiology beyond the disease state, its outcome head uses both conditional evidence and continuous features. A separate risk decomposition characterizes why improved state evidence need not translate directly into improved endpoint prediction.

In summary, (i) we formulate cohort variation through selection and measurement in a shared physiological model and, under explicit coordinate and alignment assumptions, identify a minimal linear selection statistic and the conditional disease evidence up to a global state permutation (Section 3); (ii) we develop DISCERN, which estimates this evidence without disease-state labels and combines it with continuous physiology for outcome prediction (Section 4); and (iii) on controlled enrollment shifts, conditioning on the estimated selection statistic restores state recovery under reversed enrollment where random conditioning does not, and on cross-database ICU mortality DISCERN improves AUROC over matched ERM in both directions (Section 5).

2 Problem formulation

Data-generating model. Let $S \in \{0, \dots, K-1\}$ be the latent disease state, B the background factors that pre-exist the disease (age, genetics), and D the physiology that the disease shapes. The physiology is generated by mechanisms that are the same in every cohort, $D_j = f_j(D_{\mathrm{pa}(j)}, B, S, \epsilon_j)$ along a directed acyclic graph, so background factors enter the physiology as parents (age enters neurodegeneration). We write $Z = (B, D)$ for the full latent vector and keep B apart from D because a cohort can change the two differently: selection can associate B with S, but it cannot alter how S and B shape D. Modality m observes a subset of it, $X^{(m)} = g_m(Z_{A_m}, \xi_m;\, c_m(e))$, with private variation ξ_m; the overlaps of the A_m define the shared, pairwise and private blocks of an *information lattice*. The recorded endpoint $Y \sim p(y \mid s, z;\, c_Y(e))$ is one more partial view, not the state, and can depend on physiology beyond S.

How cohorts differ: selection and measurement. A cohort e generates its data as

$$p_e(s, b, d, x, y) = \underbrace{\pi_e(s)\, p_e(b \mid s)}_{\text{selection}}\; p(d \mid b, s)\; \prod\nolimits_m p\big(x^{(m)} \mid z_{A_m}; c_m(e)\big)\; p\big(y \mid s, z; c_Y(e)\big). \quad (1)$$

Only two things depend on e. *Selection* is whom the cohort enrolls: its prevalence π_e and how strongly its enrollment ties background to disease state, $p_e(b \mid s)$; a memory clinic enrolls cases older than its controls, a population cohort does not. *Measurement*, $c(e) = (c_1(e), \dots, c_M(e), c_Y(e))$, is how the cohort records them: scanners, assays and label protocol. The mechanisms $p(d \mid b, s)$ are fixed. Selection is what corrupts state evidence: the B–S association belongs to the sampling law, not to the disease, yet the physiology inherits it, so a pooled learner sees it in every modality that observes D. We therefore treat selection as the shift to be removed, hold acquisition fixed for identification, and carry label-protocol drift as a separate term of the outcome risk (Proposition 3). Orthogonal to both is the *mechanism regime*: additive, where the state shifts physiological levels, or with interactions, where it also changes couplings and part of the evidence lives in covariances.

Problem. We observe source cohorts $e \in \mathcal{E}_{\mathrm{src}}$ with modalities and endpoint but never S, and a target cohort with unlabeled modalities only. We seek a function of the modalities whose evidence about S does not depend on selection and is identifiable without observing S, and a predictor of Y

that uses it. To state what can be recovered we assume common coordinates $z \in \mathbb{R}^d$ for the latent vector in which each cohort is a Gaussian mixture over states,

$$p_e(z) = \sum_{s=0}^{K-1} \pi_{es}\, \mathcal{N}(z; \mu_{es}, \Sigma_s), \qquad \pi_{es} > 0,\ \Sigma_s \succ 0, \tag{2}$$

with components distinct within each cohort and state-specific covariances shared across cohorts (A1). Cohort-specific weights and means carry selection; shared covariances carry the mechanisms, including interactions. Let $U = \Gamma^\top Z$ collect the directions along which the state-conditional densities change across cohorts (the *selection statistic*, constructed in Section 3). For a fixed positive reference prior π we define the *conditional disease evidence* and the *reference posterior*

$$T_s(z) = \log \frac{p_e(z \mid s)}{p_e(z \mid 0)} - \log \frac{p_e(U(z) \mid s)}{p_e(U(z) \mid 0)}, \qquad q_\pi^{\mathrm{ref}}(s \mid z) = \mathrm{softmax}_s\{\log \pi_s + T_s(z)\}, \tag{3}$$

with $T_0 = 0$. The first term is the evidence a cohort's own state model would use; the second is the part of it that selection created. Their difference keeps the background as a covariate rather than as evidence. In Figure 1a, where selection acts on the background, the linear-Gaussian instance makes U an invertible linear function of B once source enrollment varies along every background factor (Appendix B.3), and then $T_s(z) = \log p(d \mid b, s)/p(d \mid b, 0)$ is exactly the likelihood ratio of the fixed mechanisms: a laboratory value is read relative to age, while the age–disease association that selection created is removed. The learner need not know which coordinates are background: U is identified from how state densities change across source cohorts. The reference posterior generally differs from the cohort's Bayes posterior $p_e(s \mid z)$, which uses selection as evidence and does not transfer. We further assume (A2) at least three source cohorts that differ in enrollment, (A3) one of the alignment conditions of Proposition 1, and, for transfer, (A4) that the target's state-conditional laws are Gaussian with the same Σ_s and shifts in $\mathcal{U}$.

3 Identifying disease evidence without observing the state

We first align the latent states across cohorts, then identify the selection statistic and the evidence that remains after conditioning on it, show why estimating this evidence requires joint observations and information-preserving representations, and relate it to prediction of the recorded endpoint. The identification results assume the common coordinates of Equation (2); recovering such coordinates from raw modalities is a separate requirement. Full statements and proofs appear in Appendix B.

3.1 Aligning unknown states across cohorts

Each cohort's Gaussian mixture identifies its components only up to a permutation (Yakowitz & Spragins, 1968), and independent permutations in different cohorts do not define a common disease variable. Either of the following conditions establishes the correspondence.

Proposition 1 (Global state alignment)**.** *Under Equation (2), states can be aligned across source cohorts up to one global permutation if the shared covariances $\{\Sigma_s\}$ are pairwise distinct. For the three-state family with a common covariance, alignment also holds, up to a global exchange of the two end states, if*

$$\mu_{e0} = -\kappa_e \nu, \qquad \mu_{e1} = a, \qquad \mu_{e2} = 2a + \kappa_e \nu, \tag{4}$$

where a and ν are linearly independent and at least three source cohorts have distinct values of κ_e.

Distinct covariances provide a unique match for each state. In Equation (4), a is both a component mean and the arithmetic centre of every cohort's three component means. After centring at a, the remaining means lie on two distinct parallel lines, each containing at least three observed points. These lines determine the correspondence up to their exchange, without knowledge of κ_e or a zero-enrollment cohort. Alignment assigns consistent state identities, not their clinical interpretation.

3.2 The selection statistic and identifiable conditional evidence

Once states are aligned, their changes across cohorts identify which statistic to condition on. For a reference source cohort e_0, define

$$\mathcal{U} = \mathrm{span}\left\{\Sigma_s^{-1}(\mu_{es} - \mu_{e_0 s}) : e, s\right\}, \qquad \Gamma \text{ a full-column-rank basis of } \mathcal{U}. \tag{5}$$

These natural-parameter changes differ from mean differences when physiology is correlated.

Theorem 1 (Identifiable conditional evidence)**.** *Assume Equation (2) and an alignment condition of Proposition 1. In the common coordinates, $\mathcal{U}$ is identifiable from the source laws and is the smallest linear space whose associated statistic determines every same-state density ratio between source cohorts. With $U = \Gamma^\top Z$, the conditional law of Z given (U, S) is common to all source cohorts. Consequently, the reference posterior in Equation (3) is identifiable up to one global state permutation, with π permuted accordingly; the evidence T is identified with the corresponding choice of baseline state. These conclusions extend to targets satisfying (A4).*

The key identity is that the same-state log-density ratio between two cohorts is linear in z with coefficient in $\mathcal{U}$, so it depends only on U and conditioning on U cancels it. If V completes U to invertible coordinates, the remaining conditional density is a shared kernel $h_s(v \mid u)$, giving $T_s(z) = \log h_s(v \mid u) - \log h_0(v \mid u)$ for any basis of $\mathcal{U}$. Useful evidence requires variation beyond the selection statistic: if $\mathcal{U} = \mathbb{R}^d$, then $T = 0$ and $q_\pi^{\mathrm{ref}} = \pi$. Source variation must also cover the target shifts stipulated by (A4), and retaining fewer directions than $\dim \mathcal{U}$, as a low-rank estimator may do, is an approximation. Appendix B.3 gives sufficient structural conditions for rank one, which a single scalar enrollment parameter does not by itself ensure.

3.3 Reading the modalities jointly and without loss

Evidence can be lost in two ways. First, fusing per-modality state predictions discards cross-modal dependence, which can carry all state information: if two Gaussian views have correlation $+\rho$ in one state and $-\rho$ in the other, both single-view posteriors are constant, whereas the joint Bayes error is $\arccos(\rho)/\pi < 1/2$ (Proposition 4, Appendix B.7). The state model therefore retains covariance terms. Second, the encoder must preserve the information the evidence depends on. Write $\tilde{Z} = (Z, \xi_{1:M})$ and call the coordinates observed by exactly the modalities in a the block of a.

Proposition 2 (Evidence-preserving representations)**.** *Fix the acquisition context. Suppose the view maps are diffeomorphisms onto their images, every coordinate of Z is observed by at least one modality, and $\tilde{Z}$ has a smooth positive density on an open, simply connected product domain. (i) Under the known-dimension and population optimization conditions of Yao et al. (2024, Thm. 3.2), the content shared by a subset of views is recoverable up to a diffeomorphism. View geometry alone permits an exact-support block to be reparameterized conditionally on blocks observed by strict supersets. (ii) A common diffeomorphism F of Z preserves T when the conditioning statistic is transported as $U \circ F^{-1}$. If an encoder permits exact reconstruction of all modalities, then T and q_π^{ref} are functions of its output. Retaining only content shared by all modalities can lose it.*

Shared content for a subset includes the blocks shared by larger subsets. This motivates organizing the representation from broadly shared to partially shared and private information. Reconstruction supplies a sufficient condition for preserving evidence that alignment alone may omit. The population statement does not imply that finite reconstruction error, or a particular residual objective, identifies the exact blocks. Moreover, preserving evidence does not ensure that it is expressed in the Gaussian coordinates needed by our estimator (Appendix B.8).

3.4 From disease evidence to a recorded endpoint

The reference posterior excludes the target's selection gate $g_s(u) = P_e(S = s \mid U = u)$ and generally differs from its Bayes posterior; for a target satisfying (A4), the excess state log-loss satisfies

$$R_e^S(q_\pi^{\mathrm{ref}}) - R_e^S(p_e) \le \mathbb{E}_U \,\mathrm{KL}\big(g(U) \,\|\, \pi\big) = I_e(S; U) + \mathrm{KL}\big(P_e(S) \,\|\, \pi\big). \tag{6}$$

The bound vanishes when S and U are independent and π matches target prevalence; Appendix B.5 gives the exact discrepancy, which need not increase monotonically with an enrollment parameter. For observed inputs, assume $Z = \Phi(X)$ in the fixed acquisition context and write $q_\pi^{X,\mathrm{ref}}(x) = q_\pi^{\mathrm{ref}}(\cdot \mid \Phi(x))$, $p_e^X(x) = P_e(S = \cdot \mid X = x)$ and $\eta_e(x) = P_e(Y = 1 \mid X = x)$.

Proposition 3 (State-to-task decomposition)**.** *Suppose, pointwise in x, that $|\eta_e - h(p_e^X, r_e)| \le d_e$ for continuous task information r_e and a map h shared across cohorts and (L_S, L_R)-Lipschitz in its two arguments. Let $\hat{f} = \hat{h}(\hat{q}, \hat{r})$ be a head with approximation error $\epsilon_h = |\hat{h}(\hat{q}, \hat{r}) - h(\hat{q}, \hat{r})|$. Then*

$$|\hat{f} - \eta_e| \le A_e := \epsilon_h + L_S \|\hat{q} - q_\pi^{X,\mathrm{ref}}\|_1 + L_S \|q_\pi^{X,\mathrm{ref}} - p_e^X\|_1 + L_R \|\hat{r} - r_e\| + d_e,$$

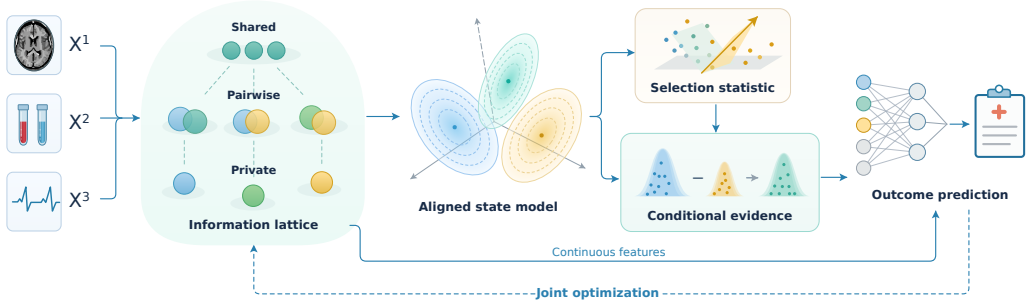


Figure 2: **Joint learning of conditional disease evidence.** An information-lattice encoder splits the modalities into shared, pairwise, and private blocks; a source-cohort state model supplies the natural-parameter shifts that define the selection statistic; conditional evidence (joint minus marginal log-likelihood ratio) and continuous features predict the outcome; all parts are trained jointly.

and the excess binary Brier risk over η_e is at most $\mathbb{E}_e[A_e(X)^2]$.

The terms separate head approximation, estimation of the reference posterior, its discrepancy from the target posterior (linked to Equation (6) by Pinsker's inequality; Appendix B.6), continuous task information, and residual endpoint mismatch, including label-protocol drift. Better state evidence directly reduces the second term but need not remove the others, so large state gains can coexist with small outcome gains; this motivates the continuous path to the outcome in Section 4, where the decomposition describes a coupled approximation problem rather than separately trained stages.

4 DISCERN: Jointly Learning Conditional Disease Evidence

DISCERN (DISease-state Content idEntification for Robust multimodal learNing) combines multimodal representation learning, latent-state modeling, and outcome prediction in a jointly optimized framework (Figure 2). Section 3 specifies the evidence to estimate; the learning objective couples this estimate to a representation that preserves multimodal information and supports the recorded endpoint. Disease-state labels are never required: recorded outcomes supervise the prediction head and, through joint optimization, shape the representation and state model.

Representing overlapping multimodal information. Let E_ϕ be an information-lattice encoder with output $\hat{z} = E_\phi(X)$. Each block $\hat{z}_a$ corresponds to information available to the modalities indexed by $a \subseteq [M]$. For three modalities, the lattice includes one shared block, three pairwise blocks, and three private blocks. Blocks are extracted from coarse to fine: the shared block is estimated first, pairwise blocks explain the residuals $\delta_m = \bar{x}_m - G_m^{\mathrm{sh}}(\hat{z}_{\mathrm{sh}})$, and private blocks represent the remaining variation. The information objective is

$$\mathcal{L}_{\mathrm{info}} = \sum_{|a|\geq 2} \mathcal{L}_a^{\mathrm{align}} + \sum_m \|\bar{x}_m - G_m(\hat{z}_{a\ni m})\|^2 + \sum_m \left[\|\delta_m\|^2 + \left\|\delta_m - G_m^{\mathrm{pw}}(\hat{z}_{|a|=2,\, a\ni m})\right\|^2\right] + \lambda_{\mathrm{var}}\mathcal{L}_{\mathrm{var}}. \tag{7}$$

Alignment encourages consistent estimates of shared blocks within each source cohort, reconstruction encourages retention of information beyond those blocks, and the residual terms organize the partially shared information. The variance penalty discourages collapse. Reconstruction is motivated by Proposition 2, with the caveats noted there; it stays active during joint training.

Modeling latent states across source cohorts. We use a Gaussian state model

$$p_\psi(\hat{z} \mid e) = \sum_{s=0}^{K-1} \pi_{es}\, \mathcal{N}(\hat{z}; \mu_{es}, \Sigma_s), \tag{8}$$

where cohort-specific means and mixture weights accommodate selection, while the covariance Σ_s of each state is shared across cohorts. In the clinical implementation, a block-ordered Gaussian

structural equation model parameterizes the covariances through state-dependent slopes. EM on source features initializes the state model, with consistent state correspondence across cohorts motivated by Proposition 1; the model then enters joint optimization through the feature-density regularizer

$$\mathcal{L}_{\text{dens}} = -\mathbb{E}_{(X,e)\sim\mathcal{D}_{\text{src}}} \log p_\psi(E_\phi(X) \mid e). \tag{9}$$

It encourages coordinates compatible with the state model; lacking a raw-input change-of-variables term, it is a feature regularizer rather than a normalized generative likelihood of the modalities.

Estimating the selection statistic and conditional evidence. At the current state-model parameters, let $\bar{\mu}_s = \sum_e w_e \mu_{es}$, where w_e are normalized source weights. We form the natural-parameter shift matrix with columns $\Sigma_s^{-1}(\mu_{es} - \bar{\mu}_s)$ and obtain a basis Γ from its leading singular directions. The experiments retain one direction. The resulting statistic $U = \Gamma^\top \hat{z}$ estimates the source selection variation characterized by Theorem 1. Using the reference family $\mathcal{N}(\bar{\mu}_s, \Sigma_s)$, define $\hat{\ell}_s(A) = \log \hat{p}(A \mid s) - \log \hat{p}(A \mid 0)$. The conditional evidence and reference posterior are

$$\begin{aligned} \hat{T}_s(\hat{z}) &= \hat{\ell}_s(\hat{z}) - \hat{\ell}_s(\Gamma^\top \hat{z}), \\ \hat{q}_\pi(s \mid \hat{z}) &= \text{softmax}_s\Big(\log \pi_s + \hat{T}_s(\hat{z})\Big), \end{aligned} \tag{10}$$

where π is a fixed positive reference prior. Subtracting the marginal evidence removes the contribution associated with the estimated selection statistic while retaining the conditional relations among physiological quantities, including state-dependent interactions through the covariance terms. The basis Γ is recomputed once per epoch outside gradient tracking and held fixed in between.

Joint optimization with the recorded outcome. The outcome predictor combines the conditional state log odds with continuous features $r_\phi(X)$:

$$\hat{p}_{\phi,\psi,\theta}(Y = 1 \mid X) = h_\theta\left(r_\phi(X), \left\{\log \frac{\hat{q}_\pi(s \mid E_\phi(X))}{\hat{q}_\pi(0 \mid E_\phi(X))}\right\}_{s=1}^{K-1}\right). \tag{11}$$

The continuous path retains task information beyond the discrete state, as motivated by Proposition 3; in the clinical model it also includes the context branch described in Appendix D. With binary cross-entropy denoted by $\mathcal{L}_{\text{task}}$, training minimizes

$$\min_{\phi,\psi,\theta} \quad \mathcal{L}_{\text{task}} + \lambda_{\text{info}}\mathcal{L}_{\text{info}} + \lambda_{\text{dens}}\mathcal{L}_{\text{dens}}, \tag{12}$$

including the reconstruction decoders among the representation parameters. Outcome gradients pass through both the continuous and conditional-evidence paths, coupling the encoder, state model, and head. The information and density terms act as regularizers ($\lambda_{\text{info}} = 0.01$ and $\lambda_{\text{dens}} = 0.001$ in the clinical experiments). Joint training requires no observed S, although the recorded Y informs the learned states through these gradients. Training starts from a supervised warm-up of the backbone and EM initialization of the state model, with model selection on source data only (Appendix D); at deployment all components are frozen and use only the target patient's inputs.

5 Experiments

The experiments follow the identification chain: the generator's own coordinates, where Equation (2) holds and S is hidden during fitting but available for evaluation (Theorem 1); learned multimodal features in their place (Proposition 2); and recorded clinical endpoints, where no state is observed (Proposition 3). The first two isolate the conditional-evidence estimator with the representation supplied or held fixed; only the clinical study runs the full joint training of Section 4. Hyperparameters and checkpoints are selected on source data throughout. Some clinical targets were inspected in earlier development rounds, so the clinical comparisons are exploratory.

Benchmark. One structural model with three states and ten latent factors generates three partially overlapping modalities through leaky-affine views with private variation. Twelve source cohorts (2,000 training and 400 validation observations each) differ in how strongly enrollment ties the root background factors to the state, as in Figure 1a, and in label style. Four target laws keep, remove,

Table 1: **State recovery across enrollment laws** (state AUROC on common coordinates, mean±SD over five generated studies). Targets keep, remove, reverse or strengthen the training enrollment. Bold: better of the two deployable estimators; the reference oracle reads the true mechanism.

Regime	State estimator	same	off	reversed	stronger
additive	uncorrected latent-state model	$0.839_{\pm.318}$	$0.771_{\pm.042}$	$0.043_{\pm.048}$	$\mathbf{0.995}_{\pm.007}$
additive	**DISCERN conditional evidence**	$\mathbf{0.884}_{\pm.048}$	$\mathbf{0.880}_{\pm.050}$	$\mathbf{0.883}_{\pm.053}$	$0.880_{\pm.049}$
additive	reference oracle	$0.882_{\pm.050}$	$0.882_{\pm.050}$	$0.882_{\pm.050}$	$0.882_{\pm.050}$
with interactions	uncorrected latent-state model	$\mathbf{0.939}_{\pm.089}$	$0.769_{\pm.073}$	$0.265_{\pm.139}$	$\mathbf{0.984}_{\pm.012}$
with interactions	**DISCERN conditional evidence**	$0.869_{\pm.056}$	$\mathbf{0.849}_{\pm.047}$	$\mathbf{0.825}_{\pm.067}$	$0.870_{\pm.059}$
with interactions	reference oracle	$0.854_{\pm.049}$	$0.855_{\pm.046}$	$0.856_{\pm.045}$	$0.853_{\pm.050}$

reverse or strengthen the training enrollment, and the generator also changes label style and site, so target cohorts differ in both selection and measurement, with enrollment on the main axis. Two mechanism regimes, additive and with interactions, each have five generated studies with 10,000 test observations per target law. The recorded endpoint Y also depends on continuous factors and on cohort-specific label terms, the $c_Y(e)$ of Equation (1). State AUROC averages the binary AUROCs for $S > 0$ and $S > 1$ of the posterior mean severity; source S fixes one global scoring permutation.

5.1 State recovery across enrollment laws

Table 1 and Figure 1b test Theorem 1 on common coordinates, comparing the uncorrected state model, the conditional evidence of DISCERN (three aligned fits averaged) and a reference oracle that reads the true mechanism with a fixed population prior. Reversal exposes the shortcut: the uncorrected model falls to 0.043 and 0.265 state AUROC, and conditioning raises it to 0.883 and 0.825. The conditional evidence is flat across the four laws and tracks the reference oracle, with the largest gap (0.031) under reversal with interactions. The fourth law shows the price in Equation (6): a target that selects strongly makes selection itself evidence, which the uncorrected readout exploits (0.995, 0.984) and the conditional evidence removes (0.880, 0.870).

Attribution. A gain under reversal could come from any change to the state model. Five further studies per regime read one fitted mixture in three ways: uncorrected, conditioned on a random unit direction, and conditioned on the estimated direction (Figure 4, Appendix C). Under reversal the estimated direction reaches 0.860 and 0.861 state AUROC, against 0.151 and 0.270 for a random direction and 0.115 and 0.252 uncorrected, improving over both in every study. The correction is attributable to the selection statistic of Theorem 1, not to conditioning as such.

5.2 From state evidence to the recorded endpoint

By Proposition 3, large state gains need not carry over to Y. On common coordinates under reversal, the head reading continuous Z with the averaged conditional evidence reaches 0.780 and 0.812 outcome AUROC against 0.779 and 0.810 for matched ERM, and a head that reads only the state falls to 0.697 and 0.744 (Table 3). Physiology carries task information that no state does.

Learning the coordinates is the harder step. On three paired studies a label-free 13-dimensional lattice replaces the supplied Z, and state recovery fails: under reversal its state AUROC is 0.118 and 0.306, below chance like the uncorrected model on common coordinates, and an invertible calibration fitted jointly with the mixture (Appendix C.5) leaves it at 0.135 and 0.307 although it improves source feature NLL from 5.24 to 4.23 and from 5.93 to 4.94. The outcome head matches but does not improve on the matched lattice ERM head, 0.850 and 0.797 against 0.863 and 0.806 AUROC, within one SD (Table 4). The information is not gone: diagnostic probes trained on separate target-labeled samples read S from the frozen lattice at 0.868 and 0.832 under the selection-off law, close to 0.903 and 0.859 on the true Z. Consistent with Proposition 2, the reconstruction-trained lattice keeps the evidence but not necessarily in coordinates satisfying Equation (2); target views also change site, which the proposition holds fixed, so the two causes are confounded.

Table 2: **Cross-database ICU mortality** (E→M: eICU to MIMIC-IV; M→E: reverse). AUROC and AUPRC average physical hospitals; NLL pools target patients; mean±SD over three seeds. †: source-trained baselines with a different backbone; matched rich ERM shares the DISCERN backbone and inputs.

Method	AUROC E→M	AUROC M→E	AUPRC E→M	AUPRC M→E	NLL E→M	NLL M→E
ERM early fusion†	$0.808_{\pm.002}$	$0.783_{\pm.015}$	$0.467_{\pm.012}$	$0.394_{\pm.027}$	$0.342_{\pm.009}$	$0.373_{\pm.057}$
ERM late fusion†	$0.793_{\pm.011}$	$0.792_{\pm.001}$	$0.473_{\pm.018}$	$0.411_{\pm.004}$	$0.351_{\pm.029}$	$0.347_{\pm.042}$
V-REx†	$0.804_{\pm.009}$	$0.778_{\pm.015}$	$0.461_{\pm.010}$	$0.388_{\pm.032}$	$0.340_{\pm.020}$	$0.368_{\pm.032}$
GroupDRO†	$0.810_{\pm.002}$	$0.778_{\pm.015}$	$0.468_{\pm.014}$	$0.389_{\pm.030}$	$0.332_{\pm.013}$	$0.378_{\pm.037}$
CORAL†	$0.805_{\pm.003}$	$0.774_{\pm.017}$	$0.462_{\pm.015}$	$0.383_{\pm.030}$	$0.332_{\pm.013}$	$0.362_{\pm.034}$
SWAD-style†	$\mathbf{0.820}_{\pm.001}$	$0.790_{\pm.013}$	$0.485_{\pm.008}$	$0.405_{\pm.023}$	$0.323_{\pm.008}$	$0.349_{\pm.045}$
Matched rich ERM	$0.782_{\pm.017}$	$0.777_{\pm.011}$	$0.462_{\pm.022}$	$0.398_{\pm.012}$	$0.383_{\pm.030}$	$0.372_{\pm.024}$
DISCERN	$0.818_{\pm.018}$	$\mathbf{0.812}_{\pm.016}$	$\mathbf{0.502}_{\pm.020}$	$\mathbf{0.431}_{\pm.016}$	$\mathbf{0.316}_{\pm.016}$	$\mathbf{0.328}_{\pm.028}$

5.3 Cross-database ICU mortality

We evaluate hospital mortality between eICU (Pollard et al., 2018) and MIMIC-IV (Johnson et al., 2023) in both directions, with a 48-hour observation window, a 6-hour gap and fixed patient-disjoint partitions (Appendix D). The databases differ in selection (one academic centre against 57 hospitals) and in measurement (charting and laboratory practice), so the clinical variant departs from Section 2: acquisition is not fixed, and the outcome-fine-tuned encoder makes its states label-informed. All rich variants share a 38-dimensional physiology lattice, a context branch for the remaining inputs, the same supervised warm-up and source-only checkpoint selection, so the matched ERM isolates the mechanism branch; historical baselines share the targets but not the backbone.

Table 2 shows that DISCERN reaches 0.818 and 0.812 AUROC, 0.036 and 0.035 above the matched ERM, with the best AUPRC and NLL in both directions; NLL, a proper score, falls from 0.383 and 0.372 for the matched ERM to 0.316 and 0.328, so the gain is not confined to ranking. In AUROC, DISCERN is the best method into eICU, 0.020 above the strongest historical baseline, and 0.002 below a weight-averaging recipe on a different backbone into MIMIC. Among the matched ablations (Table 6, Appendix D), information learning alone accounts for 0.006 of the gain and the conditional branch for the rest, while the uncorrected and no-edge controls stay at the information-only level. The gain holds for every assumed number of latent factors, while synthetic state recovery peaks at the true one (Appendix C.6; exploratory ADNI study: Appendix E).

6 Conclusion

This paper shows that enrollment shortcuts in medical multimodal data can be separated from disease evidence in a principled way, without observing the disease state. From a generative account in which cohorts differ in selection and measurement but share disease mechanisms, we showed that unobserved states can be aligned across cohorts, that enrollment-induced natural-parameter changes identify a minimal selection statistic, and that conditioning on it yields disease evidence identifiable without state labels and invariant to enrollment shifts within the identified family. The same analysis specifies what an estimator must preserve, namely cross-modal dependence and information not shared by all modalities, and explains why state evidence and endpoint prediction are related but distinct goals. DISCERN turns these results into one jointly trained model, and experiments confirm the predictions: under reversed enrollment, conditioning on the estimated statistic lifts state recovery from below chance to near the oracle, unlike a random direction, and on cross-database ICU mortality DISCERN improves AUROC, AUPRC and NLL over matched ERM in both directions.

The analysis also shows where the next gains lie. A reconstruction-trained lattice already retains the state information but does not yet express it in the common Gaussian coordinates our identification results assume; representations that do would extend the guarantees end to end. Selection statistics beyond rank one, a clinical anchor for the learned states, and more seeds on unseen targets would further strengthen the clinical findings, which remain exploratory (Appendix **??**). More broadly, separating what enrollment contributes from what the disease contributes offers a principled route to multimodal models whose disease evidence remains valid when enrollment changes.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets. The authors designed the structural causal model of Figure 1a, including the latent states, background factors, mechanisms, modality views, and source and target enrollment laws; OpenAI Codex implemented, from this specification, the code that generates the synthetic data and runs the synthetic experiments of Section 5 and Appendix C. The remaining required-disclosure tasks (translation, dataset cleaning, and qualitative data analysis) are not applicable to this work. Additionally, we used Claude (Anthropic) and ChatGPT (OpenAI) to edit the paper for readability and length, and to check the manuscript for consistency among text, tables, cross-references, and notation. We have reviewed all AI-assisted work: the authors checked the generated code against the specified model and target laws, and checked every edited passage and every reported number against the derivations and experimental results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with generative AI.

REPRODUCIBILITY STATEMENT

Assumptions are stated in Section 2 and proofs are given in Appendix B. The synthetic generator, target laws, seeds and sample budgets are in Appendix C; the clinical cohort definitions, partitions and training protocols are in Appendices D and E. The code is available at: https://anonymous.4open.science/r/DISCERN-AF7F.

REFERENCES

Julian N. Acosta, Guido J. Falcone, Pranav Rajpurkar, and Eric J. Topol. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022.

Jerome Allyn, Matthieu Oliver, Raphaël Cerveau, Nicolas Allou, Tristan Barennes, and Cyril Ferdynus. Multi-source data integration through pooling and transfer learning improves generalizability and specialization of deep learning models for ICU mortality and length of stay prediction: a four-database external validation study. *Critical Care*, 30:364, 2026. doi: 10.1186/s13054-026-06034-5.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019.

Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. Detecting shortcut learning for fair medical AI using shortcut testing. *Nature Communications*, 14:4314, 2023.

Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11:3673, 2020.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Chong Lam Cheong. Research on AI security strategies and practical approaches for risk management. *Journal of Computer, Signal, and System Research*, 2(7):98–115, 2025. doi: 10.71222/17gqja14.

Chong Lam Cheong. Study on risk assessment methods and multi-dimensional control mechanisms in AI systems. *European Journal of AI, Computing & Informatics*, 2(1):31–46, 2026. doi: 10.71222/58dr7v22.

Haoyue Dai, Yiwen Qiu, Ignavier Ng, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Latent variable causal discovery under selection bias. In *Proceedings of the 42nd International Conference on*

Machine Learning, volume 267 of *Proceedings of Machine Learning Research*, pp. 12161–12178. PMLR, 2025.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, 2025.

Shaza Elsharief, Saeed Shurrab, Baraa Al Jorf, Leopoldo Julian Lechuga Lopez, Krzysztof J. Geras, and Farah E. Shamout. MedMod: Multimodal benchmark for medical prediction tasks with electronic health records and chest X-ray scans. In *Proceedings of the Sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pp. 781–803. PMLR, 2025.

Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.

Jean-Philippe Fortin, Nicholas Cullen, Yvette I. Sheline, Warren D. Taylor, Irem Aselcioglu, Philip A. Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J. McGrath, Melvin McInnis, Mary L. Phillips, Madhukar H. Trivedi, Myrna M. Weissman, and Russell T. Shinohara. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

Paul Hager, Martin J. Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23924–23935, 2023.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 479–503, 2022.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110:303–348, 2021.

Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3:136, 2020.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In *Advances in Neural Information Processing Systems*, volume 34, pp. 10944–10956, 2021.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 859–868, 2019.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Clifford R. Jack, Jr., Heather J. Wiste, Stephen D. Weigand, David S. Knopman, Prashanthi Vemuri, Michelle M. Mielke, Val Lowe, Matthew L. Senjem, Jeffrey L. Gunter, Mary M. Machulda, Brian E. Gregg, V. Shane Pankratz, Walter A. Rocca, and Ronald C. Petersen. Age, sex, and APOE $\varepsilon 4$ effects on memory, brain structure, and β-amyloid across the adult life span. *JAMA Neurology*, 72(5):511–519, 2015. doi: 10.1001/jamaneurol.2014.4821.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023.

W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Pritish Kamath, Akilesh Tangella, Danica J. Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 4069–4077, 2021.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2207–2217, 2020.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations (ICLR)*, 2023.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5:171, 2022.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 5637–5664, 2021.

Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 11455–11472, 2022.

Konstantinos Kontras, Thomas Strypsteen, Christos Chatzichristos, Paul Pu Liang, Matthew B. Blaschko, and Maarten De Vos. Balancing multimodal training through game-theoretic regularization. In *Advances in Neural Information Processing Systems*, 2025.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 5815–5826, 2021.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie E. Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Proceedings of the First Conference on Causal Learning and Reasoning (CLeaR)*, pp. 428–484, 2022.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10): 1–42, 2024.

Zihan Liang, Ziwen Pan, and Ruoxuan Xiong. Causal representation learning from multimodal clinical records under non-random modality missingness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28791–28808. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1465.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 6781–6792, 2021.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 4114–4124, 2019.

Malte Londschien, Manuel Burger, Gunnar Rätsch, and Peter Bühlmann. Domain generalization and adaptation in intensive care with anchor regression. *RSS: Data Science and Artificial Intelligence*, 2(1):udag001, 2026. doi: 10.1093/rssdat/udag001.

Zhou Lu. A theory of multimodal learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18177–18186, 2022.

Bret Nestor, Matthew B. A. McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C. Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks. In *Proceedings of the 4th Machine Learning for Healthcare Conference (MLHC)*, pp. 381–405, 2019.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL)*, pp. 151–159, 2020.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8238–8247, 2022.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018.

Aahlad Puli, Lily H. Zhang, Eric K. Oermann, and Rajesh Ranganath. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. In *International Conference on Learning Representations*, 2022.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B*, 83 (2):215–246, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Luis R. Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M. Wiberg, Michael L. Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 5:149, 2022.

Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, 2022.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2):345–352, 2020.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3118–3127, 2019.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision (ECCV) Workshops*, pp. 443–450, 2016.

Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, Eran Segal, Eric P. Xing, and Kun Zhang. Causal representation learning from multimodal biomedical observations. In *International Conference on Learning Representations*, 2025.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59723–59760. PMLR, 2025.

Sana Tonekaboni, Sam Freesun Friedman, Xinyi Zhang, Mahnaz Maddah, and Caroline Uhler. A representation fusion framework for decoupling diagnostic information in multimodal learning. *npj Digital Medicine*, 8:765, 2025. doi: 10.1038/s41746-025-02144-6.

Fanny Tranchellini, Youssef Farag, Catherine Jutzeler, and Lakmal Meegahapola. Evaluating deep learning sepsis prediction models in ICUs under distribution shift: a multi-centre retrospective cohort study. *npj Digital Medicine*, 9:306, 2026. doi: 10.1038/s41746-026-02364-4.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable domain generalization via invariant-feature subspace recovery. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23018–23033, 2022.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705, 2020.

Julia K. Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Holger A. Haenssle. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology*, 155(10):1135–1141, 2019.

Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8):1065–1070, 2021.

Sidney J. Yakowitz and John D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations (ICLR)*, 2024.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, 2025.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 25407–25437, 2022.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, 2018.

Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yiwen Song, Cai Xu, Jingjing Zhou, Aparna Joshi, Shravan Dheep, Mani Malek, Hamid Palangi, Joon Baek, Rick Pereira, and Karthik Narasimhan. ShieldGemma 2: Robust and tractable image content moderation. *arXiv preprint arXiv:2504.01081*, 2025. Model and Transformers implementation: `https://huggingface.co/docs/transformers/en/model_doc/shieldgemma2`.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3Care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2418–2428, 2022.

Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pp. 279–290, 2021.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 12979–12990, 2021.

A EXACT EXAMPLES OF ENROLLMENT SHORTCUTS AND JOINT EVIDENCE

These examples give exact population errors for specified mechanisms (Figure 3). They illustrate Sections 2 and 3; they are not fitted DISCERN or ERM performance.

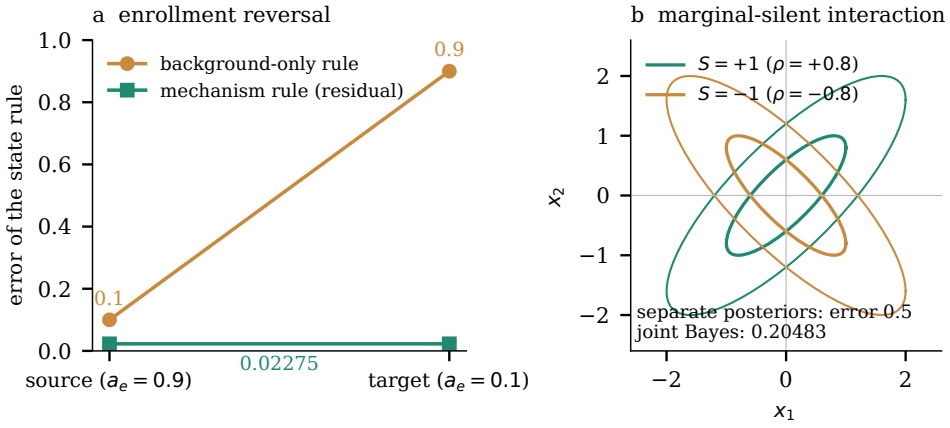


Figure 3: **Analytical illustrations.** (a) An enrollment reversal changes the error of a background-only rule from 0.1 to 0.9, while a rule that uses the specified mechanism through the residual has error 0.02275 in both cohorts. (b) Contours at Mahalanobis radii one and two for two Gaussian states with correlations $+0.8$ and -0.8. Each single-view posterior is uninformative, whereas the joint interaction yields error 0.20483. Contours are analytical densities; no error bars apply.

Enrollment-induced shortcut. This is the selection of Figure 1a with one background factor. Let population signs $S, B \in \{-1, +1\}$ be independent and balanced, and let the inclusion indicator R satisfy

$$P(R = 1 \mid S, B, e) = \begin{cases} a_e, & B = S, \\ 1 - a_e, & B \neq S. \end{cases} \tag{13}$$

Then $P(R = 1) = 1/2$, $P(S = +1 \mid R = 1, e) = 1/2$ and $P(B = S \mid R = 1, e) = a_e$. The rule $h_B = B$ has selected-cohort error $1 - a_e$: 0.1 for the source law $a_e = 0.9$ and 0.9 for the target law $a_e = 0.1$. Let $D = S + B + \epsilon$ with $\epsilon \sim \mathcal{N}(0, 0.5^2)$ independent of (S, B) and of the inclusion noise, inclusion depending only on (S, B, e). With the parents and the mechanism known,

$$h_{\text{mech}} = \text{sign}(D - B), \qquad P(h_{\text{mech}} \neq S \mid R = 1, e) = \Phi(-2) \simeq 0.02275. \tag{14}$$

Conditioning on B preserves the disease mechanism while removing the selected-cohort association from its evidence, as in the conditional sub-model evidence of Appendix B. The mechanism and parents are given here; this discrete example lies outside the Gaussian alignment theorem, and target selection can itself carry useful evidence, which is the price in Equation (6).

Interaction invisible to separate posteriors. In a separate two-view model take balanced S, independent standard Gaussian noises ϵ_1, ϵ_2, and

$$X_1 = D_1 = \epsilon_1, \qquad X_2 = D_2 = \rho S D_1 + \sqrt{1 - \rho^2}\, \epsilon_2, \qquad \rho = 0.8. \tag{15}$$

Both single-view state-conditionals are standard normal, so each single-view Bayes posterior equals $1/2$, and any rule that uses only those two posterior values has error $1/2$. In contrast,

$$\log \frac{p(x_1, x_2 \mid S = +1)}{p(x_1, x_2 \mid S = -1)} = \frac{2\rho x_1 x_2}{1 - \rho^2}, \qquad R^*_{\text{joint}} = \frac{\arccos \rho}{\pi} \simeq 0.20483. \tag{16}$$

The Bayes rule is $\text{sign}(X_1 X_2)$. The loss concerns separate-posterior compression, not all late fusion; in Figure 1a the laboratory modality $X^{(2)}$ observes both D_1 and D_2, so this particular bottleneck does not apply there.

B FULL STATEMENTS AND PROOFS

Unless stated otherwise, Equation (2) holds in common coordinates $z \in \mathbb{R}^d$, with positive component weights, distinct components within each cohort, and positive-definite state-specific covariances Σ_s shared across cohorts. We use Γ for a basis of the selection subspace, reserving B for background variables. All identification statements concern population laws. No claim of finite-sample recovery or convergence of the joint training objective is needed for these results.

B.1 FINITE MIXTURES AND PROPOSITION 1

Proposition 1 (full statement). *The components of each source mixture are identified up to a permutation. If the shared covariances are pairwise distinct, the components can be matched across all source cohorts up to one global permutation. For the three-state common-covariance family in Equation (4), linear independence of a and ν and three distinct source values of κ_e identify the central state and match the end states up to their global exchange.*

Proof. We recall the finite-Gaussian-mixture argument (Yakowitz & Spragins, 1968). If two minimal mixtures agree, combining identical components gives $\sum_{j=1}^{J} \alpha_j \mathcal{N}(z; \mu_j, \Sigma_j) = 0$ for distinct parameter pairs. Choose v such that the pairs $(v^\top \Sigma_j v, v^\top \mu_j)$ are all distinct. Such a v exists because each pairwise failure is contained in a proper polynomial zero set. Taking moment-generating functions along v yields

$$\sum_{j=1}^{J} \alpha_j \exp\left(\tfrac{1}{2} t^2 v^\top \Sigma_j v + t v^\top \mu_j\right) = 0.$$

Order the pairs first by projected variance and then by projected mean. Divide by the exponential for the largest pair and let $t \to +\infty$. All other terms vanish, so its coefficient is zero. Repeating proves linear independence and hence identification of the unordered component parameters and weights.

When the shared covariances are distinct, a given covariance has exactly one match in each cohort. Matching by covariance leaves only one global ordering of the states unspecified.

For Equation (4), the arithmetic mean of the three unordered component means is a, regardless of their mixture weights. Linear independence makes the component at a unique. After subtracting a, the other means lie on

$$L_+ = a + \operatorname{span}(\nu), \qquad L_- = -a + \operatorname{span}(\nu).$$

These lines are distinct and parallel, and each contains at least three distinct observed points. Any line through three points in their union must contain two points from one of them, hence coincide with it; it cannot then intersect the other parallel line. The two lines and their component assignments are therefore determined, up to their exchange. Neither the values of κ_e nor a source cohort with $\kappa_e = 0$ need be known. □

B.2 PROOF OF THEOREM 1 AND ITS ESTIMATOR CORRESPONDENCE

Theorem 1 (full statement). *Under Equation (2) and an alignment condition of Proposition 1, the subspace $\mathcal{U}$ in Equation (5) is identified in the common coordinates. A linear statistic $A^\top Z$ determines every same-state source density ratio if and only if $\mathcal{U} \subseteq \operatorname{col}(A)$. Let $r = \dim \mathcal{U}$, choose a basis $\Gamma \in \mathbb{R}^{d \times r}$, and choose C so that $Z \leftrightarrow (U, V) = (\Gamma^\top Z, C^\top Z)$ is invertible. Then $p_e(v \mid u, s) = h_s(v \mid u)$ is independent of the source cohort. The evidence in Equation (3) equals $\log h_s(v \mid u) - \log h_0(v \mid u)$. The reference posterior is identified up to the global state permutation, and the evidence with the corresponding baseline. The same conditional kernels apply to any Gaussian target with covariance Σ_s in state s and $\Sigma_s^{-1}(\mu_{ts} - \mu_{e_0 s}) \in \mathcal{U}$ for every state.*

Proof. Define $\beta_{es} = \Sigma_s^{-1}(\mu_{es} - \mu_{e_0 s})$. Expanding the Gaussian quadratic forms gives

$$\frac{p_e(z \mid s)}{p_{e_0}(z \mid s)} = \exp\left(\beta_{es}^\top z - c_{es}\right), \qquad c_{es} = \tfrac{1}{2}\left(\mu_{es}^\top \Sigma_s^{-1} \mu_{es} - \mu_{e_0 s}^\top \Sigma_s^{-1} \mu_{e_0 s}\right). \tag{17}$$

Aligned mixture parameters determine these vectors and their span. Changing the reference cohort does not change the span, because every difference between two cohort means is a difference of their contrasts with e_0.

Since $\beta_{es} \in \mathcal{U}$, the right-hand side of Equation (17) is a positive function $a_{es}(u)$. The common change-of-coordinates Jacobian cancels, so

$$p_e(u, v \mid s) = a_{es}(u) p_{e_0}(u, v \mid s), \qquad p_e(u \mid s) = a_{es}(u) p_{e_0}(u \mid s).$$

The second identity follows by integration over v. Dividing gives $p_e(v \mid u, s) = p_{e_0}(v \mid u, s) = h_s(v \mid u)$. Factoring the densities in (u, v) then proves the expression for T. Equivalently, the conditional law of Z is the same on each affine conditioning fibre.

For minimality, suppose a ratio is measurable with respect to $A^\top Z$ but its vector β_{es} is outside $\mathrm{col}(A)$. After removing redundant columns of A, a nondegenerate Gaussian conditioned on $A^\top Z$ has full support on its affine fibre. There is $w \in \ker(A^\top)$ with $\beta_{es}^\top w \neq 0$, so $\beta_{es}^\top Z$ has positive conditional variance and cannot be measurable with respect to $A^\top Z$. This contradiction proves necessity. Conversely, $\mathcal{U} \subseteq \mathrm{col}(A)$ makes every expression in Equation (17) a function of $A^\top Z$, proving sufficiency. This is minimality for the specified ratio family, not identification of a uniquely defined biological nuisance.

Replacing Γ by ΓH, with invertible H, replaces U by $H^\top U$. Its state density ratios are unchanged because their common Jacobian cancels. A state permutation τ gives

$$T'_s = T_{\tau(s)} - T_{\tau(0)}, \qquad \pi'_s = \pi_{\tau(s)}, \qquad q^{\mathrm{ref}\prime}_{\pi'}(s \mid z) = q^{\mathrm{ref}}_{\pi}(\tau(s) \mid z).$$

Thus the posterior is permuted, while the evidence is also recentered at its newly chosen baseline.

For a target satisfying (A4), Equation (17) holds with $e = t$ and $\beta_{ts} \in \mathcal{U}$, giving the same cancellation. This is an assumption on the target family, not a property established for arbitrary targets from source data. When $r = 0$, U is constant and no correction is required. When $r = d$, V is empty, $T = 0$, and $q^{\mathrm{ref}}_{\pi} = \pi$. □

Why the centred estimator implements the same definition. Section 4 uses $\bar{\mu}_s = \sum_e w_e \mu_{es}$ with $w_e \geq 0$ and $\sum_e w_e = 1$. This does not change the population subspace:

$$\mathrm{span}\{\Sigma_s^{-1}(\mu_{es} - \bar{\mu}_s) : e, s\} = \mathrm{span}\{\Sigma_s^{-1}(\mu_{es} - \mu_{e_0 s}) : e, s\} = \mathcal{U}. \tag{18}$$

Each centred contrast is a linear combination of reference-cohort contrasts, and subtracting the centred contrast for e_0 gives the converse. Moreover, $\Sigma_s^{-1}(\bar{\mu}_s - \mu_{e_0 s}) \in \mathcal{U}$, so the Gaussian reference family $\mathcal{N}(\bar{\mu}_s, \Sigma_s)$ has the same conditional kernels h_s. It is a single Gaussian for each state, not the pooled mixture of that state's source densities. For $0 < r \leq d$, its induced marginal on U gives

$$\begin{aligned} T_s(z) = \; & \log \frac{\mathcal{N}_d(z; \bar{\mu}_s, \Sigma_s)}{\mathcal{N}_d(z; \bar{\mu}_0, \Sigma_0)} \\ & - \log \frac{\mathcal{N}_r(u; \Gamma^\top \bar{\mu}_s, \Gamma^\top \Sigma_s \Gamma)}{\mathcal{N}_r(u; \Gamma^\top \bar{\mu}_0, \Gamma^\top \Sigma_0 \Gamma)}, \qquad u = \Gamma^\top z. \end{aligned} \tag{19}$$

The covariance in each marginal is $\Gamma^\top \Sigma_s \Gamma$; it is positive definite for full-column-rank Γ. The joint and marginal terms must come from the same state-conditional reference family. Equation (10) is this construction with learned parameters and coordinates. Exact invariance requires the full subspace; truncating it to a smaller rank generally leaves some selection dependence. Periodically refreshing the basis during joint training changes the estimate, not the population identity.

Conditional sub-model evidence. If background coordinates B are available, remove linear redundancies from (B, U) to obtain W, and choose V_D so that $Z \leftrightarrow (W, V_D)$ is invertible. Since W determines U, the same cancellation gives the cohort-invariant ratio

$$T_s^{V_D \mid W} = \log \frac{p_e(V_D \mid W, s)}{p_e(V_D \mid W, 0)} = \log \frac{p_e(V_D, W \mid s)}{p_e(V_D, W \mid 0)} - \log \frac{p_e(W \mid s)}{p_e(W \mid 0)}. \tag{20}$$

Conditioning on the larger W can remove additional evidence; this quantity need not equal the original T_s unless the added coordinates contribute no further conditional state evidence. For a known acyclic graph, state-independent factors cancel from the joint ratio, but the marginal normalizer must still be retained. This identity does not identify the graph. In $D = aS + \gamma B + \epsilon$, for example, the conditional mechanism is expressed through $D - \gamma B$, so interpreting D requires retaining B as a covariate.

B.3 Connection to the structural model and the rank-one setting

Let $M_s = I - A_s$ and consider the linear Gaussian instance

$$Z = M_s^{-1}(o_s + \kappa_e \ell_s r_0 + \epsilon_s), \qquad \epsilon_s \sim \mathcal{N}(0, \Omega_s),$$

where M_s is invertible, Ω_s is positive diagonal, r_0 is a fixed enrollment direction, and only κ_e varies across cohorts. Here r_0 denotes a vector, distinct from the rank $r = \dim \mathcal{U}$. Then $\mu_{es} = M_s^{-1}(o_s + \kappa_e \ell_s r_0)$ and $\Sigma_s = M_s^{-1}\Omega_s M_s^{-\top}$, giving

$$\Sigma_s^{-1}(\mu_{es} - \mu_{e_0 s}) = (\kappa_e - \kappa_{e_0})\ell_s M_s^{\top}\Omega_s^{-1} r_0. \tag{21}$$

Provided the source values of κ_e vary, the exact subspace is

$$\mathcal{U} = \operatorname{span}\left\{\ell_s M_s^{\top}\Omega_s^{-1} r_0 : s = 0, \ldots, K-1\right\}. \tag{22}$$

A single enrollment parameter therefore yields rank one only when its nonzero state-specific natural-parameter directions are collinear. State-dependent propagation and innovation variances can otherwise produce several directions.

One sufficient structural condition is that r_0 is supported on root background coordinates. Their rows in A_s are zero; because Ω_s is diagonal, $A_s^{\top}\Omega_s^{-1} r_0 = 0$, so the direction reduces to $\Omega_s^{-1} r_0$. If the innovation variances are equal over the selected roots within each state, this vector is proportional to r_0. With at least one nonzero ℓ_s and a nonzero source contrast, $\mathcal{U} = \operatorname{span}(r_0)$. The variances may differ between states. Selection on a single root coordinate is a special case.

More generally, suppose enrollment changes root-background means, all such natural contrasts remain in the background coordinate subspace, and the source contrasts span that entire subspace. Then U is an invertible linear function of B and

$$T_s(z) = \log \frac{p(d \mid b, s)}{p(d \mid b, 0)}.$$

A sufficient excitation condition is that, in at least one state, the root-background mean contrasts span all background coordinates; the relevant diagonal covariance scaling is invertible. Without full coverage, U need not recover all of B, and the interpretation as the entire conditional physiological mechanism is not justified. The theorem still applies to its identified subspace and targets satisfying (A4).

B.4 Common transformations and nonlinear limits

For a common affine change $Z' = LZ + \tau$ with invertible L,

$$\mu'_{es} = L\mu_{es} + \tau, \quad \Sigma'_s = L\Sigma_s L^{\top}, \quad \beta'_{es} = L^{-\top}\beta_{es}, \quad \mathcal{U}' = L^{-\top}\mathcal{U}.$$

Choosing $\Gamma' = L^{-\top}\Gamma$ gives $U' = U + \Gamma^{\top}L^{-1}\tau$. Common Jacobians and the common translation of U cancel from the state density ratios. Distinct-covariance matching and the centred parallel-line geometry are preserved; the zero-offset parametrization in Equation (4) need not remain literal after translation. Truncated Euclidean singular vectors and a fixed numerical ridge need not have this exact affine equivariance.

For a common diffeomorphism F, the transformed density is

$$p_e^F(F(z) \mid s) = \frac{p_e(z \mid s)}{|\det J_F(z)|}, \qquad U^F = U \circ F^{-1}. \tag{23}$$

The Jacobian cancels from every state likelihood ratio, while $U^F(F(Z)) = U(Z)$ has exactly the same conditional distribution. Both terms of Equation (3), hence T and the reference posterior, are unchanged. A nonlinear F need not preserve Gaussianity or make U^F linear. Evidence preservation is therefore weaker than the coordinate condition required by the Gaussian estimator.

Likewise, an invertible calibration cannot recreate information already lost by an encoder: a measurable bijection with measurable inverse preserves the generated sigma-algebra and unrestricted Bayes risks. It may improve a restricted readout by changing the coordinates in which that readout operates. Gaussian-mixture identifiability alone does not establish identifiability of an unknown

nonlinear observation map. Without restrictions on the component family, any density $p(z)$ admits latent explanations

$$\pi_s = \int p(z) w_s(z)\, dz, \qquad p(z \mid s) = \frac{p(z) w_s(z)}{\pi_s},$$

for arbitrary positive functions w_s summing to one. These explanations have the same observed law and different state posteriors. Additional views that only invertibly re-express the same Z do not remove this ambiguity.

B.5 Reference posterior and the exact target discrepancy

Let π be a positive reference prior. For any probability law ν_U of U, define the reference population

$$P^{\mathrm{ref}}_{\nu_U}(S = s, du, dv) = \pi_s \nu_U(du) h_s(v \mid u)\, dv. \tag{24}$$

Here S and U are independent before V is drawn. Bayes' rule gives

$$q^{\mathrm{ref}}_\pi(s \mid u, v) = \frac{\pi_s h_s(v \mid u)}{\sum_j \pi_j h_j(v \mid u)},$$

independently of the chosen ν_U. This defines the reference posterior even when no observed cohort has this sampling law.

For a target e with these conditional kernels, write

$$g_s(u) = P_e(S = s \mid U = u), \qquad m_g(v \mid u) = \sum_s g_s(u) h_s(v \mid u), \qquad m_\pi(v \mid u) = \sum_s \pi_s h_s(v \mid u).$$

Define $R^S_e(q) = \mathbb{E}_e[-\log q(S \mid Z)]$ and assume the displayed expectations are finite. At fixed u, the KL divergence between the joint laws $g_s(u)h_s(v \mid u)$ and $\pi_s h_s(v \mid u)$ is $\mathrm{KL}(g(u)\|\pi)$. Factoring instead in the order V then S gives

$$\begin{aligned} \Delta^S_e &:= R^S_e(q^{\mathrm{ref}}_\pi) - R^S_e(p_e) \\ &= \mathbb{E}_{U,V}\, \mathrm{KL}\big(p_e(\cdot \mid U, V) \,\|\, q^{\mathrm{ref}}_\pi(\cdot \mid U, V)\big) \\ &= \mathbb{E}_U[\mathrm{KL}(g(U)\|\pi) - \mathrm{KL}(m_g(\cdot \mid U)\|m_\pi(\cdot \mid U))]\,. \end{aligned} \tag{25}$$

Nonnegativity of the second KL term proves Equation (6). Also,

$$\mathbb{E}_U \mathrm{KL}(g(U)\|\pi) = I_e(S; U) + \mathrm{KL}(P_e(S)\|\pi). \tag{26}$$

This separates state–selection dependence from prevalence mismatch. If $g(U) = \pi$ almost surely, both the bound and the actual excess risk vanish. Otherwise the reference posterior can differ from the target posterior. The actual gap is zero precisely when the two posteriors agree almost surely; in the positive Gaussian family this requires $g(U) = \pi$ almost surely. In particular, switching off an enrollment parameter is not by itself enough: it must also give state-independent U and the chosen reference prevalence. Neither the bound nor its decomposition proves monotonicity of the actual excess risk along an arbitrary enrollment path.

The construction extends to any positive reference gate $a_s(u)$ by replacing π_s with $a_s(u)$ in the reference posterior and the KL expressions. Comparing two such upper bounds alone does not order their actual risks.

B.6 Proof of Proposition 3 and the continuous-information limit

Definition on observed inputs. For the exact link to Equation (6), require a measurable recovery map $Z = \Phi(X)$ in the fixed acquisition context and the observation-model condition $S \perp X \mid Z$. The latter follows from the factorization in Section 2, where the measurement law depends on Z but not additionally on S. Then

$$p^X_e(X) = p_e(\cdot \mid Z), \qquad q^{X,\mathrm{ref}}_\pi(X) = q^{\mathrm{ref}}_\pi(\cdot \mid Z) \quad \text{almost surely.}$$

Without recoverability, the pointwise identification of these objects is not available; a posterior on raw inputs requires marginalizing over the unobserved coordinates.

Proof of Proposition 3. Suppress the argument x. Insert $h(\hat{q}, \hat{r})$ and $h(p_e^X, r_e)$ between $\hat{f}$ and η_e. By the definition of ϵ_h, the Lipschitz property, and the residual bound,

$$|\hat{f} - \eta_e| \le \epsilon_h + L_S \|\hat{q} - p_e^X\|_1 + L_R \|\hat{r} - r_e\| + d_e.$$

Inserting $q_\pi^{X,\mathrm{ref}}$ into the state difference and applying the triangle inequality gives A_e. For $R_e^Y(f) = \mathbb{E}_e[(f(X) - Y)^2]$, conditioning on X yields

$$R_e^Y(\hat{f}) - R_e^Y(\eta_e) = \mathbb{E}_e[(\hat{f}(X) - \eta_e(X))^2] \le \mathbb{E}_e[A_e(X)^2].$$

This uses the scalar binary Brier convention; summing squared errors over both class probabilities multiplies both sides by two. □

Under the recoverability and measurement conditions above, Pinsker's inequality and Equation (25) give the precise connection between state log-loss and the third term:

$$\begin{aligned} \mathbb{E}_e \|p_e^X(X) - q_\pi^{X,\mathrm{ref}}(X)\|_1^2 &\le 2\,\Delta_e^S \\ &\le 2\,\mathbb{E}_U \mathrm{KL}(g(U) \| \pi). \end{aligned} \tag{27}$$

The logarithms are natural. The log-loss gap and the posterior norm are distinct quantities; Equation (27) connects them under the stated conditions.

Correspondence with the jointly trained head. Equation (11) uses log odds rather than probabilities as numerical inputs. Its composite probability-space head is

$$\hat{h}(q, r) = h_\theta\left(r, \left\{\log \frac{q_s}{q_0}\right\}_{s=1}^{K-1}\right), \qquad \log \frac{\hat{q}_\pi(s \mid \hat{z})}{\hat{q}_\pi(0 \mid \hat{z})} = \log \frac{\pi_s}{\pi_0} + \hat{T}_s(\hat{z}).$$

The positive Gaussian densities and positive prior ensure these odds are defined. This substitution puts the method in the form of Proposition 3. Lipschitz continuity is assumed for the comparison map h; it is not inferred merely from implementing a neural head on log odds. The term ϵ_h measures approximation of that comparison map at the learned inputs. Joint updates can change all estimated quantities, so the decomposition does not assert that training affects only one term. The residual d_e includes label-protocol drift and any failure of (p_e^X, r_e) to describe the endpoint through a common h.

What a state-only outcome head can lose. Let $Q = q_\pi^{X,\mathrm{ref}}(X)$ be the only input to an unrestricted predictor. Since Q is measurable with respect to X,

$$\inf_f R_e^Y(f(Q)) - R_e^Y(\eta_e) = \mathbb{E}_e \operatorname{Var}(\eta_e(X) \mid Q). \tag{28}$$

The minimizer is $f(Q) = \mathbb{E}_e[\eta_e(X) \mid Q]$. Adding a deterministic representation $R = r(X)$ reduces this oracle risk by

$$\mathbb{E}_e\Big[(\mathbb{E}_e[\eta_e(X) \mid Q, R] - \mathbb{E}_e[\eta_e(X) \mid Q])^2\Big] \ge 0,$$

by orthogonality of nested conditional expectations. This is a within-target information statement; source-trained heads need not attain either oracle. It explains why retaining continuous physiology can matter even with exact reference evidence.

B.7 Interaction example and averaging

Section 3 uses the following example to show that reducing each modality to its own state posterior before fusion can discard all state evidence.

Proposition 4 (Marginal-silent interaction)**.** *Let two zero-mean, unit-variance Gaussian views have correlation $+\rho$ in one state and $-\rho$ in the other, where $0 < \rho < 1$, with equal state priors. Both single-view posteriors are constant, so every classifier based only on those posteriors has error $1/2$. The joint log-likelihood ratio, with the positive-correlation state in the numerator, is $2\rho x_1 x_2/(1-\rho^2)$, and the joint Bayes error is $\arccos(\rho)/\pi$.*

Proof of Proposition 4. Let the two zero-mean state covariances be

$$\Sigma_+ = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad \Sigma_- = \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}, \qquad 0 < \rho < 1.$$

Both single-view laws are $\mathcal{N}(0,1)$ in both states. With equal priors, each single-view posterior is therefore the constant vector $(1/2, 1/2)$, and a classifier receiving only these vectors has error $1/2$.

The determinants coincide and expanding the inverse-covariance quadratic forms gives

$$\log \frac{p_+(x_1,x_2)}{p_-(x_1,x_2)} = \frac{2\rho x_1 x_2}{1-\rho^2}.$$

The joint Bayes classifier chooses the positive-correlation state when $x_1 x_2 > 0$. Under that state, set $A = (X_1 + X_2)/\sqrt{2}$ and $B_* = (X_1 - X_2)/\sqrt{2}$. These variables are independent Gaussians with variances $1+\rho$ and $1-\rho$. For independent standard normals G_1, G_2, the error probability is

$$P_+(X_1 X_2 < 0) = P\left(\left|\frac{G_1}{G_2}\right| < \sqrt{\frac{1-\rho}{1+\rho}}\right) = \frac{2}{\pi}\arctan\sqrt{\frac{1-\rho}{1+\rho}} = \frac{\arccos\rho}{\pi}.$$

Here G_1/G_2 is standard Cauchy. Reflection of one view gives the same error under the other state, proving the stated Bayes error. The limitation concerns fusion of the marginal posterior outputs, not every classifier that receives the original views. □

For completeness, averaging probabilistic predictions obeys a different, purely convexity-based statement. For weights $\alpha_m \geq 0$ summing to one, the log loss of $\bar{q} = \sum_m \alpha_m q_m$ satisfies

$$-\log \bar{q}(Y \mid X) \leq \sum_m \alpha_m [-\log q_m(Y \mid X)],$$

and squared loss satisfies the analogous inequality. These bounds compare an average with the weighted average of its constituent losses; they do not guarantee improvement over the best constituent and cannot recover information absent from every constituent output.

B.8 Evidence-preserving representations: proof of Proposition 2

Shared content and exact-support blocks. Include private variation in $\tilde{Z} = (Z, \xi_{1:M})$ and let $\tilde{A}_m$ index the coordinates observed by modality m. For a nonempty modality subset a, define

$$I(a) = \bigcap_{m \in a} \tilde{A}_m, \qquad C(a) = \bigcup_{a' \supsetneq a} I(a'), \qquad J(a) = I(a) \setminus C(a).$$

The union is empty when $a = [M]$. The set $I(a)$ is all content shared by the views in a; $J(a)$ contains only coordinates observed by exactly those views. These objects must not be conflated.

For the positive shared-content result we invoke Yao et al. (2024, Thm. 3.2) with its full setup: smooth invertible view maps, a smooth positive latent density on the stipulated product domain, the correct shared-content dimension, smooth encoders into a unit cube of that dimension, and a global optimum of the population alignment-minus-entropy objective. The conclusion identifies $\tilde{Z}_{I(a)}$ up to a smooth invertible map. The result does not by itself separate $J(a)$ from $C(a)$, or prove identification for a finite-sample surrogate.

Proof of the remaining claims in Proposition 2. To see the conditional ambiguity, choose a smooth change

$$(\tilde{z}_{J(a)}, \tilde{z}_{C(a)}) \longmapsto (\psi(\tilde{z}_{J(a)}; \tilde{z}_{C(a)}), \tilde{z}_{C(a)}),$$

whose full extension Ψ is a diffeomorphism on the latent domain; all other coordinates are unchanged. Such transformations include blockwise diffeomorphisms and admissible conditional triangular changes. If $m \notin a$, it observes no coordinate in $J(a)$ and its inputs are unchanged. If $m \in a$, it observes all coordinates in $J(a) \cup C(a)$ and its view map can be composed with the corresponding restriction of Ψ^{-1}. The transformed model has exactly the same observed views. Thus view geometry alone cannot rule out this conditional reparameterization. This is an ambiguity of the unrestricted smooth view model; additional distributional restrictions can reduce it.

For evidence invariance, Equation (23) shows that the state-density Jacobian cancels and that the transported statistic is the same random variable. Hence both likelihood-ratio terms in Equation (3) are unchanged, proving part (ii)'s transformation claim.

For reconstruction, the inverse view maps and coverage of all coordinates give a measurable map $Z = \Phi(X)$ at the fixed acquisition context. Let $R = E(X)$ and suppose a decoder G satisfies $G(E(X)) = X$ almost surely on the population under consideration. Then

$$Z = \Phi(G(R)), \qquad T(Z) = (T \circ \Phi \circ G)(R), \qquad q_\pi^{\mathrm{ref}}(\cdot \mid Z) = q_\pi^{\mathrm{ref}}(\cdot \mid \Phi(G(R))).$$

The evidence and posterior are therefore measurable functions of R. For a claim on a target population, the reconstruction identity must hold there as well; reconstruction of finitely many training examples is insufficient.

Finally, let $S \in \{0, 1\}$, let $Z_{\mathrm{sh}} \sim \mathcal{N}(0, 1)$ be independent of S, and let $Z_{12} = S + \epsilon$ with independent $\epsilon \sim \mathcal{N}(0, 1)$. Take three views with identity maps: the first two observe $(Z_{\mathrm{sh}}, Z_{12})$ and the third observes only Z_{sh}. All cohorts have the same state-conditional laws, so U is trivial and $T_1(Z) = Z_{12} - 1/2$. This nonconstant evidence cannot be recovered from the content Z_{sh} shared by all three views. The example isolates preservation of evidence; it does not require solving a cross-cohort component alignment problem. □

The information objective in Equation (7) is motivated by these distinctions. Alignment encourages common content, reconstruction encourages retaining the rest, and residual terms organize partially shared information after broadly shared blocks. Ordinary subtraction in observation space does not establish a unique conditional block decomposition under arbitrary nonlinear view maps. Exact reconstruction is sufficient, but not necessary, to preserve a particular T; a lossy statistic could retain that evidence alone. Conversely, an injective encoder may preserve all information while presenting it in coordinates that violate Equation (2). These limits distinguish information preservation, coordinate identification, and the behavior of the joint estimator in Equation (12).

C Synthetic benchmark: protocol and supplementary results

C.1 Evidence levels and comparability

The common-coordinate study supplies the generator's ten-dimensional Z to all compared state estimators. This removes representation error but does not reveal S during fitting; it is a mechanism-stage experiment, not an end-to-end raw-input result. The learned-view study uses the original multimodal observations and an estimated 13-dimensional lattice. The clinical study is a separate, label-informed architecture with 38 physiology dimensions. These three studies must not be pooled into one effect estimate or presented as runs of an identical estimator.

The five common-coordinate seeds are 58013, 58027, 58043, 58061 and 58073; the learned-view study uses the first three in both regimes, retaining the paired generated studies; the attribution study uses five separately generated studies with seeds 31415, 27182, 16180, 14142 and 17320. A seed in simulation is one generated study with multiple source and target observations. Tables report sample standard deviations across seeds, not standard errors, and small numbers of seeds are not interpreted as significance tests. All four target laws are reported: stronger selection is informative because a method can benefit from selection under one law and fail under its reversal. Equality of an evidence function across cohorts does not imply exactly equal marginal AUROC, because the distribution of conditioning covariates and the label mechanism can change task difficulty; we therefore do not impose a flat-AUROC validity criterion.

C.2 Generator and source fitting

The generator has three latent states, ten factors and three leaky-affine observation views with private variation. Source enrollment acts on the generator's selection-related root factors, which play the role of B in Figure 1a, and is strong and heterogeneous across the twelve cohorts; label styles, the measurement of Y, are heterogeneous. Target-law generation changes enrollment as stated in Section 5 together with site and label style under the original law definitions. The recorded binary Y uses a score involving continuous latent factors (including z_0 and $0.7z_2$), cohort-dependent terms

that do not involve S, private variation and site effects, which together play the role of $c_Y(e)$ and the label noise in Equation (1); it is not replaced with a label generated only from S, because a clean state task and an observed outcome task test different claims.

The state mixture uses $K = 3$, at most 150 EM iterations, two restarts and covariance ridge 10^{-3}. Source-validation density selects common versus state-specific covariance. Common-covariance states are aligned by the parallel-line estimator of Equation (4), with a check that the permutations preserve the observation density; distinct state covariances provide the alternative alignment of Proposition 1. Finite-sample EM and geometric fitting approximate the population construction, and the generator is not claimed to satisfy Equation (4) exactly. The natural-shift approximation keeps one direction. For averaging, the original fit and two within-cohort bootstrap fits are matched through source-validation posterior agreement and combined with equal weights. True S is excluded from all these choices; the single global source-S scoring permutation is determined after state models and outcome heads are fitted.

C.3 Outcome heads on common coordinates

Outcome heads have twelve inputs (ten coordinates and two state-evidence slots), two hidden layers of width 64 and a binary output. The ERM control fills the evidence slots with zero; state-only ablations zero the continuous coordinates. AdamW uses learning rate 10^{-3}, weight decay 10^{-4}, batch size 512, gradient clipping at 5, at most 40 epochs and patience 6 on source cohort-average NLL. The style variants add a centred cohort-specific linear logit and intercept, initialized at zero; deployment always uses the shared head. No weight averaging is used in these heads. Table 3 gives the proper scores under reversal. For state recovery under reversal, single corrected-fit state NLL is 0.787 and 0.891 (additive, with interactions), and averaged corrected-fit NLL is 0.781 and 0.883; these are modest improvements, distinct from the large correction-versus-uncorrected contrast of Table 1. A joint criterion, requiring both that state averaging lowers state NLL relative to the single corrected fit and that the averaged-evidence-plus-style head lowers outcome NLL relative to ERM, holds in 1/5 additive and 2/5 interaction studies.

Table 3: Recorded-endpoint prediction under reversed enrollment, given common Z: mean±SD over five generated studies.

Regime	Head	AUROC	AUPRC	NLL	Brier
additive	ERM	$0.779_{\pm.019}$	$0.878_{\pm.042}$	$0.538_{\pm.053}$	$0.181_{\pm.019}$
additive	+ conditional average	$0.780_{\pm.021}$	$0.878_{\pm.040}$	$0.532_{\pm.044}$	$0.179_{\pm.016}$
additive	+ conditional + style	$0.778_{\pm.013}$	$0.877_{\pm.040}$	$0.529_{\pm.051}$	$0.177_{\pm.020}$
additive	state only	$0.697_{\pm.066}$	$0.824_{\pm.062}$	$0.602_{\pm.091}$	$0.204_{\pm.031}$
with interactions	ERM	$0.810_{\pm.073}$	$0.884_{\pm.087}$	$0.515_{\pm.145}$	$0.168_{\pm.051}$
with interactions	+ conditional average	$0.812_{\pm.068}$	$0.887_{\pm.084}$	$0.513_{\pm.143}$	$0.168_{\pm.050}$
with interactions	+ conditional + style	$0.810_{\pm.071}$	$0.884_{\pm.088}$	$0.513_{\pm.147}$	$0.168_{\pm.052}$
with interactions	state only	$0.744_{\pm.047}$	$0.852_{\pm.098}$	$0.566_{\pm.119}$	$0.192_{\pm.045}$

C.4 Attribution study

Figure 4 tests whether the correction comes from the estimated selection direction rather than from conditioning as such. Five separately generated studies use the same twelve-source sample budgets, $K = 3$, rank-one correction, covariance-family selection and four target laws. Within each study a single fitted mixture supplies all three readouts of Figure 4; the random unit direction is fixed by the study seed plus 99013. Source S supplies only the final scoring permutation. Reference KL is $\mathbb{E}_Z \, \mathrm{KL}(q^{\mathrm{ref}}(\cdot \mid Z) \| \hat{q}(\cdot \mid Z))$ averaged over target observations, not actual-target excess log-loss. The aggregate file retains all 240 planned method–regime–law–seed rows, including oracle and linear-outcome diagnostics. These data are separate from the five averaged-fit studies and are not pooled with them. Their absolute levels are therefore not comparable with Table 1: with interactions, the estimated direction reaches 0.861 here against 0.825 ± 0.067 for the averaged fit there, a difference within one between-study SD that reflects different generated studies rather than an effect of averaging. Within the studies of Table 1, averaging does not lower performance: reversed-law state NLL is 0.891 for the single corrected fit and 0.883 for the average (Appendix C, outcome

heads). Per study, the estimated direction exceeds both other readouts in all five studies of each regime; across studies, its state AUROC is 0.860 ± 0.010 and 0.861 ± 0.033, against 0.151 ± 0.242 and 0.270 ± 0.178 for the random direction and 0.115 ± 0.188 and 0.252 ± 0.174 uncorrected. The attribution study is designed for the within-mixture comparison of the three readouts, not for absolute levels.

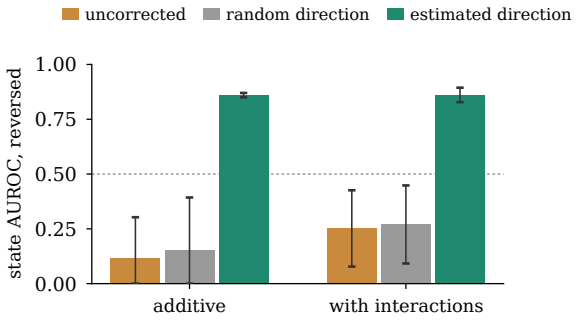


Figure 4: **Attribution of the estimated direction.** State AUROC under reversed enrollment for three readouts of the same fitted mixture: uncorrected, conditioned on a random unit direction, and conditioned on the estimated selection direction. The comparison tests whether the gain comes from the estimated direction; it is not an ablation of the number of latent states. Bars show means; error bars indicate sample standard deviations across five independently generated studies (truncated at 0 where one SD exceeds the mean).

C.5 Learned representations, calibration and diagnostic probes

The lattice has atom dimensions $(2, 1, 1, 1, 3, 2, 3)$ with hidden width 32. Per-modality standardization uses source training statistics; site FiLM context is zero in both source and target, so the lattice receives no information about the target's site change, which Proposition 2 holds fixed. The information objective is Equation (7): within-cohort contrastive alignment, reconstruction, conditional-residual reconstruction and 0.1 times a variance-floor penalty; AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 256, at most 30 epochs and patience 5 on source-validation information loss. No S or Y supervises the encoder.

For the calibration study we freeze the learned features $\tilde{z}$ and fit a shared, blockwise invertible map F jointly with the mixture using the Jacobian-normalized feature likelihood

$$\log p_e(\tilde{z}) = \log \sum_s \pi_{es}\, \mathcal{N}\big(F(\tilde{z}); \mu_{es}, \Sigma_s\big) + \log |\det J_F(\tilde{z})|, \tag{29}$$

so that $z = F(\tilde{z})$; the Jacobian normalizes the feature density but cancels from posterior ratios. Calibration first standardizes features with the base mixture's source mean and scale. Three block-diagonal LU layers have positive diagonal factors and positive-slope piecewise-affine activations; blocks never mix different information atoms, and shifted kinks are initialized with unit slopes for an exact identity initialization. The affine control fixes activation slopes to one; the mixture-only control fixes the entire transform. All refinements start from the same fitted density; Adam with learning rate 10^{-3}, batch size 512, clipping at 5, at most 40 epochs and patience 6 on source-validation feature NLL, with the initial model included as a candidate. The outcome head recipe matches the common-coordinate heads; all lattice-based heads have the same input width, and a calibrated-coordinate ERM control measures a coordinate change without added state evidence. Raw-concatenated-input ERM with the same hidden widths reaches reversed-selection outcome AUROC 0.792 and 0.812. Table 4 gives the learned-view results.

Table 4: Learned-view interface under reversed enrollment: the same frozen 13-dimensional lattice is used by all heads; DISCERN variants add state evidence. Mean±SD over three paired studies. Lattice + ERM outputs no state.

Method	S AUROC, add.	S AUROC, inter.	Y AUROC, add.	Y AUROC, inter.	Y NLL, add.	Y NLL, inter.
Lattice + ERM	–	–	$0.863_{\pm.029}$	$0.806_{\pm.104}$	$0.507_{\pm.085}$	$0.622_{\pm.262}$
DISCERN base	$0.118_{\pm.095}$	$0.306_{\pm.083}$	$0.850_{\pm.034}$	$0.797_{\pm.107}$	$0.502_{\pm.092}$	$0.669_{\pm.295}$
DISCERN mixture refinement	$0.125_{\pm.076}$	$0.276_{\pm.095}$	$0.849_{\pm.032}$	$0.798_{\pm.104}$	$0.514_{\pm.103}$	$0.678_{\pm.310}$
DISCERN affine calibration	$0.130_{\pm.080}$	$0.279_{\pm.099}$	$0.850_{\pm.023}$	$0.796_{\pm.107}$	$0.495_{\pm.063}$	$0.673_{\pm.303}$
DISCERN nonlinear calibration	$0.135_{\pm.084}$	$0.307_{\pm.077}$	$0.849_{\pm.021}$	$0.795_{\pm.105}$	$0.522_{\pm.080}$	$0.683_{\pm.311}$

Nonlinear calibration improves source-validation feature-density NLL from 5.240 to 4.233 (additive) and from 5.925 to 4.939 (with interactions), while state NLL under reversal (12.72 ± 4.83 base versus 12.89 ± 4.93 calibrated; 7.62 ± 2.49 versus 7.79 ± 1.90) does not improve; no seed has both primary NLL endpoints improved. Fitting the feature density more accurately is insufficient to identify the intended state family. No DISCERN variant improves outcome AUROC over Lattice + ERM; affine calibration gives the lowest additive outcome NLL (0.495 against 0.507), within one SD.

Diagnostic probes are fitted only after deployed models, checkpoints and predictions are saved. Linear multinomial probes select $C \in \{0.1, 1, 10\}$ by their own validation NLL; neural probes have two width-64 hidden layers, at most 30 epochs and patience 6. Source-trained probes use source S and evaluate all target laws; on the frozen lattice they obtain source-validation AUROC 0.934 and 0.889 but reversed-target AUROC 0.173 and 0.652, so source-trained probes can themselves fail under reversal. Target-trained probes use 5,000 independent observations from a reserved target-adaptation stream (4,000 training, 1,000 validation) and evaluate the existing 10,000 target-test samples; they obtain 0.868 and 0.832 on the selection-off law and 0.993 and 0.958 on the reversed law, where the high values can exploit target-specific selection. Oracle-Z selection-off values are 0.903 and 0.859. These results support residual state readability, not information-theoretic sufficiency or unsupervised disease identification; true target labels never initialize, select or update a deployed method.

Numerical checks verified inverse round trips, analytic versus automatic-differentiation Jacobians, atom block structure, identity initialization and one-dimensional density normalization. The main run completed six cases, yielding 240 state-metric rows, 240 outcome rows and 432 diagnostic-probe rows; thirty previous-baseline replay comparisons passed; 29 frozen source and protocol file hashes and 162 saved checkpoint hashes were verified, and diagnostic probes left deployed checkpoints unchanged.

C.6 Number of assumed latent factors

The structural state model assumes a number N of latent factor nodes, which clinical data do not reveal. Figure 5 varies $N \in \{2, 3, 4, 5\}$. Panels (a) and (b) use a separate generator that follows the structural model of Figure 1a with three states and three latent physiology factors, so that $N = 3$ is correct, in the additive and interaction regimes; the coordinates are supplied as in Section 5.1, the ERM reference is a transformer on the same coordinates, and each point averages three seeds. Panel (c) uses the clinical variant of Appendix D, where the number of factors is unknown, and reports the AUROC gain over ERM in each direction.

State recovery depends on the assumed number. State AUROC peaks at the true $N = 3$ in both regimes and falls to about chance or below when N is misspecified, lowest at $N = 2$ and $N = 5$ in the additive regime. Even at $N = 3$ it is weak, about 0.54 (additive) and 0.64 (with interactions), well below Table 1; the generator and the state model differ from that benchmark, so the levels are not comparable. The outcome gain is less sensitive: its mean is positive at every N in both regimes, between about 3 and 5 points, but the additive error bars include zero at $N = 4$ and $N = 5$, and the additive gain is largest at $N = 2$, where state AUROC is below chance. This is the pattern of Proposition 3: the head also reads the continuous coordinates, and state recovery improves only one term of the bound. Because the reference is a transformer rather than the matched head of Table 3, these gains are not comparable with the near-zero gains there.

On ICU the gain lies between about 4.5 and 5 points at every N in both directions, and its variation across N, under half a point, is of the order of the spread across seeds, so the clinical result does not

hinge on the assumed number. These gains exceed the 3.6 and 3.5 points of Table 2 over matched ERM; the sweep is a sensitivity analysis and is not pooled with Table 2.

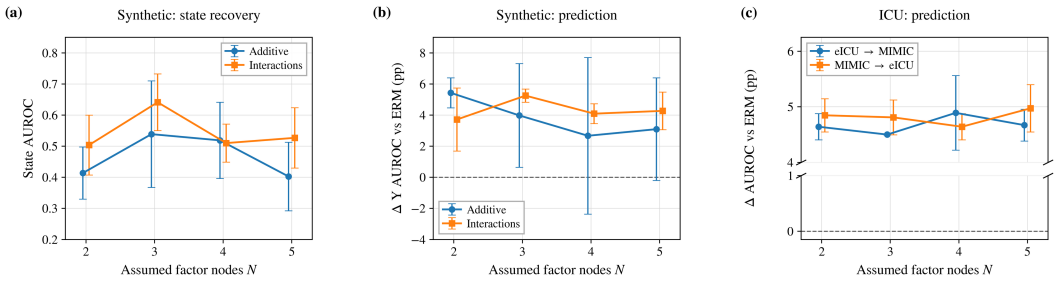


Figure 5: **Number of assumed latent factors.** (a) State AUROC and (b) outcome AUROC gain over a transformer ERM, in percentage points, on a generator with three states and three latent factors, when DISCERN assumes N factor nodes. (c) AUROC gain over ERM in the ICU study, where the number of factors is unknown (broken axis). Points are means over three seeds.

D ICU EXPERIMENT: PROTOCOL AND SUPPLEMENTARY RESULTS

D.1 COHORT ACCOUNTING

Both directions preserve a fixed patient split and feature contract: 48-hour observations, a 6-hour gap and same-admission hospital mortality. The eICU-to-MIMIC target has 9,922 stays from 9,922 patients at one physical hospital; its 23 care-unit and time groups are not 23 hospitals. The MIMIC-to-eICU target has 26,763 stays from 22,982 patients at 57 physical hospitals. The saved configurations identify MIMIC-IV 2.0 and eICU-CRD 2.0. Eligibility requires age at least 18 and ICU length of stay from 54 to 240 hours, with a half-hour rounding tolerance. The retained source caches contain 18,419 MIMIC stays and 58,391 eICU stays; these are not the target-test counts. eICU preprocessing removes 11 stays with unknown sex and 869 stays to resolve cross-hospital patient duplicates; where cross-hospital time ordering is unavailable, the patient-to-hospital assignment uses the lexicographically smallest stay key. Training and validation use fixed patient-disjoint role plans (Table 5); target adaptation partitions are held separately and unused in this round. The cache enforces a 48-hour event-time window, but prospective record availability and chunk-order Glasgow Coma Scale reconstruction have not been verified, and the length-of-stay eligibility limits applicability beyond this population.

Table 5: Source role sizes in the fixed clinical patient plans (ICU stays). Validation excludes separately reserved adaptation data; target cohorts are the same across seeds.

Direction	Seed	Source train	Source validation
E→M	42	18,555	6,039
E→M	3407	21,952	4,089
E→M	2026	20,461	4,679
M→E	42	7,692	1,944
M→E	3407	7,063	2,176
M→E	2026	8,509	1,691

D.2 ARCHITECTURE AND TRAINING

The physiology views contain the first eleven demographic inputs, bedside measurements and laboratory features, excluding the creatinine, lactate, bilirubin and culture-site families reserved for construct checks. Their seven information blocks have dimensions $(4, 2, 2, 2, 4, 8, 16)$, giving 38 continuous coordinates. A separate context path receives the excluded laboratory columns, remaining static features, medications and procedures, with widths 128 and 64. The mortality head reads physiology and context through a width-128 layer; the mechanism branch adds a linear readout

of two standardized state log-odds coordinates, initially weighted zero. All prediction inputs are available to every rich variant.

Five rich variants share the backbone: matched ERM; information only, the lattice without a state branch; uncorrected state, state evidence without conditioning; conditional state (DISCERN); and no edges, the block-ordered mechanism with its structural edges removed. They share a five-epoch supervised warm-up with AdamW at 10^{-3}, then independent optimizers at 3×10^{-4}, weight decay 10^{-4}, batch size 256, clipping at 5 and at most 30 epochs, with patience 6 on source-validation mean group AUROC and a 10^{-4} improvement threshold; the warm-up checkpoint is an explicit candidate, and a fixed five-epoch averaging window around the best epoch is considered with source-only batch-norm recalculation and source validation choosing raw versus averaged weights (a local weight-averaging option, not canonical SWAD). The information objective has weight 0.01 and the mechanism density regularizer 0.001. The mechanism initializes with $K = 3$, two EM restarts, 100 iterations and ridge 0.01; the block order is imposed. State covariance parameters, encoder, context and outcome head update together; the selection basis is recomputed once per epoch outside gradient tracking. The feature density is a regularizer without a raw-input Jacobian, not an identified generative likelihood of clinical observations. The deployed mechanism subtracts the rank-one marginal evidence, and the source structure audit retains all 38 coordinates in all six fits. Because gradients from the supervised endpoint reach the physiology encoder, the clinical state model is label-informed through its encoder even though its initialization fit excludes Y; the design does not establish label-free state identification.

D.3 Matched ablations

Table 6 isolates the mechanism choices on the shared backbone. Appendix C.6 varies the number of latent factor nodes the mechanism assumes.

Table 6: Matched-architecture ICU ablations, mean±SD over three seeds; all prediction inputs are available to every variant.

Variant	AUROC E→M	AUROC M→E	AUPRC E→M	AUPRC M→E
Matched rich ERM	$0.782_{\pm.017}$	$0.777_{\pm.011}$	$0.462_{\pm.022}$	$0.398_{\pm.012}$
Information only	$0.788_{\pm.018}$	$0.783_{\pm.015}$	$0.470_{\pm.020}$	$0.402_{\pm.015}$
Uncorrected state	$0.788_{\pm.018}$	$0.782_{\pm.016}$	$0.470_{\pm.020}$	$0.401_{\pm.016}$
DISCERN	$\mathbf{0.818}_{\pm.018}$	$\mathbf{0.812}_{\pm.016}$	$\mathbf{0.502}_{\pm.020}$	$\mathbf{0.431}_{\pm.016}$
No edges	$0.788_{\pm.018}$	$0.782_{\pm.016}$	$0.470_{\pm.020}$	$0.401_{\pm.016}$

D.4 Historical baselines

The historical rows of Table 2 are earlier source-trained implementations evaluated on the same fixed target cohorts. Early fusion uses a two-layer width-256 GELU network with dropout 0.1; late fusion adds the logits of five such modality networks trained jointly and has a different parameter count. V-REx, GroupDRO and CORAL use local objectives based on Krueger et al. (2021); Sagawa et al. (2020); Sun & Saenko (2016); ERM and the weight-averaging model sample up to 256 observations per step, while the environment-based objectives sample up to 16 hospitals with up to 32 observations each, so compute is not exactly matched. V-REx uses weight 1 for five epochs and 10 thereafter; GroupDRO a log-weight step of 0.01; CORAL alignment weight 1. All use Adam at 10^{-3} with weight decay 10^{-4}, at most 60 epochs, patience 8 and source-validation pooled AUROC, source-fitted standardization and no target adaptation. The SWAD-style baseline's backbone and schedule differ from the rich architecture, which is why the matched controls are necessary. These comparisons provide context; they cannot by themselves isolate the mechanism branch.

D.5 Construct probes and post-fit diagnostics

Construct probes use source-selected ridge regression with $\alpha \in \{0.1, 1, 10, 100\}$ and age and sex controls; they read physiology or state evidence, never the context representation. The held-out laboratory families can nevertheless enter the mortality context branch. Creatinine R^2 is 0.307 and 0.222 from the continuous representation and 0.059 and -0.075 from state evidence, against 0.512

and 0.437 for a source-fitted 38-dimensional PCA reference; these proxies are not disease states, and measured-only target subgroups may be selected by measurement practices. Post-fit diagnostics find that the first natural-shift direction explains 37.1 to 57.6% of the fitted squared singular-value energy of the clinical mechanisms; this is neither a causal dimension estimate nor proof that a larger rank would help. Training loss commonly falls while validation NLL increases, and early stopping often selects the warm-up or an early epoch; checkpoints before the warm-up were not selectable in this protocol.

E ADNI: PROTOCOL AND COMPLETE RESULTS

Endpoint and cohort. The ADNI experiment classifies contemporaneous CN, MCI and AD diagnoses; it does not predict progression. The cohort contains 1,807 patients: 1,264 source-training, 272 source-validation and 271 target patients, where within each diagnosis the oldest 15% define an age-shift target, an age-based selection shift. This diagnosis-controlled holdout had been evaluated previously, so the analysis is exploratory. A fixed, label-blind patient-hash rule reserves 54 validation patients and 54 target patients as separate unlabeled context sets; the remaining 218 source-validation patients select downstream checkpoints, and all 20 methods share the same 217-patient primary target (56 CN, 113 MCI, 48 AD). Roles are patient-disjoint. Source environments are source-training age tertiles (420, 415 and 429 patients), not hospitals.

Shared representation and replication. All methods reuse one upstream source-supervised encoder selected on source validation (encoder seed 2026), whose four 16-dimensional views represent T1 MRI, demographics, behaviour and genetics/history, each with an availability indicator; all retained patients have MRI, and image inputs were screened for harmful content with ShieldGemma 2 (Zeng et al., 2025) as a routine data-integrity check (Cheong, 2025; 2026). Standardization uses source training only. Seeds 42, 3407 and 2026 repeat downstream fitting, not upstream MRI training, so sample SD reflects downstream variability on a shared encoder and cohort. Some clinical measurements use nearest visits rather than strictly pre-MRI visits; neither prospective availability nor an independent source-validation test is claimed.

Identity of the variant. These results use the earlier joint categorical DISCERN extension: a linear lattice, three Gaussian latent components, a tempered label-assisted EM objective with weight 0.25, class-pair mean-relevance filtering across age environments, and a categorical mixture head. It differs from the common-coordinate conditional-evidence estimator and from the clinical rich-encoder experiment, does not instantiate their estimator, and its tempered objective is not a normalized joint likelihood. Strict source deployment removes complementary-coordinate shared style and source-environment head intercepts while retaining source posterior offsets; common-style variants retain the shared style term; target variants estimate posterior-branch centring from the separate 54-patient unlabeled context. Target labels are never used for fitting or checkpoint selection.

Training and reporting. Supervised networks have joint softmax outputs and are selected on source-validation multiclass NLL; MLPs use two 256-unit GELU layers, dropout 0.1, Adam at 10^{-3} with weight decay 10^{-4}, at most 60 epochs and patience 8. DG settings are fixed local implementations. State EM uses at most 200 iterations; categorical heads use 600 Adam updates at learning rate 0.02. All 60 primary method–seed cells are complete and all 24 state-EM traces report convergence, which is not evidence of identification. An archived paired diagnosis-stratified bootstrap (2,000 patient draws of the mean three-seed probability vectors) conditions on fixed models, omits retraining uncertainty and is not multiplicity-adjusted; for example, target strict DISCERN minus early ERM has an ensemble macro-AUROC difference of -0.020 with 95% interval $[-0.063, 0.022]$. Primary no-environment DISCERN predicts no AD cases in any seed, so some per-class precision values are undefined; all four metrics in Table 7 have three valid seeds. The evidence supports a bounded exploratory result, not a clinical improvement claim.

Table 7: Complete ADNI primary results on the same 217 target patients: mean±SD over three downstream fits sharing one upstream encoder. AUROC and AUPRC are macro one-versus-rest (↑); NLL and Brier are lower-is-better; Brier sums squared probability errors over the three classes. The lower group accesses 54 separate unlabeled target-context patients.

Method	Macro AUROC	Macro AUPRC	NLL	Brier
Source-only prediction; no target adaptation				
ERM early fusion	$0.738_{\pm.004}$	$0.587_{\pm.002}$	$0.917_{\pm.043}$	$0.555_{\pm.008}$
ERM late fusion	$0.747_{\pm.008}$	$0.596_{\pm.011}$	$0.909_{\pm.010}$	$0.551_{\pm.006}$
ERM MRI only	$0.582_{\pm.023}$	$0.419_{\pm.018}$	$1.022_{\pm.019}$	$0.619_{\pm.012}$
ERM tabular only	$0.734_{\pm.007}$	$0.586_{\pm.016}$	$0.897_{\pm.020}$	$0.543_{\pm.005}$
IRM-style	$0.728_{\pm.010}$	$0.567_{\pm.015}$	$0.923_{\pm.064}$	$0.563_{\pm.026}$
V-REx	$0.736_{\pm.008}$	$0.583_{\pm.011}$	$0.940_{\pm.011}$	$0.551_{\pm.005}$
GroupDRO	$0.736_{\pm.009}$	$0.580_{\pm.012}$	$0.906_{\pm.039}$	$0.557_{\pm.018}$
CORAL	$0.736_{\pm.010}$	$0.580_{\pm.013}$	$0.901_{\pm.034}$	$0.553_{\pm.015}$
Lattice ERM (wide)	$0.706_{\pm.015}$	$0.553_{\pm.027}$	$0.919_{\pm.016}$	$0.557_{\pm.009}$
Lattice ERM (narrow)	$0.744_{\pm.004}$	$0.576_{\pm.012}$	$0.941_{\pm.008}$	$0.549_{\pm.004}$
Lattice ERM (deep)	$0.698_{\pm.016}$	$0.558_{\pm.013}$	$0.980_{\pm.046}$	$0.582_{\pm.021}$
DISCERN source strict	$0.711_{\pm.019}$	$0.538_{\pm.029}$	$0.977_{\pm.025}$	$0.579_{\pm.004}$
DISCERN source common-style	$0.717_{\pm.002}$	$0.564_{\pm.002}$	$1.052_{\pm.009}$	$0.627_{\pm.005}$
Independent unlabeled target context; 54 patients				
Target-normalized ERM	$0.734_{\pm.012}$	$0.579_{\pm.012}$	$0.956_{\pm.047}$	$0.567_{\pm.011}$
DISCERN target strict	$0.709_{\pm.021}$	$0.535_{\pm.035}$	$0.973_{\pm.033}$	$0.577_{\pm.009}$
DISCERN target common-style	$0.715_{\pm.002}$	$0.563_{\pm.002}$	$1.053_{\pm.011}$	$0.628_{\pm.006}$
DISCERN no-environment	$0.726_{\pm.000}$	$0.586_{\pm.000}$	$0.891_{\pm.000}$	$0.543_{\pm.000}$
DISCERN affine posterior	$0.716_{\pm.003}$	$0.563_{\pm.004}$	$1.054_{\pm.014}$	$0.629_{\pm.007}$
DISCERN deep strict	$0.710_{\pm.005}$	$0.588_{\pm.020}$	$0.919_{\pm.027}$	$0.560_{\pm.022}$
DISCERN deep common-style	$0.761_{\pm.017}$	$0.602_{\pm.011}$	$0.937_{\pm.031}$	$0.529_{\pm.013}$

F Related work

Medical multimodal learning. Fusion architectures for imaging and electronic health records span early, late and hybrid designs (Huang et al., 2020; Acosta et al., 2022; Kline et al., 2022); MedFuse handles paired and unpaired time series and radiographs (Hayat et al., 2022), HAIM combines tabular, temporal, text and image embeddings (Soenksen et al., 2022), and contrastive objectives pre-train imaging encoders from clinical tables (Hager et al., 2023). Missing-modality transformers (Ma et al., 2022; Zhang et al., 2022) and modality-balancing objectives (Wang et al., 2020; Peng et al., 2022) address in-distribution failure modes; surveys organize the space (Stahlschmidt et al., 2022; Baltrušaitis et al., 2019; Liang et al., 2024). Recent work disentangles shared from modality-specific diagnostic information in paired ECG and cardiac MRI (Tonekaboni et al., 2025), treats modality missingness as informative through a latent health state, with experiments on MIMIC-IV and eICU (Liang et al., 2025), and benchmarks EHR and chest X-ray fusion within one health system (Elsharief et al., 2025). These works assume representative training cohorts and evaluate on held-out patients of the same cohort or health system. Shortcut learning in medical imaging (Zech et al., 2018; DeGrave et al., 2021; Winkler et al., 2019; Oakden-Rayner et al., 2020; Geirhos et al., 2020; Brown et al., 2023) and deployment failures (Wong et al., 2021; Nestor et al., 2019; Finlayson et al., 2021) document the cost. Huang et al. (2021) explain when multiple modalities provably outperform a single one by the quality of a shared latent representation; Lu (2023) give a learning-theoretic account based on heterogeneity between modalities and a learnable connection between them. More recent analyses show that contrastive alignment captures mainly redundant information and can miss unique and synergistic terms (Dufumier et al., 2025), that whether alignment helps depends on the balance of redundant and unique information (Tjandrasuwita et al., 2025), and that modality competition can be rebalanced by a game-theoretic regularizer (Kontras et al., 2025). These address in-distribution learning; none models how a disease generates the modalities or how their evidence changes across cohorts.

Robust machine learning. IRM (Arjovsky et al., 2019), V-REx (Krueger et al., 2021), GroupDRO (Sagawa et al., 2020), CORAL (Sun & Saenko, 2016), LISA (Yao et al., 2022), JTT (Liu et al., 2021), DFR (Kirichenko et al., 2023), SWAD (Cha et al., 2021) and test-time adaptation (Iwasawa & Matsuo, 2021; Wang et al., 2021) form the standard baselines; DomainBed (Gulrajani & Lopez-Paz, 2021) and WILDS (Koh et al., 2021) show how hard ERM is to beat, and Rosenfeld et al. (2021); Kamath et al. (2021) show when invariance criteria fail. NuRD (Puli et al., 2022) and invariant-feature subspace recovery (Wang et al., 2022) are the closest statistical targets; anchor regression (Rothenhäusler et al., 2021), invariant causal prediction (Peters et al., 2016; Rojas-Carulla et al., 2018) and conditional variance penalties (Heinze-Deml & Meinshausen, 2021) condition on observed labels or anchors. Clinical DG frameworks (Zhang et al., 2021) and harmonization (Johnson et al., 2007; Fortin et al., 2018) address site effects; Subbaswamy et al. (2019); Subbaswamy & Saria (2020); Castro et al. (2020) argue for causal modelling of clinical shifts. In critical care, multi-database studies find that internal validation overestimates external performance and that site-specific features act as shortcuts for mortality prediction (Allyn et al., 2026), that sepsis models degrade across MIMIC-IV, eICU and HiRID with the best adaptation strategy depending on target data (Tranchellini et al., 2026), and that anchor regression with the database as anchor aids transfer across nine ICU databases (Londschien et al., 2026). Causal representation learning builds on identifiability results for finite Gaussian mixtures (Yakowitz & Spragins, 1968), nonlinear ICA with auxiliary variables (Hyvärinen & Morioka, 2016; Hyvärinen et al., 2019; Khemakhem et al., 2020), mixture priors with piecewise-affine decoders (Kivva et al., 2022), content and style from augmentations (von Kügelgen et al., 2021; Zimmermann et al., 2021; Daunhawer et al., 2023), multi-view partial observability (Yao et al., 2024; Sturma et al., 2023), sparse cross-modal causal structure in biomedical observations (Sun et al., 2025), partial disentanglement for adaptation (Kong et al., 2022) and sparsity (Lachapelle et al., 2022), is unified by viewing identification as preserving known invariances (Yao et al., 2025), and is limited by the impossibility of unsupervised disentanglement without inductive bias (Locatello et al., 2019). Latent structure can remain identifiable under selection bias in linear Gaussian models (Dai et al., 2025), a setting close to our enrollment shift but aimed at causal discovery rather than transfer. Our result concerns a function of the latent vector, the evidence about an unobserved state that survives enrollment changes, and its alignment and coordinate premises are explicit.