PRUDENT: REPRESENTING EVIDENCE-BASED CLINICAL DECISION MAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline reinforcement learning for clinical treatment recommendation typically combines observed practice, clinical knowledge, and outcome-based objectives into a single optimisation problem, with the resulting policy determining the recommended action. While this formulation provides a natural way to optimise treatment, it obscures distinctions that evidence-based medicine requires at the point of care, including where evidence comes from, how strong it is, where it applies, and when it is silent. We instead take the clinical decision, rather than the optimised policy, as the object of study. We introduce **PRUDENT**, an evidence-based decision framework in which routine clinical practice anchors the decision while guidelines and learned policies remain explicit evidence sources, each retaining its strength, coverage, and ability to remain silent. PRUDENT characterises the decision context along two continuous dimensions, measuring how atypical the current state is relative to routine precedent and how strongly the evidence sources that speak disagree. Together, these dimensions form a decision plane that distinguishes routine decisions, known controversy, consensus extrapolation, and the true edge. PRUDENT represents each evidence source as a transport route from routine practice toward the action it supports, using conditional flow matching to learn these routes. The decision context governs how far departure from routine is warranted, while evidence strength limits how far each source may be followed. The resulting candidates are arbitrated separately for each action dimension using only the sources that speak, and the final action is projected onto the clinically feasible set. On MIMIC-IV, the decision plane identifies a consensus-extrapolation region comprising 22% of decisions in which patients have lower SOFA and lactate and higher mean arterial pressure, receive substantially less fluid and vasopressor treatment, yet experience $1.76\times$ the mortality of patients in the routine region. These findings are observational and require no counterfactual estimate. In controlled simulation with known optimal control, PRUDENT adapts its departure from routine across five regimes, moving away when routine practice becomes unreliable while remaining close when it remains appropriate. Together, these results show how evidence-based clinical decision making can be represented without collapsing heterogeneous evidence into a single optimised policy.

1 INTRODUCTION

A decade of work on treatment recommendation in critical care shares one commitment. Clinical knowledge of every kind is converted into a scalar objective, a policy is optimised against it, and the maximiser is reported as the recommendation (Komorowski et al., 2018; Raghu et al., 2017; Peng et al., 2018). Guidelines enter as reward shaping or as a constraint, observed practice enters as a behavioural regulariser, and the resulting policy is evaluated by estimating its value under the data collecting behaviour. The formulation is inherited from reinforcement learning, where a single objective is the definition of the problem rather than a modelling choice.

What scalarisation discards. A clinical recommendation is not a number. It carries a source, a strength grade recording how confident the evidence base is, a coverage condition naming the states to which it applies, and an applicability judgement about the individual patient. Once that recom-

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.

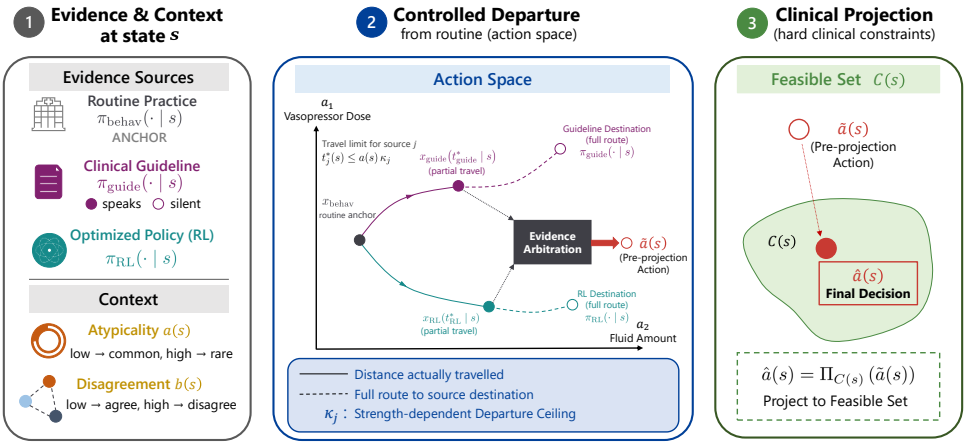


Figure 1: **PRUDENT represents a clinical decision rather than optimising a policy.** Routine practice is the origin of every decision. Each evidence source proposes a destination when it speaks and may be silent on a given action dimension. The context diagnosis governs warranted departure from routine practice, while evidence strength caps how far a source may be followed. The resulting decision remains attributable to the evidence that produced it.

mendation has been folded into a reward term, none of the four survives. A guideline statement graded weak and a value estimate fitted to one hospital's records become indistinguishable contributions to one scalar, and the decision that results cannot afterwards be attributed to either. Silence is discarded along with strength: a source that has nothing to say about the present patient becomes indistinguishable from one that endorses whatever the other sources propose.

This is not how evidence-based practice is understood by its own proponents. Its founding statement defines the field as the conscientious, explicit and judicious use of current best evidence, with judgement applied at the point of care rather than compiled away in advance (Sackett et al., 1996). GRADE exists because strength is a property of evidence that must remain visible to the clinician using it (Guyatt et al., 2008). A representation faithful to this account must keep the sources apart until the decision is made.

Our position. We take the object of study to be the decision rather than the policy, and call the resulting framework PRUDENT, for PRincipled Unified Decision Evidence represeNTation. The name reflects the judicious weighing of evidence against the situation at hand. Evidence sources are represented as distributions over the same action space, each retaining its grade, coverage, and right to say nothing. Routine clinical practice is the anchor of every decision, so evidence justifies controlled departure from customary care rather than replacing it. Before evidence is weighed, the decision context is diagnosed along two axes that answer a prior question: whether the present patient lies where routine practice is well supported by precedent, and whether the sources that speak agree. Neither implies the other, and departure from routine care is warranted for different reasons in each case.

An optimised policy occupies no privileged position in this account. It is one evidence source, subject to the same treatment as a guideline recommendation and to a ceiling on how far the decision may follow it. This substantive commitment makes the reliability of an optimisation a property to be examined, rather than the standard against which all other evidence is measured.

Contributions. **A representation of evidence-based clinical decision making.** Clinical evidence is represented as explicit decision entities rather than collapsed into a reward, with source, strength, coverage and silence all retained through to the decision. We give the requirements such a representation must satisfy and an operator that meets them (Sections 3 and 4.1).

A context-dependent arbitration architecture. Routine practice serves as the decision anchor and other sources justify controlled deviation from it. A diagnosis of context governs warranted departure, while evidence strength enters through distinct mechanisms that limit how far and how strongly a source may influence the decision (Sections 4.2 to 4.4).

Evidence that the representation captures clinically meaningful structure. On MIMIC-IV the representation isolates a region containing 22% of decisions in which patients present with lower SOFA, lower lactate and higher mean arterial pressure, receive fluid and vasopressors substantially less often, and yet die at 1.76 times the rate of the routine region. In a simulated environment with known optimal control, the context-dependent gate adapts across settings in which the anchor degrades with rarity, fails for unobserved or random reasons, remains correct, or is worse on common states; no single fixed departure distance serves all of them (Section 5).

2 Related Work

Clinical decision optimisation. Treatment recommendation in critical care is commonly formulated as offline reinforcement learning over rewards constructed from clinical surrogates and outcomes. The AI Clinician learned discretised vasopressor and fluid doses from MIMIC (Komorowski et al., 2018), with subsequent work extending treatment optimisation to continuous state spaces and alternative value estimators (Raghu et al., 2017; Peng et al., 2018). A central difficulty is that offline policies may propose actions unsupported by the observed data (Levine et al., 2020), motivating behaviour-constrained and conservative methods (Fujimoto et al., 2019; Kumar et al., 2020), as well as approaches that avoid querying out-of-distribution actions during training (Kostrikov et al., 2022). Off-policy value estimates in clinical cohorts can be sensitive to the estimator and cohort construction (Jeter et al., 2019), motivating methodological guidance for healthcare RL (Gottesman et al., 2019) and broader clinical caution in interpreting learned treatment policies (Saria, 2018).

Representing medical knowledge and attribution. Clinical guidelines and expert knowledge have been incorporated into clinical decision models through guideline-shaped rewards and safety constraints (den Hengst et al., 2024; Wang et al., 2024), while recent work has explored LLMs for extracting structured recommendations from narrative guidelines (Hsu et al., 2025). In these approaches, knowledge typically enters the learning or decision pipeline as rewards, constraints, or structured rules. Its source, strength and coverage are therefore not generally retained as explicit components of the resulting decision. Interpretability methods can attribute a completed prediction to its inputs (Lundberg & Lee, 2017), while interpretable-by-design approaches instead make the model structure itself the explanation (Rudin, 2019). PRUDENT follows the latter principle at the level of evidence composition.

Combining evidence distributions. Multiple policies are commonly handled by selecting one according to an estimated value or by mixing them, neither of which preserves the role of each source in the resulting decision. Moreover, summaries of disagreeing distributions may fall between modes where no individual source provides strong support. Conditional flow matching provides transport paths between distributions (Lipman et al., 2023; Tong et al., 2024); we use this machinery to represent controlled movement from routine practice toward an evidence source. The broader gap addressed here is representational: how to combine heterogeneous clinical evidence while retaining source identity, graded strength, coverage, and silence until the decision is made.

3 Problem Formulation

Computational treatment recommendation is usually posed as policy optimisation: given a reward r assembled from clinical surrogates and outcomes, one seeks $\pi^\star = \arg\max_\pi \mathbb{E}_\pi[\sum_t \gamma^t r(s_t, a_t)]$. Heterogeneous clinical knowledge may enter through the reward, constraints, or regularisation toward a reference policy. In each case, the evidence is absorbed into the optimisation problem, after which its source, grade, coverage and applicability are generally no longer recoverable from the resulting policy.

We instead take the decision as the object of study. Let $\pi_{\text{behav}}(\cdot \mid s)$ denote routine clinical practice and $\{(\pi_j(\cdot \mid s), g_j)\}_{j=1}^J$ further evidence sources over the same action space $\mathcal{A}$, each carrying a

strength grade g_j. With $\mathcal{C}(s) \subseteq \mathcal{A}$ the clinically admissible actions, define

$$\mathcal{D}(s) = \Pi_{\mathcal{C}(s)}\Big(\mathcal{M}\big(\pi_{\mathrm{behav}}(\cdot \mid s), \{(\pi_j(\cdot \mid s), g_j)\}_{j=1}^{J}, \mathrm{ctx}(s)\big)\Big), \tag{1}$$

where $\mathrm{ctx}(s)$ diagnoses the decision context, $\mathcal{M}$ is an arbitration map, and $\Pi_{\mathcal{C}(s)}$ enforces hard clinical constraints. Routine practice is the origin of the decision, while an optimised policy is one evidence source rather than the target.

Reading Equation 1 as a model of evidence-based practice imposes three requirements. **Anchoring.** A source silent at s must alter neither the decision nor its diagnosed context. Writing $\mathcal{E}$ for the evidence set and π' for a source silent at s, we require

$$\mathcal{D}(s; \mathcal{E} \cup \{\pi'\}) = \mathcal{D}(s; \mathcal{E}), \qquad \mathrm{ctx}(s; \mathcal{E} \cup \{\pi'\}) = \mathrm{ctx}(s; \mathcal{E}). \tag{2}$$

Silence is therefore absence from the arithmetic rather than agreement with routine practice. **Attribution.** The decision record must expose how far the decision moved, the direction in which each source pushed it, and the weight each source received. **Strength.** Evidence strength must have pre-specified consequences that differ by kind, rather than acting only as a tunable scalar weight. Section 4 constructs an operator satisfying these requirements.

4 PRUDENT

4.1 REPRESENTING EVIDENCE

Every source is a conditional distribution over the same action space, so that sources can be compared without being converted into terms of a common objective. The action is two dimensional, comprising fluid volume and maximum vasopressor dose over the next four hours, each scaled to $[0, 1]$. We use two evidence sources alongside routine practice.

Routine practice π_{behav} is estimated by behavioural cloning of observed care and serves as the anchor of Equation 1: every evidence route begins from routine practice rather than competing with it as another destination. The guideline source π_{guide} represents graded recommendations compiled from SSC 2021 (Evans et al., 2021). When an applicable rule specifies an action, it induces a distribution over that action whose precision depends on its GRADE strength; the full construction is given in Appendix C.1.

Silence is represented per action dimension. If no guideline rule bears on dimension d at state s, the corresponding guideline distribution falls back to π_{behav} for route construction, while a firing mask records that the guideline did not speak. This distinction is preserved downstream (Section 4.3). The optimised source π_{rl} is learned by offline reinforcement learning and enters the representation as another evidence source. Training details for both learned sources are given in Appendix D.

4.2 DIAGNOSING THE DECISION CONTEXT

Two distinct situations warrant departure from routine practice. The patient may lie where routine practice is poorly constrained by precedent, so the anchor is less reliable; or the evidence sources may point in different directions, so a decision must be taken rather than read off. Neither implies the other, so we diagnose the context along two continuous axes.

Atypicality. We measure how far the current state lies from recorded practice by the empirical CDF position of a ridge-regularised Mahalanobis distance,

$$D(s) = \sqrt{(s - \hat{\mu})^{\top}(\hat{\Sigma} + \varrho I)^{-1}(s - \hat{\mu})}, \qquad a(s) = \hat{F}_D(D(s))\,, \tag{3}$$

against a reference sample drawn once from the training split and thereafter frozen. Rank calibration places $a(s)$ on $[0, 1]$, so it reads as the fraction of precedent lying closer to the centre of the recorded state distribution.

Evidence disagreement. We measure the dispersion of the source means $\mu_j(s) = \mathbb{E}_{\pi_j}[a \mid s]$ over the sources that speak. Let $S_d(s) \subseteq \{0, \ldots, J\}$ denote the sources defined on action dimension d at state s, including routine practice and any applicable evidence source. We define

$$b_{\mathrm{raw}}(s) = \sqrt{\sum_{d=1}^{2} \frac{1}{|S_d(s)|} \sum_{j \in S_d(s)} \left(\frac{\mu_j(s)_d - \bar{\mu}_d(s)}{\varsigma_d}\right)^2}, \qquad b(s) = \hat{F}_b(b_{\mathrm{raw}}(s))\,, \tag{4}$$

where $\bar{\mu}_d(s)$ is the mean over $S_d(s)$ and ς_d is the training-set scale of dimension d. Routine practice is included because disagreement is defined relative to departmental practice as well as among the other sources. Restricting the calculation to sources that speak prevents guideline silence from being counted as agreement with routine practice.

The two signals are close to orthogonal on held-out states (Spearman $\rho = +0.13$). Neither signal is thresholded; the partition of their plane in Section 5.3 is purely presentational.

4.3 Arbitration as controlled deviation

When sources disagree, the mixture $\sum_j \alpha_j \pi_j$ is multimodal and its mean can fall where every source assigns low probability. We therefore represent combination as transport from routine practice toward each evidence source. For source j, a conditional velocity field carries samples from π_{behav} to π_j; writing $x_j(t \mid s)$ for the position at time t of a sample $a_0 \sim \pi_{\text{behav}}(\cdot \mid s)$, t measures how far the decision departs from routine practice toward that source. Conditional flow matching provides one realisation of this transport (Lipman et al., 2023). The conditional paths are linear, so the learned routes are close to straight lines (Appendix A.2); the commitment defended in Section 5.5 is to transport against averaging, not to the expressiveness of the field. The travelled distance is

$$t_j^*(s) = \begin{cases} \min\left\{ \arg\max_{t \in [0,1]} \widehat{V}(s, x_j(t \mid s)),\ a(s)\,\kappa_j \right\}, & \text{if a value signal is available,} \\ a(s)\,\kappa_j, & \text{otherwise,} \end{cases} \tag{5}$$

where $a(s)$ controls departure according to the decision context and κ_j is the source-specific ceiling fixed by evidence strength. On MIMIC-IV, where we introduce no counterfactual value estimate, the second case applies; the value-guided case is evaluated in the simulated environment of Section 5.4.

Transport and adoption are separated per action dimension. A route may be defined on every dimension, but it contributes to the decision only where its source speaks. Let $\mathcal{E}_d(s) \subseteq \{1, \dots, J\}$ denote the evidence sources speaking on dimension d. Their arbitration weights are

$$\alpha_{k,d}(s) = \operatorname*{softmax}_{k \in \mathcal{E}_d(s)} \ell_k(s), \qquad \ell_k(s) = \big(0.5 + b(s)\big) \log \rho_{g_k} - c(s)\nu_k(s), \tag{6}$$

where ρ_{g_k} is the prior associated with the source grade, $\nu_k(s)$ is the GRADE-weighted violation score of route k, and

$$c(s) = c_0 + (c_1 - c_0)a(s)(1 - b(s)), \qquad c_0 = 0.2,\ c_1 = 1.0.$$

The softmax is taken only over the speaking set, so a silent source receives zero weight on that dimension. The disagreement term modulates differences between the grade priors, while $c(s)$ increases the penalty for rule violations when the patient is atypical but the available evidence appears to agree. Appendix A.1 shows the prior structure confines the weights among speaking sources to a narrow band, so the results are insensitive to the exact prior values.

The arbitrated action is

$$\tilde{a}_d(s) = \sum_{j=1}^{J} \alpha_{j,d}(s)\, x_j(t_j^*(s) \mid s)_d. \tag{7}$$

It reduces to the routine action a_0 when no evidence source speaks or when all travelled distances vanish. Because silent sources are excluded from both the disagreement and arbitration speaking sets, adding a silent source leaves the diagnosed context and the decision unchanged, satisfying the anchoring requirement of Equation 2.

4.4 Evidence Strength and Clinical Projection

Evidence strength enters the representation through distinct mechanisms rather than a single scalar weight. The GRADE weight controls the precision of guideline recommendations, while the grade-specific ceiling κ_j limits how far a route may travel and the prior ρ_{g_j} controls how strongly a speaking source enters arbitration:

grade	strong	weak	best practice	none	data derived
κ_j	1.00	0.50	0.40	0.00	0.80
ρ_{g_j}	1.00	0.55	0.45	0.05	0.70

(8)

These quantities, like c_0 and c_1 of Equation 6, are fixed in advance rather than fitted to optimise downstream performance. The complete account of the four strength channels, including dose precision and rule selection, is given in Appendix A.3.

The final step enforces clinical admissibility. Let $\mathcal{C}(s)$ denote the set of actions satisfying the hard constraints applicable at state s. The final decision is

$$\hat{a}(s) = \Pi_{\mathcal{C}(s)}(\tilde{a}(s)) \,. \tag{9}$$

Rules admitted to $\mathcal{C}(s)$ are enforced as feasibility conditions; evidence strength instead governs the construction, permitted departure, and relative influence of the evidence routes. This separation prevents clinical constraints from being traded against the other components of the decision.

5 Experiments

5.1 Setup and evaluation stance

We study vasopressor and fluid dosing in MIMIC-IV (Johnson et al., 2023) over ICU stays with circulatory failure, discretised into four-hour windows and split by patient. The cohort targets haemodynamic compromise rather than imposing a Sepsis-3 infection criterion; full cohort construction is given in Appendix B. On MIMIC-IV we report no counterfactual estimate of outcomes under alternative doses and no off-policy value for the arbitrated policy, since the reliability of such estimates in this setting remains contested (Gottesman et al., 2019; Jeter et al., 2019). Our real-data claims are therefore descriptive or structural. Because each patient contributes multiple windows sharing one 90-day outcome, all reported proportions use stay-clustered bootstrap intervals.

5.2 Q1: Do the evidence sources actually disagree?

Arbitration is useful only if treatment decisions are genuinely contested. Among $2{,}000$ closely matched pairs of windows from different stays, clinicians disagreed about whether to give fluid in 13.9% of pairs (95% CI $[11.0, 17.0]$) and vasopressor in 19.2% ($[15.8, 23.0]$). The structure is action-specific: disagreement about fluid decreases with severity while disagreement about vasopressor increases, with a contrast of -20.5 percentage points ($[-32.8, -8.4]$). The two disagreement indicators are also nearly independent within matched pairs ($\phi = -0.023$), supporting the per-action representation of Section 4.2.

Across the represented evidence sources, 90-day mortality rises from 26.9% to 43.0% across disagreement quartiles even though the upper two quartiles share a median SOFA of 7.0. Of the 93 double-annotated SSC 2021 recommendation blocks, two yield state-conditional action rules, covering 22.7% of held-out fluid decisions (Figure A1; Appendices F.6 and C).

5.3 Q2: Does the diagnosis find states where routine practice should be challenged?

To describe the results in clinically readable terms, we partition the plane spanned by the two signals at the median of each axis and name the four regions *routine*, *known controversy*, *consensus extrapolation*, and *true edge*. The partition is presentational, since the framework applies no threshold to either signal. Ninety-day mortality rises from 20.7% in the routine region to 47.7% at the true edge (Figure 2), with the ordering preserved under stay-clustered intervals and at every partition threshold in $[0.3, 0.7]$. Across five reference seeds, the consensus-extrapolation region occupies $22.1 \pm 0.3\%$ of states at $36.2 \pm 0.5\%$ mortality (Appendices F.3 and F.4).

The consensus extrapolation region is the substantive finding. Compared with known controversy, these patients have lower SOFA (6.0 vs. 7.0), lower lactate and higher mean arterial pressure, receive fluid (63.8% vs. 92.6%) and vasopressors (11.9% vs. 38.2%) markedly less often, yet die 10.7 percentage points more frequently, 1.76 times the mortality of the routine region. This gap is not simply a restatement of recorded severity: after adjustment for the full state, membership in consensus extrapolation rather than known controversy carries an odds ratio of 1.49 (95% CI $[1.18, 1.86]$).

In the two low-disagreement regions the guideline never fires, speaking on none of the $11{,}188$ routine states and none of the $8{,}860$ consensus-extrapolation states, so their low disagreement reflects

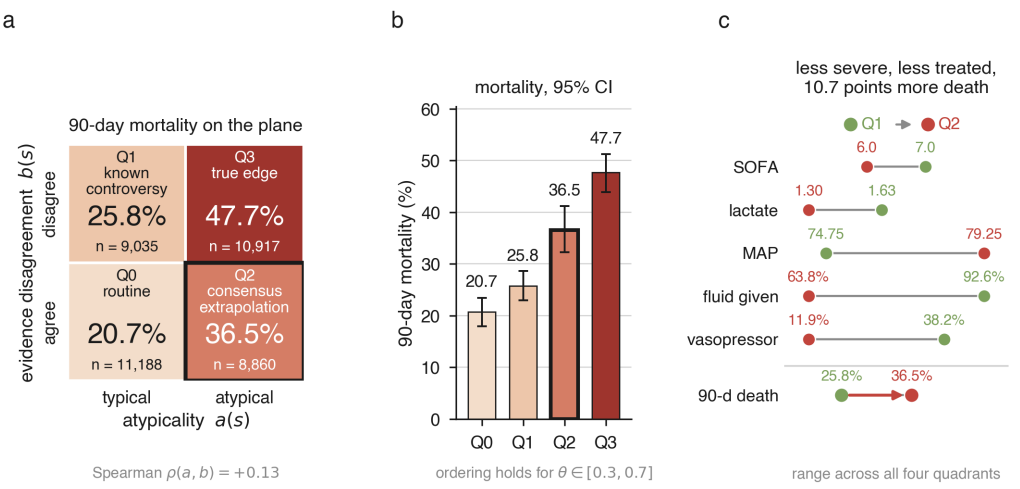


Figure 2: **The decision plane.** Atypicality $a(s)$ and evidence disagreement $b(s)$ are close to orthogonal (Spearman $\rho = +0.13$; $n = 40{,}000$ held-out windows). **(a)** Median splits yield four clinically readable regions with 90-day mortality overlaid. **(b)** Mortality across the four regions, with stay-clustered intervals. **(c)** Consensus-extrapolation patients appear less severe and receive less treatment than known-controversy patients, yet die 10.7 percentage points more frequently.

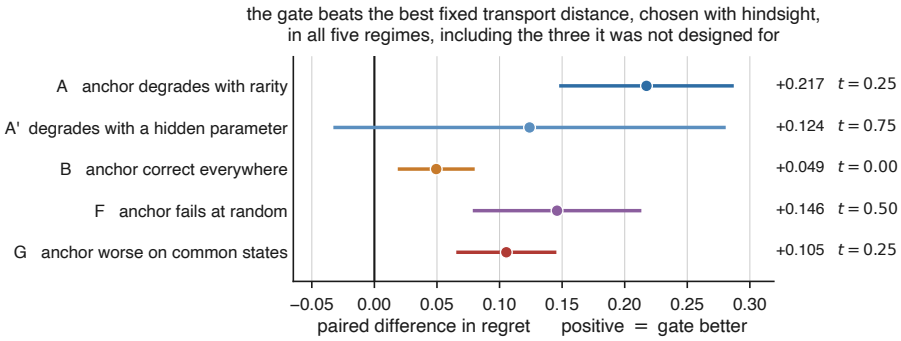


Figure 3: **The gate makes a situational judgement.** Paired difference in regret against the best fixed transport distance, chosen with hindsight in each regime; positive values mean the gate has lower regret. Regret is measured against the dose the optimal control proposes at the same state, executed the same way (Appendix E.3), over 200 simulated patients and 48 draws each, paired within patient. Regimes A and A′ follow the premise encoded by the context factor, B removes the need to depart, and F and G break or invert that premise. The full comparator set, with Wilcoxon p-values, is reported in Table A9.

concordance among the sources that speak rather than agreement with the guideline (Appendix F.7). The region is stable to alternative training of the optimised source, occupying 21.7% of states at 35.7% mortality under a naively trained variant despite a 2.4-fold change in the raw disagreement scale (Appendix F.5).

5.4 Q3: Does the gate move when it should and stay when it should not?

A gate that always departs and one that never departs can each pass only part of the intended test. We therefore evaluate five regimes in a simulated environment with known optimal control. The anchor degrades with rarity in A; degrades with an unobserved pharmacokinetic parameter rather than with $a(s)$ in A′; is correct everywhere in B; fails at random in F; and is worse on *common* states in G. Thus A directly instantiates the premise encoded by the context factor, while A′, B, F and G

progressively decouple, remove or reverse it, together testing whether the gate stays adaptive when atypicality is informative, uninformative or directionally misleading. This experiment exercises the value-guided branch of the gate.

Against the best fixed transport distance chosen with hindsight in each regime, the gate achieves lower paired regret in all five (Figure 3): by 0.217 in A (95% CI $[0.148, 0.287]$), 0.124 in A$'$, 0.049 in B, 0.146 in F, and 0.105 in G. The best fixed distance itself varies from $t = 0$ in B to $t = 0.75$ in A$'$, showing why no single departure distance serves all regimes. Removing the value signal increases regret in every regime (0.271, 0.253, 0.401, 0.209 and 0.327 in A, A$'$, B, F and G; all $p < 10^{-11}$). In regime B, where the anchor is correct, value search remains at the anchor while the value-free distance $a(s)\kappa_j$ is strictly positive, separating the context-scaled distance as a ceiling rather than a target.

Oracle comparisons clarify the role of that ceiling. Selecting the best source using ground truth approximately matches the gate when the anchor degrades and outperforms it when the anchor remains reliable; selecting a source using the same exact value signal improves regret by 0.272 in regime A. The comparison exposes what the transport ceiling costs under exact value information: bounding source influence can only restrict an otherwise correct choice, so the ceiling earns its place where value estimates are imperfect.

5.5 Q4: Why transport rather than averaging?

With two sources on opposite sides of the optimum, sampling from the mixture yields a return of -5.300, compared with -5.110 for transport at $t = 0.5$, a paired improvement of 0.190 (95% CI $[0.128, 0.252]$, $p = 1 \times 10^{-9}$). The within-patient dispersion of the sampled action is 1.176 for the mixture and 0.210 for transport, with 60.6% of mixture draws exhibiting a bimodal configuration. Yet the expected actions of the two methods coincide ($p = 0.73$), showing that the advantage is distributional rather than a shift in the mean action: transport avoids alternating between competing modes and yields substantially more concentrated decisions (Appendix F.8). On MIMIC-IV, the arbitrated action lies more than two nats below every route's density on only 3.3% of states (Appendix F.11).

5.6 Q5: Can a decision be attributed to its sources?

Exact reconstruction provides a first check: the arbitrated action is recovered from the recorded quantities to a residual of 2×10^{-16}, so no information required to reproduce the decision is lost. We test the stronger notion of attribution by intervening on the representation. With the anchor draw held fixed, adding an evidence source that is silent everywhere leaves the decision, disagreement signal, and region assignment exactly unchanged across $8{,}000$ states. The result is also practically relevant: the guideline is silent on 77.4% of fluid decisions and on every vasopressor decision. The learned field respects the mask as well: where the guideline speaks its route travels a mean of 386.7 mL and exceeds 100 mL on 87.7% of states, while where it is silent it travels a mean of 14.6 mL and exceeds 100 mL on 0.4%.

Source removal further exposes how attribution interacts with evidence coverage. Removing the guideline changes the fluid decision by 35.8 mL on average, and for that source its recorded weight predicts the movement to within 3×10^{-10}; where its weight is zero, the decision does not move (Figure A3, Appendix F.11). Removing the optimised source produces a larger mean movement of 95.6 mL, while a weight-only prediction has an error of 45.5 mL because the guideline is then the only remaining source and is silent on most states, causing the decision to fall back to the anchor. The intervention therefore makes both components of the representation visible: weights describe the influence of sources that speak, while coverage determines what evidence remains when a source is removed.

5.7 Q6: Do explicit clinical constraints matter?

Because evaluating the constraint channel requires outcomes under alternative actions, we test it in simulation. Across 400 patients, a constraint binds on 45.5% of decisions. Hard projection reduces regret from 0.148 to 0.112 and eliminates violations, whereas a soft penalty yields regret 0.145 and a violation rate of 35.6%, nearly unchanged from 35.9% without constraints. These results support

treating clinical admissibility as a final projection rather than another scalar penalty that can be traded against the other components of the decision (Appendix F.9).

6 DISCUSSION

Keeping evidence sources explicit changes what a clinical decision can record. The clearest example is silence: in the consensus-extrapolation region, the compiled guideline fires on none of the 8,860 states, so low disagreement reflects agreement among the sources that speak rather than three-way concordance. More broadly, only two of 93 guideline recommendation blocks compile into state-conditional action rules. Treating this missing coverage explicitly prevents absent guidance from being silently converted into a default opinion. The same representation makes the reliability of an optimised policy inspectable: propensity-weighted and naively trained variants change the scale of raw disagreement by a factor of 2.4 yet identify regions of similar size and mortality after rank calibration, while simulation shows that an exact-value oracle can outperform capped arbitration. PRUDENT therefore treats an optimised policy as evidence whose coverage and permitted influence can be examined, rather than as the objective against which all other evidence is judged. This is also what a conservative off-policy estimate would have to assume away. On this cohort such an estimate rests on an effective sample size of 0.27 (Appendix D.2), so it would resolve by fiat the very reliability the representation is built to keep visible.

The transport ceiling and value search serve distinct roles. Evidence strength limits how far a source may influence the decision, whereas a value signal, when available, selects how far to travel within that allowance. When routine practice is already correct, the value-free distance $a(s)\kappa_j$ permits departure, while value search can remain at the anchor. Separating permission from selection allows evidence strength to be specified independently of downstream optimisation and distinguishes a source that speaks but recommends little movement from one that is silent.

7 LIMITATIONS

Our findings on MIMIC-IV are observational rather than counterfactual. The mortality gradients identify clinically distinct regions of the recorded state space but do not establish what would have happened under different treatment. Unobserved factors, including goals of care, comorbidity burden and individual contraindications, may affect both treatment choice and prognosis. We therefore evaluate treatment improvement only in simulation, where optimal control is known; the MIMIC-IV results support structural and attribution claims.

The compiled guideline is sparse and action-asymmetric: only two of 93 annotated recommendation blocks yield state-conditional rules, covering 22.7% of held-out fluid decisions and providing no vasopressor-dose opinion. Some regions are therefore informed primarily by routine practice and the optimised source. PRUDENT represents this missing coverage explicitly, but broader guideline coverage and additional evidence sources are needed to test the generality of heterogeneous arbitration. MIMIC-IV and simulation exercise complementary components of the framework rather than serving as interchangeable evaluations.

Our treatment of silence is strict, but our treatment of firing range is not: a rule fires wherever its predicate holds, and whether that predicate correctly delimits the population the recommendation was written for is never asked. Appendix F.7 bounds this asymmetry; it does not reach the main finding, but it leaves an alternative reading of the true-edge region that the available data cannot exclude.

AI USE STATEMENT

The guideline compilation of Section 4.1 uses a large language model, and we state precisely where.

The gold standard is human and was produced first. The guideline was segmented programmatically into 93 blocks and GRADE levels were parsed from the source text; two annotators then labelled every block independently against a written handbook, the annotation workbook contains no model output of any kind, and neither annotator saw a model proposal at any point. Seven genuine dis-

agreements were adjudicated and the resulting gold standard was frozen before any compilation was run.

Claude Sonnet was then prompted with few shot examples to compile the same 93 blocks independently. Its output is scored against the frozen gold standard (Appendix C) and supplies the predicates and action specifications used downstream. Two departures from a mechanical reading are recorded in Appendix C.3: a lactate condition introduced through the prompt, and one predicate condition removed by hand in the constraint channel.

The same model family assisted with drafting and editing prose in this manuscript and with writing analysis and plotting code; all such code was read by the authors and every reported number was regenerated from the final code base. Generative AI tools were not used to generate synthetic data, to develop the framework or its operator requirements, to formulate or refine the hypotheses under test, to design the experiments or the simulated environment, to implement the methods themselves, or to interpret the results. Translation and dataset cleaning are not applicable to this work. All numerical results, statistical claims and interpretations are the authors' responsibility and were verified against the result files behind every figure. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses MIMIC-IV (Johnson et al., 2023), a de-identified critical care database distributed under credentialed access. We accessed it under the applicable data use agreement and made no attempt to re-identify individuals or to link records to external sources. No new patient data were collected and no intervention was administered.

The findings reported on MIMIC-IV are observational. The mortality gradients across the decision plane describe regions of the recorded state space; they do not establish what would have happened under different treatment, and the elevated mortality of the consensus-extrapolation region must not be read as evidence that those patients should have received more fluid or vasopressor. We report no counterfactual estimate on real data for this reason, and evaluate whether controlled departure improves decisions only in simulation, where the optimal control is known.

PRUDENT is a representation of how heterogeneous evidence composes into a decision, not a deployable decision support system. Its compiled guideline covers two recommendation blocks and offers no vasopressor-dose opinion, its optimised source is trained on one hospital system, so the size and boundaries of the regions it identifies may reflect local practice. Unobserved factors including goals of care and individual contraindications are also absent from the state. Any clinical use would require prospective validation, broader evidence coverage and audit of the optimised source.

REPRODUCIBILITY STATEMENT

The cohort criterion, the split and the window construction are given in Appendix B, with per-stage counts in Table A1; that appendix also specifies the state features and their normalisation groups, the frozen context reference samples and action scales, the subsampling convention under which counts should be compared, and the action space bounds. Architectures, optimisers, hyperparameters, stopping epochs and divergence criteria for all three learned components are in Appendix D. The guideline distribution, the annotation protocol with its inter-annotator agreement, the two compiled rules and the two recorded departures from a mechanical reading are in Appendix C, with the rules themselves in Table A4 and per-source coverage in Table A3. The two bootstrap procedures behind every reported interval are specified in Appendix A.4. Appendix G records the seeding discipline required to reproduce the traced decisions and lists the artefacts that will be released upon publication.

REFERENCES

Floris den Hengst, Martijn Otten, Paul Elbers, Frank van Harmelen, Vincent François-Lavet, and Mark Hoogendoorn. Guideline-informed reinforcement learning for mechanical ventilation in

critical care. *Artificial Intelligence in Medicine*, 147:102742, 2024. doi: 10.1016/j.artmed.2023.102742.

Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M. Coopersmith, Craig French, Flávia R. Machado, Lauralyn McIntyre, Marlies Ostermann, Hallie C. Prescott, et al. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021. *Critical Care Medicine*, 49(11):e1063–e1143, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062. PMLR, 2019.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019.

Gordon H. Guyatt, Andrew D. Oxman, Gunn E. Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, and Holger J. Schünemann. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650):924–926, 2008.

Hsing-Yu Hsu, Lu-Wen Chen, Wan-Tseng Hsu, Yow-Wen Hsieh, and Shih-Sheng Chang. Extracting clinical guideline information using two large language models: Evaluation study. *Journal of Medical Internet Research*, 27:e73486, 2025. doi: 10.2196/73486.

Russell Jeter, Christopher Josef, Supreeth Shashikumar, and Shamim Nemati. Does the artificial intelligence clinician learn optimal treatment strategies for sepsis in intensive care? *arXiv preprint arXiv:1902.03271*, 2019.

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):1, 2023. doi: 10.1038/s41597-022-01899-x.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Xuefeng Peng, Yi Ding, David Wihl, Omer Gottesman, Matthieu Komorowski, Li-wei H. Lehman, Andrew Ross, Aldo Faisal, and Finale Doshi-Velez. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In *AMIA Annual Symposium Proceedings*, 2018.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference (MLHC)*, 2017.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

David L. Sackett, William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes, and W. Scott Richardson. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72, 1996.

Suchi Saria. Individualized sepsis treatment using reinforcement learning. *Nature Medicine*, 24(11): 1641–1642, 2018.

Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):801–810, 2016.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. Expert Certification.

Yuan Wang, Anqi Liu, Jucheng Yang, Lin Wang, Ning Xiong, Yisong Cheng, and Qin Wu. Clinical knowledge-guided deep reinforcement learning for sepsis antibiotic dosing recommendations. *Artificial Intelligence in Medicine*, 150:102811, 2024. doi: 10.1016/j.artmed.2024.102811.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

A Mathematical detail

A.1 How much the arbitration weights vary, and why

The prior ρ_{g_k} of Equation 6 does not depend on the state, so all state dependence in the weights enters through admission to the speaking set $\mathcal{E}_d(s)$, through the multiplier $0.5 + b(s)$, and through the violation term. Per-dimension masking makes the first of these dominant, and the resulting structure has two levels of very different size.

The large level is admission. A source silent on a dimension receives weight exactly zero there. The guideline is silent on every vasopressor decision, so the weights on that dimension are $[0, 1]$ at every state, and it is silent on 77.4% of fluid decisions. Between regions the weights differ by 0.116, and essentially all of that is the difference between speaking and not speaking.

The small level is the multiplier. Among states where both sources speak, the multiplier can only scale a gap the priors already establish, and cannot create one. With the two grades used here the gap is $\log 0.70 - \log 0.55 = 0.241$, so as $b(s)$ ranges over $[0, 1]$ the logit difference is confined to $[0.121, 0.362]$ and the weight of the data derived source to $[0.530, 0.590]$. Conditional on the guideline speaking, the measured weights are 0.418 for the guideline and 0.582 for the optimised source, with a standard deviation of 0.0043 across states. The measured pair corresponds to a logit difference of 0.331, which the multiplier would produce at an equivalent $b(s)$ of 0.87. Because the softmax is not linear this is an equivalent value rather than a conditional mean, but it is consistent with disagreement being high on precisely those states where the guideline speaks, since a fired target sits near the action ceiling and far from routine practice.

Reported over all states rather than conditionally, the guideline's mean weight on the fluid dimension is 0.0916, which is 0.418 multiplied by the firing rate. That figure measures coverage rather than attribution and should not be read as the guideline receiving nine percent of the decision where it speaks.

Making the second level genuinely state dependent would require the prior to enter additively rather than multiplicatively, or a precision term that varies across states. We report the present form and its consequence. Read as a sensitivity analysis, the closed form above is stronger than an empirical sweep: for any prior values with the same ordering, the weight of the data-derived source is confined

to an interval whose width is set by the log-ratio of the priors, here $[0.530, 0.590]$. Perturbing ρ_{g_j} within its grade ordering therefore cannot change which source dominates a dimension, only the margin; the quantity the results do depend on is admission to the speaking set, which is determined by coverage rather than by any fitted parameter.

A.2 Recovering the guideline dose target from route geometry

Under the linear conditional paths of the transport construction the route endpoint satisfies $x_j(t_j^*) = a_0 + t_j^*(\bar{a}_j - a_0)$, so the target the guideline proposed can be recovered as $\bar{a}_j = a_0 + (x_j(t_j^*) - a_0)/t_j^*$ wherever $t_j^* > 0$. This inversion assumes the learned field transports along a straight line; the agreement reported below is therefore also a check of that assumption.

Applying it to the fired states gives a distribution with no point mass at the action ceiling, 1.1% of targets lying within one percent of it, and a standard deviation of 0.107 in normalised units. A deterministic target of $\min(30 \cdot \text{weight}, \text{ceiling})$ would place every patient above $67.9\,\text{kg}$ exactly at the ceiling, producing a point mass there. The observed spread is instead consistent with sampling from the truncated guideline distribution, whose standard deviation at the weak grade is $\sigma_0(1 - \lambda w_{\text{weak}}) = 0.102$. Roughly one quarter of windows carry a weight below $67.9\,\text{kg}$, and on those the recovered target does vary with the patient; on the remainder the dose channel sits at its ceiling and conveys no individual information. The agreement between the recovered spread (0.107) and the theoretical value under linear transport (0.102) serves a second purpose: had the learned field bent the routes appreciably, the linear inversion would not recover the truncated guideline distribution this closely. The routes the field actually learns are therefore close to straight lines, which is why the claims of this paper are stated for transport against averaging (Section 5.5) rather than for the expressiveness of the field; conditional flow matching is retained because it provides a single state-conditional construction that applies uniformly to any source distribution, including multimodal or constrained sources for which a straight line between means is not defined.

A.3 The Four Channels of Evidence Strength

Strength enters at four places with different mathematical characters.

Dose precision, continuous. In Equation 13, $\sigma_k = \sigma_0(1 - \lambda w_k)$ makes a strongly graded rule contribute a narrower component and therefore a larger share where several rules bear on the same dimension. The same continuous weight w_k is reused in the violation term ν_j, so breaches of stronger rules count more when routes are compared.

Transport ceiling, discrete and hard. The grade-specific κ_j bounds how far the decision may travel toward source j and is fixed independently of the observed state.

Attribution prior, discrete and soft. The prior ρ_{g_j} determines the source's contribution to the arbitration logit before state-dependent disagreement and rule violations are taken into account. The ceiling and prior are deliberately separate: a zero ceiling forbids movement, whereas a small non-zero prior leaves a source available to arbitration when it speaks. Sources without a GRADE level use the data-derived category; its placement between weak and strong evidence is a pre-specified modelling choice.

Rule selection, binary and hard. Restricting the compiled rule set by grade yields a stratified policy for which recommendations are admitted. In the present guideline set this mechanism cannot be meaningfully ablated because the two compilable rules occupy different grades and the strong rule contributes a constraint rather than a dose recommendation.

The violation score. Each rule binding at s defines a feasible set on the action space. The violation score of route j is the sum of the GRADE weights of the rules its endpoint breaches,

$$\nu_j(s) = \sum_{k \in \mathcal{K}(s)} w_k \mathbb{1}\big[\, x_j(t_j^* \mid s) \notin \mathcal{C}_k(s) \,\big], \tag{10}$$

where $x_j(t_j^* \mid s)$ is the mean of that route's endpoint over the n draws. Violation is binary rather than graded by distance: the compiled constraints state whether an action dimension is zero, not by how much it departs, so there is no displacement to measure. The score is evaluated once at the endpoint

rather than integrated along the path, and contributions from several binding rules are summed, so with the two compiled rules it takes values in $\{0,\ 0.40,\ 0.85,\ 1.25\}$.

Two properties of this term should be stated. It enters only the soft penalty of Equation 6, where it expresses a preference among routes that are all feasible; feasibility itself is enforced by the projection of Equation 9 and is never left to ν. And it is currently a single scalar per route, aggregated across action dimensions, whereas the constraints are themselves per dimension. A finer form $\nu_{j,d}$ is the natural generalisation; the two coincide in the configuration reported here, since the soft penalty is not enabled on MIMIC-IV and $\nu \equiv 0$ there.

The GRADE weight. The continuous weight w is indexed by strength *and* quality, unlike the ceiling and prior of Equation 8, which are indexed by strength alone:

	high	moderate	low	very low
strong	1.00	0.85	0.70	0.60
weak	0.65	0.55	0.40	0.30

(11)

with 0.50 for a best-practice statement carrying no quality rating and 0.30 where no level can be matched. The two compiled rules occupy two of these cells: SSC0005 is weak with low quality ($w = 0.40$) and SSC0009 is strong with moderate quality ($w = 0.85$). Through $\sigma_k = \sigma_0(1 - \lambda w_k)$ with $\sigma_0 = 0.15$ and $\lambda = 0.8$ these give component widths of 0.102 and 0.048; the first is the value recovered from route geometry in Appendix A.2.

Keeping the two tables separate is deliberate. The weight w expresses how much a rule should concentrate the dose it proposes and how much its breach should count, which depends on the quality of the evidence behind it as well as on the strength of the recommendation. The ceiling and prior express how far the decision may travel toward a source and how loudly it speaks in arbitration, which we tie to strength alone.

A.4 Interval construction

Three figures carry intervals and two different bootstraps produce them, because the dependence structures differ.

Window-level proportions use a one-way stay bootstrap. Mortality by region and by disagreement quartile is a Bernoulli observation per decision window, and a patient contributes about eleven windows that share a single 90-day outcome. The intraclass correlation is therefore 1.0 and the design effect 11.1, so we resample stays with replacement, $2{,}000$ times. The resulting intervals are between three and four and a half times wider than the Wilson intervals (Wilson, 1927) they replace, the ratio being largest in the consensus-extrapolation region, which has the fewest distinct patients per window of any region.

Matched pairs use a stay-level node bootstrap. A matched pair is different: its two windows come from different patients by construction, so the pair is internally independent, but one patient appears in 2.9 pairs on average and can appear at either end of different pairs. The dependence is therefore dyadic rather than nested. We resample stays and give a pair (A, B) the weight $c_A c_B$, where c counts how many times that stay was drawn. This widens the intervals by roughly a factor of two relative to treating pairs as independent.

We do not use a likelihood ratio test in place of either bootstrap, because its reference distribution assumes independent observations and would reintroduce the assumption being corrected.

Matching cannot be reconstructed from the printed parameters. An early version of the pair analysis rebuilt the matching rather than loading it, and obtained 723 pairs instead of $2{,}000$ with a vasopressor disagreement rate of 22.7% instead of 19.2%, enough to support a different conclusion. The cause is that which query rows survive the same-stay exclusion changes the distance distribution and hence the percentile threshold. We therefore store the pair indices with the results, and the analysis loads them; they will be included in the release.

How the pairs were matched. For each query window in the test split we retrieve the eight nearest neighbours under Euclidean distance on the full 47-dimensional standardised state, exclude every candidate from the same stay, and keep the closest survivor. A pair is admitted when that distance

Table A1: **Cohort construction.** Counts are stays and, where given, four hour decision windows. The split is by `subject_id` at 70/15/15 with seed 0, so no patient contributes to more than one split.

	Stays	Windows
Adult first ICU admissions	65,366	
less non-positive length of stay	−12	
After construction	65,354	1,414,413
less stays with no rows after the action shift	−276	
After the action shift	65,078	1,349,059
After the cohort criterion	24,777	774,213
train	17,343	545,406
validation	3,717	114,960
test	3,717	113,847

falls below the tenth percentile of the first-neighbour distance distribution, which is 2.978 here, and the first $2{,}000$ admitted pairs are retained. The retrieval is exact rather than an approximate index.

The five variables reported in the matching-quality summary — mean arterial pressure, lactate, SOFA, heart rate and creatinine — are used to *check* that matched windows are clinically comparable. They do not enter the distance, which uses all 47 dimensions.

The percentile threshold is computed on the distance distribution that survives the same-stay exclusion, so which query rows survive that exclusion changes the threshold. This is why the matching cannot be rebuilt from the printed parameters and why the pair indices are stored with the results.

B Dataset and pipeline

B.1 Cohort construction

We use MIMIC-IV (Johnson et al., 2023), restricted to adult first ICU admissions and discretised into four hour decision windows.

The cohort criterion is

$$\text{SOFA} \geq 2 \quad \text{and} \quad \big(\text{Lactate} > 2 \ \text{ or } \ \texttt{max_dose_vaso} > 0\big),$$

applied at the stay level. We do not call this a Sepsis-3 cohort. Sepsis-3 (Singer et al., 2016) requires suspected or documented infection together with an acute rise in SOFA of at least two points, and our criterion contains no infection component and uses an absolute rather than a relative SOFA threshold. What it selects is patients with organ dysfunction accompanied by circulatory compromise, whether or not that compromise is infectious in origin. We adopt it because the decision problem studied here, namely fluid and vasopressor dosing, is defined by circulatory compromise rather than by the presence of infection, and because infection adjudication in MIMIC-IV requires choices that would themselves need defending. The consequence is a broader and less severely ill population than a Sepsis-3 cohort: 90 day mortality is 27.0% in the training split against 18.5% in all adult first admissions. Results should be read as concerning circulatory failure in the ICU.

B.2 Features and normalisation

The state has 47 dimensions, comprising 32 features standardised linearly, 11 standardised after the transform $x \mapsto \log(0.1 + x)$, and 4 binary features centred at zero. Scalers are fitted on the training split alone.

Which group a feature belongs to matters for the guideline dose channel, which converts a mL/kg rule using the inverse transform of patient weight. Weight is in the linear group. A log group feature given a linear inverse transform returns values that remain clinically plausible while being wrong, and no range check would catch it.

Table A2 lists the features in each group.

Table A2: **The state.** The 47 dimensions by normalisation group. Linear features are standardised directly; log-group features are standardised after $x \mapsto \log(0.1 + x)$; binary features are centred at zero. Scalers are fitted on the training split alone. Group membership is not cosmetic: the guideline dose channel inverts the transform of patient weight to convert a mL/kg rule, and a log-group feature given a linear inverse transform returns values that remain clinically plausible while being wrong.

group	features
linear (32)	age, weight, GCS, heart rate, systolic BP, mean BP, diastolic BP, respiratory rate, temperature, FiO_2, potassium, sodium, chloride, glucose, magnesium, calcium, haemoglobin, white cell count, platelets, PTT, PT, pH, PaO_2, $PaCO_2$, base excess, bicarbonate, lactate, SOFA, SIRS, shock index, PaO_2/FiO_2, cumulative fluid balance
log (11)	SpO_2, urea nitrogen, creatinine, AST, ALT, total bilirubin, INR, cumulative input, input over the previous four hours, cumulative output, urine over the previous window
binary (4)	sex, mechanical ventilation, vasopressor administered, readmission

B.3 Context references

Atypicality uses a reference sample of $N = 100{,}000$ training states, from which $\hat{\mu}$ and $\hat{\Sigma}$ of Equation 3 are estimated with ridge parameter $\varrho = 10^{-3}$. Disagreement uses a reference sample of $20{,}000$ training states and an action scale $\varsigma = (0.17985, 0.17207)$ estimated on the training split. Both samples are stored as row indices and frozen, and every component loads the same reference, so the context diagnosis is identical wherever it is computed.

An earlier version of the pipeline fitted these references independently in each analysis, which produced region assignments that differed between analyses. The loader now raises an error if the source names or the reference size do not match what was frozen.

B.4 Subsampling

Analyses draw their own subsample of the test split, which has $113{,}847$ rows, using a common seed. The gate probe and the inter source disagreement analysis draw $40{,}000$ windows, the displacement analysis $20{,}000$, and the reconstruction analysis $8{,}000$. Subsampling is used because the Mahalanobis evaluation and the nearest neighbour retrieval scale with n, while $40{,}000$ already leaves between $8{,}800$ and $11{,}200$ windows in each region. Because the seed is shared, regional mortality agrees across analyses, but the counts do not: analyses drawing different n obviously differ, and two runs drawing the same n at different points in the stream differ by a few windows per region as well. Counts should therefore be compared within one run rather than across runs, and those quoted in the main text are from the run that produced Figure A4.

B.5 Action space

Actions are the intravenous fluid volume and the maximum vasopressor dose over the window, scaled to $[0, 1]$ by $\ell = (0, 0)$ and $u = (2036.37\ \mathrm{mL}, 0.6159\ \mathrm{mcg/kg/min})$, the 99th percentile of each in the training split. These are percentiles of observed practice rather than clipping limits imposed for numerical convenience, and the distinction matters when reading the guideline dose channel, whose fluid target exceeds this ceiling for any patient above 67.9 kg.

Normalised actions are clipped to $[0, 1]$ rather than dropped, so the cohort counts of Table A1 are unaffected and exactly 1.000% of windows sit at the upper boundary on each dimension. Scaling by a percentile rather than by the observed maximum is what makes this workable: the largest recorded four-hour fluid volume is $37{,}187$ mL and the largest recorded vasopressor dose is 359.7 mcg/kg/min, neither of which is physiologically possible, and either would compress the entire usable range into the bottom percent of the action space. For reference, 19.44% of windows record no fluid and 73.38% record no vasopressor, and the medians among the remainder are 200.0 mL and 0.1001 mcg/kg/min.

Table A3: **Evidence sources as decision entities.** Each source carries its own grade and is transported from routine practice by at most κ_j. Routine practice is the anchor, so no cap applies to it. The guideline is the only source that can be silent, and it is silent on one action dimension entirely. Dose coverage is the fraction of held-out windows on which the compiled predicates fire, and it is what determines the per-dimension arbitration weights of Equation 6. The last column gives the fraction of states on which each source proposes the most extreme of the three actions.

				Dose coverage (%)		Most extreme
Source	Role in the decision	Grade	κ_j	Fluid	Vasopr.	source (%)
π_{behav}	routine clinical practice (anchor)	—	—	always defined		0.61
π_{guide}	compiled SSC 2021 recommendations	weak	0.50	22.65	0.00	21.46
π_{rl}	offline RL (IPW-IQL), one evidence source	data	0.80	always defined		77.93

One consequence of that ceiling is worth recording. Patient weight carries no range check, taking values from 26.5 to 344.3 kg, and it enters the guideline dose channel directly through the mL/kg conversion. Its effect is bounded by the ceiling: 30 mL/kg exceeds 2,036 mL for any patient above 67.9 kg, and the cohort median weight is 78 kg, so an extreme weight and a typical one yield the same truncated target. In the test split 0.57% of windows carry a weight outside 25 to 250 kg. The check should still be added, but at this action scale it would not change any dose target.

B.6 Reward

The reward retains the intermediate SOFA and lactate terms of Raghu et al. (2017) together with a terminal survival term. Writing $\Delta\text{SOFA} = \text{SOFA}_{t+1} - \text{SOFA}_t$ and $\Delta\text{lac} = \text{lac}_{t+1} - \text{lac}_t$,

$$r_t = C_0 \mathbb{1}\big[\Delta\text{SOFA} = 0 \wedge \text{SOFA}_t > 0\big] + C_1 \Delta\text{SOFA} + C_2 \tanh(\Delta\text{lac}) + \mathbb{1}[t = T] \cdot \rho\,(1 - 2y_{90}), \tag{12}$$

with $C_0 = -0.025$, $C_1 = -0.125$, $C_2 = -2.0$, $\rho = 15$, and $y_{90} = 1$ for death within 90 days. The AI Clinician's purely terminal reward (Komorowski et al., 2018) is implemented as an alternative but is not used for any result reported here.

Two properties should be recorded. The known failure mode of such a reward is that it can credit a change in a score rather than a change in the patient; this is one reason the optimised source enters the representation as evidence with a bounded ceiling rather than as the objective the decision is judged against. And where SOFA or lactate is missing at t or $t+1$ the shaping terms are set to zero, which affects 67,587 of 1,349,059 windows (5.0%). The shaping signal is therefore biased toward zero exactly where measurement is sparse.

C Evidence sources and guideline compilation

C.1 Guideline Distribution

For action dimension d, let $K_d(s)$ denote the compiled guideline rules whose predicates hold at state s and that specify an action on that dimension. The guideline dose distribution is the precision-weighted product of the applicable rule components,

$$\pi_{\text{guide}}(a_d \mid s) \propto \prod_{k \in K_d(s)} \mathcal{N}\big(a_d; \bar{a}_k(s), \sigma_k^2\big)\,, \qquad \sigma_k = \sigma_0(1 - \lambda w_k), \tag{13}$$

where $\sigma_0 = 0.15$, $\lambda = 0.8$, and $w_k \in [0, 1]$ is the GRADE weight of rule k. Stronger evidence therefore yields a narrower component and a larger share of the product. If $K_d(s)$ is empty, the distribution falls back to π_{behav} on that dimension for route construction; the firing mask separately records that the guideline is silent and excludes it from downstream disagreement and arbitration.

C.2 Annotation protocol

The 2021 Surviving Sepsis Campaign guideline (Evans et al., 2021) was segmented programmatically into 93 recommendation blocks, with GRADE levels parsed from the source text rather than

Table A4: **The two rules that compile.** Predicates and action specifications as used. The lactate condition is discussed below; it does not appear in either source recommendation.

ID	GRADE	Predicate	Action
SSC0005	weak, low quality	$\texttt{age} \geq 18 \wedge \texttt{Lactate} > 2$	`input_4h`, increase, 30 mL/kg
SSC0009	strong, moderate	$\texttt{age} \geq 18 \wedge \texttt{Lactate} > 2$ $\wedge\, \texttt{MeanBP} < 65$	`max_dose_vaso`, target 65 mmHg on `MeanBP`

judged. Two annotators then labelled every block independently against a twelve page handbook containing worked examples and decision flowcharts, recording whether the block is specifiable as a state conditional action, the reason where it is not, the clinical and operational predicates, the action variable, direction and specification, and a confidence. The workbook contains human fields only.

Agreement is reported per column as unweighted Cohen's κ, with blanks treated as a distinct label: $\kappa = 0.795$ on compilability (agreement 98.9%), 0.873 on the reason for non-compilability (92.5%), and 0.795 on whether the parsed GRADE level is correct (98.9%). Seven genuine disagreements were adjudicated, comprising one on compilability, one on the GRADE check, and five on the reason field. These are agreement figures before adjudication; afterwards agreement is complete by construction and carries no information.

We do not report κ on the action columns. Its nominal value there is 0.021, which is an artefact: one annotator wrote `none` on non-compilable rows while the other left them blank, so the coefficient is measuring a filling convention. On the two rows both annotators judged compilable, the action labels agree completely, and κ on $n = 2$ carries no statistical content. A field level long form of the disagreement record contains 381 rows spanning all 93 blocks, but the overwhelming majority are this convention difference and free text notes.

C.3 The two compiled rules

Two departures from a purely mechanical reading of the guideline must be recorded, since both affect coverage and neither is visible in the compiled rule set itself.

First, the condition $\texttt{Lactate} > 2$ appears in neither source recommendation. It enters through the compilation prompt, which instructs that septic shock implies elevated lactate and a vasopressor requirement. It is therefore a modelling assumption imported into the predicate rather than a condition the guideline states, and it is one source of the coverage figures reported in the main text.

Second, SSC0009 as compiled carries the additional condition $\texttt{max_dose_vaso} > 0$, which we removed by hand in the constraint channel. Retained, it turns a rule about initiating vasopressors into one that can only be read as a rule about withdrawing them, and the constraint binds on 1.88% of decisions. Removed, it binds on 3.91% and the number of recorded violations rises from 38 to 427. This is a human modification of a compiled output and we state it rather than presenting the rule set as untouched.

C.4 Independent evaluation of the compiler

Because the gold standard was completed and frozen before any compilation was run, the compiler can be scored against it. Claude Sonnet, prompted with few shot examples, reproduced the human compilability judgement on all 93 blocks, with specificity 1.000 on the 91 blocks the annotators judged non-compilable. The low compilation rate is therefore a property of the guideline text rather than of extraction failure, which is the alternative explanation a reader would otherwise be entitled to.

The compiler was run with the `claude-sonnet-4-5` model alias at temperature 0 and a $1{,}200$ token limit, with six few-shot examples in a fixed prompt. All 93 blocks compiled, none failed, and 83 required a model call; the remainder were served from a cache of previously compiled blocks, so the rule set consumed downstream is a fixed artefact rather than one regenerated per run. The dated snapshot behind the model alias was not recorded, so an exact rerun is not guaranteed; the compiled output is released with the frozen guideline source and its hash.

Table A5: **Coverage of the compiled guideline.** **(a)** The two compiled rules cover one action dimension; the guideline offers no dose opinion on vasopressors at all. **(b)** Firing is concentrated almost entirely in the high-disagreement half of the decision plane. Panel (b) is computed on a subsample of $8{,}000$ windows rather than the $40{,}000$ of panel (a), so its rates differ slightly from the full-sample rates quoted in Appendix F.7 and its counts are not comparable with panel (a). On the full $40{,}000$ no rule fires on any Q0 or Q2 state, so the single Q2 firing recorded here is an artefact of the smaller draw. The implication is one-directional: firing almost always produces disagreement, because a fired target sits near the action ceiling and far from routine practice, but 54.6% of high-disagreement states are not fired.

(a) Firing by action dimension, held-out states

	n	fired	%
IV fluid (`input_4h`)	40,000	9,061	22.65
Vasopressor (`max_dose_vaso`)	40,000	0	0.00

(b) Firing by region of the decision plane

	n	fired	%
Q0 routine	2,301	0	0.0
Q1 known controversy	1,781	683	38.4
Q2 consensus extrapolation	1,760	1	0.1
Q3 true edge	2,158	1,070	49.6
All	8,000	1,754	21.9

The six few-shot examples are hand-written rather than drawn from the 93 blocks, so no block appears in the prompt that judges it and the evaluation set contains no example. They comprise two compilable cases and four non-compilable ones, the latter covering all four categories of reason recorded in the annotation schema.

C.5 Firing behaviour

Table A5 gives firing rates by action dimension and by region. The concentration in the high disagreement half is a consequence of the dose channel rather than a coincidence: a fired fluid target sits near the action ceiling, far above routine practice, so the source means separate and b_{raw} rises. The converse does not hold, since 54.6% of high disagreement states are not fired.

D Training details

D.1 Behavioural cloning

An MLP maps the 47 dimensional state through two hidden layers of 256 units to a softplus mean with a learned homoscedastic $\log\sigma$, trained under a Gaussian negative log likelihood with Adam at learning rate 10^{-4} and batch size 256, for at most 40 epochs with early stopping on validation likelihood and patience 3. Training stopped at epoch 11. Seed 0.

Early stopping uses likelihood rather than mean squared error deliberately. Overfitting first appears as a shrinking σ, that is as a model becoming overconfident, and π_{behav} defines the anchor distribution and the starting points of all transport routes.

D.2 Offline reinforcement learning

A LayerNorm followed by a GRU encodes a window of four states, truncated at stay boundaries, into 64 dimensions, with an auxiliary decoder predicting the next state at loss weight 0.1. The behaviour model is zero inflated, factorising each action dimension as $\pi_\beta(a_d \mid s) = (1-p_d)^{\mathbb{1}[a_d=0]}\big(p_d\,\mathcal{N}(a_d; m_d, v_d)\big)^{\mathbb{1}[a_d>0]}$ with p_d clipped to $[10^{-3}, 1-10^{-3}]$ and $\log\sigma \in [-3, 2]$.

The zero inflation is not cosmetic. Under a plain Gaussian the fitted σ collapses to its floor and the density at $a = 0$ reaches roughly 3400, so $1/\pi_\beta$ overflows and the effective sample size falls to 0.094 even with weights clipped at 10. With zero inflation it is 0.27.

Implicit Q learning (Kostrikov et al., 2022) follows, with expectile $\tau = 0.8$, advantage weighted regression coefficient $\beta = 3.0$ with the exponential capped at 100, discount $\gamma = 0.99$, Adam at 3×10^{-4}, batch size 256, and target network updates at rate 0.005. Importance weights $1/\pi_\beta$ are clipped to $[0.1, 10]$ and then self normalised, in that order. The behaviour stage runs to at most 30 epochs with patience 3, stopping at epoch 7 and retaining epoch 4; the policy stage runs a fixed 10^5 gradient steps without early stopping. Five seeds are run per configuration. A run is declared divergent, and excluded, if $|Q| > 60$, if more than 5% of actions fall outside the support, or if the effective sample size drops below 0.10.

The anchor of Appendix D.1 is fitted with a plain Gaussian while the behaviour model here is zero inflated, and the two serve different purposes. The behaviour model appears in the denominator of an importance weight, so its density at zero controls whether the weights are finite at all, and a Gaussian that cannot represent the point mass drives the effective sample size to 0.094. The anchor is never inverted. It supplies the origin of every transport route and the draws that begin them, so what matters is that its mean and dispersion describe customary care; a mis-specified probability of a zero dose shifts the starting point of routes on those states rather than destabilising an estimator. The asymmetry is deliberate but it is an approximation, and a zero-inflated anchor would represent the withhold-or-give decision explicitly.

D.3 Flow matching

The velocity field takes the action, a 32 dimensional sinusoidal embedding of t at logarithmically spaced frequencies, a 32 dimensional embedding of the source pair, and the full 47 dimensional state, through three hidden layers of 256 units with SiLU activations, to a two dimensional output, for $161{,}346$ parameters. Endpoints are cached as $40{,}000$ training states by eight draws by two source pairs, giving $640{,}000$ pairs. Adam at 3×10^{-4}, batch size 1024, at most 40 epochs with patience 4 against a ten percent holdout; training stopped at epoch 39 with a held out mean squared error of 0.02358.

Conditional flow matching is a regression objective, so no integration occurs during training. At inference the field is integrated with RK4 in 20 steps, with a per sample step size $h_i = t_i^*/20$. This is the default for every inference path. An earlier implementation averaged t^* across the batch before integrating, which silently disabled the gate; with per sample step sizes the integrator reproduces linear interpolation on a constant field to within 2×10^{-15}.

E The simulated environment

Sections 5.4, 5.5 and 5.7 measure regret in a simulated environment with a known optimal control. This appendix specifies that environment, the sense in which the optimal control is known, the scoring protocol, and the construction of the five regimes. Parameters are hard-coded in the environment module; there is no separate configuration file.

E.1 Dynamics

The environment is a two-compartment pharmacokinetic model coupled to a pharmacodynamic disease marker. With C_1 the central compartment concentration, C_2 the peripheral compartment, D the disease marker and $u \geq 0$ the dose rate,

$$\frac{\mathrm{d}C_1}{\mathrm{d}t} = -\Big(\frac{\mathrm{CL}}{V_1} + \frac{Q}{V_1}\Big)C_1 + \frac{Q}{V_2}C_2 + \frac{u}{V_1}, \tag{14}$$

$$\frac{\mathrm{d}C_2}{\mathrm{d}t} = \frac{Q}{V_1}C_1 - \frac{Q}{V_2}C_2, \tag{15}$$

$$\frac{\mathrm{d}D}{\mathrm{d}t} = k_{\mathrm{dis}}D\Big(1 - \frac{D}{D_{\max}}\Big) - k_{\mathrm{drug}}\frac{C_1}{C_1 + \mathrm{EC}_{50}}D. \tag{16}$$

Table A6: **Simulation constants and patient parameter ranges.** The ranges are pre-specified and were not tuned against any result reported here.

fixed across patients		*drawn per patient*	
step size Δt	1	clearance CL	$[0.5,\ 2.0]$
horizon	24	central volume V_1	$[5,\ 15]$
initial disease D_0	6.0	peripheral volume V_2	$[10,\ 40]$
target disease D_{target}	2.0	inter-compartment flow Q	$[0.5,\ 3.0]$
dose penalty c_{dose}	0.01	disease growth k_{dis}	$[0.05,\ 0.20]$
maximum dose $u_{\max}$	5.0	drug effect k_{drug}	$[0.3,\ 1.0]$
observation noise	0	half-effect EC_{50}	$[0.5,\ 3.0]$
		carrying capacity $D_{\max}$	$[8,\ 12]$

The marker grows logistically and is suppressed by drug effect with Michaelis–Menten saturation in C_1. The observed state is $s = (C_1, C_2, D) \in \mathbb{R}^3_{\geq 0}$ and the action is one dimensional. The dynamics are hand-written rather than learned, which is what makes the optimal control computable rather than estimated.

Trajectories are integrated with fixed-step fourth-order Runge–Kutta at $\Delta t = 1$ over a horizon of 24 steps, with all three components clipped at zero after each step. The reward is

$$r(s, u) = -\big(D - D_{\text{target}}\big)^2 - c_{\text{dose}} u^2, \tag{17}$$

evaluated at the post-transition state, and has no further terms. It is concave in u near the optimum, which matters for Section 5.5: at equal means the higher-variance action distribution has the lower expected return, so the direction of the transport-against-mixture comparison follows from Jensen's inequality and the experiment measures its size rather than its sign.

Every episode starts from $s_0 = (0, 0, D_0)$. Table A6 gives the constants and the per-patient parameter ranges. The eight parameters are drawn independently and uniformly per patient, so different patients have different optimal doses and no single action is optimal across the population.

E.2 Optimal control

The optimal control is obtained by backward dynamic programming on the known dynamics, neither analytically nor by forward search. Each state dimension is discretised into 21 points and the dose into 24 points on $[0, u_{\max}]$, a spacing of 0.217. The concentration grids run to $2u_{\max}/\text{CL} + 5$ and the disease grid to $D_{\max}$. From a terminal value of zero the recursion

$$V_h(s) = \max_u \Big[r\big(f(s, u), u\big) + V_{h+1}\big(f(s, u)\big) \Big] \tag{18}$$

is applied for $h = 23, \ldots, 0$, where f is one Runge–Kutta step and the successor value is read by nearest-grid lookup, giving a time-indexed policy table $u^\star_h(s)$.

E.3 Scoring protocol

Each patient is initialised, rolled forward ten steps under the optimal control, and the state at step six is taken as the decision state. The eleven states of that rollout form the reference on which the disagreement signal $b(s)$ is calibrated for that patient, since disagreement is defined relative to the sources built around that patient's own optimum.

Atypicality is fitted differently and deliberately so. Its reference is a population pool built from separate patients, ten states each, using the same ridge-regularised Mahalanobis construction as Equation 3 with $\varrho = 10^{-2}$. Fitting it on the patient's own trajectory would make the decision state maximally typical by construction, collapsing $a(s)$ toward zero and with it the permitted departure. The population fit mirrors the MIMIC-IV pipeline, where atypicality is fitted on the training split across patients.

Each method proposes a dose at the decision state; the dose is sampled 48 times, and every draw is executed and held for the following six steps. This is deliberate: a clinician gives one dose rather

Table A7: **The five regimes.** Each row gives the bias added to the optimal dose to form the anchor. $a(s)$ is the atypicality of the state, $z(\mathrm{CL})$ the standardised clearance of that patient, and d_A a patient-level draw. The optimiser's bias is a positive per-patient constant throughout, so the anchor and the optimiser sit on opposite sides of the optimum in A, A′ and G.

	anchor bias	optimiser bias	what it tests
A	$-2.2\,a(s)$	$\sim U(0.8, 1.5)$	the premise, made testable
A′	$-1.6\,\lvert z(\mathrm{CL})\rvert$	$\sim U(0.8, 1.5)$	the premise, without the tautology
B	0	$+1.5$	no departure needed
F	$d_A \sim \mathcal{N}(0, 1.2^2)$, constant in s	$\sim U(0.8, 1.5)$	nothing to condition on
G	$-2.2\,\big(1 - a(s)\big)$	$\sim U(0.8, 1.5)$	the premise inverted

than an expectation, and under a straight-line field the transport point at $t = 0.5$ and a mixture's expectation are the same number, so comparing expected actions would measure nothing.

What regret is measured against. The scoring baseline is the dose the optimal control proposes at the same state, executed the same way, and we report the difference in return from it. The baseline is therefore neither the optimal *sequence*, which would revise the dose at every step, nor the best constant dose over six steps. Because it freezes a policy that would otherwise adjust, a method with a continuous action can occasionally exceed it; the largest margin observed is -0.010, in regime A of Table A9, against differences between methods of 0.1 to 0.5. The coarse dose grid contributes to the same effect. This shifts every arm by the same per-patient constant and therefore leaves the paired differences, which carry every claim in Sections 5.4 to 5.7, unchanged.

Intervals and p-values measure different things. Intervals in Table A9 are normal approximations for the *mean* paired difference; p-values come from a Wilcoxon signed-rank test on the same differences. The two can disagree when the differences are heavy-tailed: the signed-rank test responds to most pairs favouring one arm, while a few extreme pairs widen the interval on the mean. The oracle selection row of regime A is an instance, with 64% of pairs favouring the arbitrated decision and $p = 3.9 \times 10^{-5}$ while the interval on the mean crosses zero.

E.4 Evidence sources and the five regimes

Three sources are built per patient, each a Gaussian policy whose mean is the optimal dose plus a regime-specific bias, clipped to $[0, u_{\max}]$, with dispersion $\sigma = 0.3$:

$$\pi_j(u \mid s) = \mathcal{N}\Big(\mathrm{clip}\big(u^\star(s) + \beta_j(s),\ 0,\ u_{\max}\big),\ \sigma^2\Big). \tag{19}$$

The *anchor* plays the role of routine practice and carries the bias that defines the regime. The *optimiser* plays the role of the learned source and carries a positive bias, constant per patient. The *rule* source plays the role of the compiled guideline: when it speaks it is close to correct, with $\beta \sim \mathcal{N}(0, 0.4^2)$, and it is silent on a fraction of states drawn per state, falling back to the anchor mean exactly as π_{guide} does in Section 4.1. That fraction is set to 0.74, chosen to approximate the 77.4% of held-out fluid decisions on which the compiled SSC rules are silent.

Table A7 gives the five anchor biases. What the five are built to separate is whether the anchor's failure is a function of the quantity the gate reads. In A it is, by construction, which makes that regime close to a tautology; a reader can see this from the constructor and should, and it is why A is not reported alone. In A′ the failure is driven by clearance, a parameter that determines the correct dose but is not a coordinate of the state, so the gate must infer it from consequences rather than read it. B removes the need to depart, F removes anything to condition on, and G inverts the premise so the gate travels least exactly where travelling most is right.

E.5 Comparators

The comparators of Table A9 are built as follows; three use ground truth and are marked as oracles there.

- **Fixed transport distance.** Interpolation from the anchor toward the mean of the evidence sources at a constant $t \in \{0, 0.25, 0.5, 0.75, 1\}$, ignoring the context. The best of the five is chosen with hindsight per regime, which is why beating it is the claim that the gate judged *when* to depart rather than that interpolation helps on average.
- **No value signal.** The same arbitration with the value-guided branch of Equation 5 removed, so the travelled distance is $a(s)\kappa_j$ exactly.
- **Selection on offline $\hat{Q}$.** A meta-policy picks one source per state by the value a k-nearest-neighbour model assigns to its mean action. That model is fitted on Monte-Carlo returns-to-go from logged patients generated by a noisy near-optimal behaviour policy with noise standard deviation 0.8, at $\gamma = 0.99$ and $k = 12$ on the standardised state-action vector. It estimates the value of the behaviour policy rather than of the optimum, and that bias is the point: it is what an analyst with logs alone would have.
- **Selection, oracle.** The same architecture with labels from the true optimal control. This is an upper bound on the selection family rather than a baseline, so losing to it is not a failure of our selector.
- **Selection on the same value signal.** Selection given the identical short-horizon value signal the gate uses, which narrows the question to whether picking or moving is better with value information held equal.
- **Best single source.** The source whose mean is closest to the optimum at the decision state, chosen with knowledge of the optimum.
- **Precision weighting.** The Gaussian product $\mu^\star = \sum_j \mu_j \sigma_j^{-2} / \sum_j \sigma_j^{-2}$, which is Bayes-optimal only when each σ_j is the error variance of an estimate of a common quantity. Here they are dispersions, which is the same incommensurability that motivates routing GRADE through a separate channel.
- **Fixed mixture.** Equal-weight mixture of the three sources, sampled by picking a component and then sampling it.
- **Constraints alone.** One optimised policy with the rule source used purely as a hard filter that forbids actions without proposing any, which is a published alternative shape for guideline use.

The value signal the gate uses here is the true short-horizon return over three steps, since the dynamics are known. It is an oracle and is labelled as such, so the claim this experiment supports is conditional: given a usable value signal, the gate regulates departure correctly. On MIMIC-IV the analogous object would be estimated and would carry extrapolation error, which is why that branch is not used there.

E.6 The constraint and the two-source configurations

Section 5.7 uses a sixth configuration in which both sources are biased in the *same* direction, the anchor by $-d_A$ with $d_A \sim U(0.8, 1.4)$ and the optimiser by a further $U(0.3, 0.8)$ below it, so the optimum lies outside their convex hull and no integration scheme can reach it by interpolation. The sources agree, so $b(s)$ is low and every uncertainty signal built on disagreement reports confidence. The bias is downward because that is what the consensus-extrapolation region looks like on MIMIC-IV: the lowest fluid and vasopressor rates of the four regions alongside the second-highest mortality.

The constraint is structurally the same object as the compiled SSC0009 rule, a guard on the state together with a forbidden region of the action space, naming no dose:

$$C(s) = \{u : u \geq \phi(s)\}, \qquad \phi(s) = \begin{cases} 0.55\, u^\star(s) & D > 3, \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

The floor is a fraction of the optimal dose rather than the optimal dose itself, so the constraint gives a direction rather than the answer; one that already knew the optimum would make the experiment vacuous. The violation magnitude is $\max(\phi(s) - u, 0)$; the hard arm projects by $u \mapsto \max(u, \phi(s))$, the one-dimensional case of the Euclidean projection $\Pi_{C(s)}$ of Equation 9; the soft arm instead subtracts the violation from the arbitration logits; the unconstrained arm applies neither.

Table A8: **Statistics of the four regions of the decision plane.** Medians are over held-out decision windows; mortality is at 90 days with a 95% stay-clustered bootstrap interval, since a patient contributes about eleven windows sharing one outcome. The number of distinct patients per region is given because it, not the number of windows, sets the precision. The treated rate is the fraction of windows in which the clinician gave a non-zero dose. Median age runs from 67 in the routine region to 63 at the true edge, and the median dose among treated windows from 131 to 341 mL and 0.070 to 0.142 mcg/kg/min. Every column also appears in Figure 2; the table is for readers who want the numbers.

				Median				Treated (%)	
Region	n	stays	%	SOFA	Lact.	MAP	90-d mortality	fluid	vaso
Q0 routine	11,188	1,998	28.0	5.0	1.30	77.75	20.7 [17.9, 23.4]	83.6	13.5
Q1 known controversy	9,035	2,275	22.6	7.0	1.63	74.75	25.8 [22.9, 28.6]	92.6	38.2
Q2 consensus extrap.	8,860	1,553	22.1	6.0	1.30	79.25	36.5 [32.3, 41.2]	63.8	11.9
Q3 true edge	10,917	2,419	27.3	8.0	2.10	74.25	47.7 [43.8, 51.2]	87.0	46.0

Section 5.5 uses a further configuration in which the two evidence sources are constant-dose policies on opposite sides of the optimum, each sampled 64 times per patient. Transport pairs each anchor draw with its own matched evidence draw and moves along the segment between them, which is why its draws are unimodal rather than merely narrower, while the mixture picks a component and then samples it. Executed returns are averaged over draws over the same six-step horizon.

F Additional experiments

F.1 Statistics of the four regions

Table A8 gives the raw numbers behind Figure 2. Every column also appears in that figure; the table is for readers who want the values rather than the picture.

F.2 Full comparison in the simulated environment

Figure 3 summarizes the comparison with the best fixed transport distance. Table A9 reports the full comparator set, per-arm regret, paired differences, and Wilcoxon signed-rank p-values across all five regimes.

F.3 Threshold sensitivity

The ordering of mortality across the four regions holds at every threshold in $[0.3, 0.7]$ (Table A10). The region shares move with the threshold, which is an arithmetic consequence of both axes being empirical CDF positions rather than evidence about the data, so the shares are reported for completeness and the ordering is what carries the claim.

F.4 Seed robustness of the context diagnosis

Across five reference seeds the consensus-extrapolation region occupies $22.1 \pm 0.3\%$ of states with a mortality of $36.2 \pm 0.5\%$. The spread is across seeds and is neither a confidence interval nor, with five seeds, an interval at all; the observed ranges are 21.7 to 22.5% and 35.4 to 36.5%.

F.5 Sensitivity to how the optimised source was trained

Replacing the propensity weighted source with a naively trained one changes the median raw disagreement by a factor of 2.4, from 0.256 to 0.106, consistent with the difference in policy drift during training. After rank calibration the naive variant places 21.7% of states in the consensus-extrapolation region at a mortality of 35.7%, both inside the ranges observed across five seeds of the weighted source. Rank calibration makes the diagnosis depend on the ordering of disagreement rather than on its magnitude.

Table A9: **Full comparison across the five regimes.** Regret is measured against exact optimal control and differences are paired within patient, so a positive difference means PRUDENT has lower regret. Only A and A′ satisfy the premise the context factor encodes; B removes the need to depart at all, and F and G break or invert that premise. Intervals are 95%; p values are from a Wilcoxon signed-rank test. † marks comparators that use ground truth and are therefore oracles. Figure 3 plots the first row of each block. Applying the constraints alone is omitted for space; it loses by 0.78 to 0.94 in every regime.

Comparator	regret	paired difference	p
Regime A, the anchor degrades as states become rarer			
PRUDENT	-0.010	—	—
best fixed distance ($t = 0.25$)	+0.208	+0.217 [+0.148, +0.287]	4.2×10^{-22}
no value signal (ablation)	+0.261	+0.271 [+0.216, +0.325]	1.3×10^{-24}
select on offline $\hat{Q}$	+0.529	+0.539 [+0.369, +0.709]	1.2×10^{-11}
select, oracle†	+0.050	+0.060 [-0.072, +0.192]	3.9×10^{-5}
select on same value signal†	-0.281	-0.272 [-0.390, -0.153]	8.3×10^{-6}
best single source†	-0.006	+0.004 [-0.137, +0.145]	0.03
precision weighting	+0.321	+0.330 [+0.264, +0.397]	9.5×10^{-19}
fixed mixture	+0.476	+0.485 [+0.421, +0.550]	7.1×10^{-28}
Regime A′, the anchor degrades with an unobserved pharmacokinetic parameter			
PRUDENT	+0.163	—	—
best fixed distance ($t = 0.75$)	+0.287	+0.124 [-0.033, +0.281]	6.3×10^{-8}
no value signal (ablation)	+0.416	+0.253 [+0.190, +0.317]	7.7×10^{-19}
select on offline $\hat{Q}$	+0.596	+0.433 [+0.189, +0.677]	1.7×10^{-10}
select, oracle†	+0.100	-0.063 [-0.261, +0.136]	0.07
select on same value signal†	-0.297	-0.459 [-0.707, -0.211]	0.06
best single source†	+0.381	+0.218 [+0.092, +0.344]	7.5×10^{-6}
precision weighting	+0.302	+0.139 [+0.011, +0.268]	5.4×10^{-8}
fixed mixture	+0.545	+0.382 [+0.275, +0.489]	1.1×10^{-17}
Regime B, the anchor is correct everywhere			
PRUDENT	+0.071	—	—
best fixed distance ($t = 0.00$)	+0.120	+0.049 [+0.019, +0.080]	3.9×10^{-4}
no value signal (ablation)	+0.471	+0.401 [+0.313, +0.489]	1.9×10^{-21}
select on offline $\hat{Q}$	+0.608	+0.538 [+0.402, +0.673]	8.7×10^{-13}
select, oracle†	+0.146	+0.075 [+0.033, +0.118]	2.1×10^{-4}
select on same value signal†	+0.042	-0.029 [-0.085, +0.028]	0.83
best single source†	+0.129	+0.059 [+0.024, +0.093]	1.3×10^{-3}
precision weighting	+0.500	+0.429 [+0.351, +0.507]	6.8×10^{-25}
fixed mixture	+0.550	+0.479 [+0.409, +0.549]	4.1×10^{-29}
Regime F, the anchor fails at random, unrelated to rarity			
PRUDENT	+0.375	—	—
best fixed distance ($t = 0.50$)	+0.521	+0.146 [+0.079, +0.213]	2.0×10^{-10}
no value signal (ablation)	+0.583	+0.209 [+0.148, +0.270]	4.2×10^{-12}
select on offline $\hat{Q}$	+0.697	+0.322 [+0.165, +0.480]	9.9×10^{-7}
select, oracle†	+0.364	-0.011 [-0.134, +0.111]	0.02
select on same value signal†	+0.113	-0.262 [-0.385, -0.138]	6.7×10^{-4}
best single source†	+0.434	+0.060 [-0.046, +0.166]	3.0×10^{-6}
precision weighting	+0.561	+0.186 [+0.113, +0.260]	3.1×10^{-9}
fixed mixture	+0.641	+0.267 [+0.196, +0.337]	1.5×10^{-14}
Regime G, the anchor is worse on common states, inverting the premise			
PRUDENT	+0.032	—	—
best fixed distance ($t = 0.25$)	+0.137	+0.105 [+0.065, +0.145]	1.2×10^{-13}
no value signal (ablation)	+0.359	+0.327 [+0.261, +0.393]	1.1×10^{-22}
select on offline $\hat{Q}$	+0.416	+0.384 [+0.184, +0.585]	1.4×10^{-10}
select, oracle†	+0.190	+0.158 [+0.002, +0.315]	1.0×10^{-3}
select on same value signal†	-0.295	-0.326 [-0.486, -0.167]	3.6×10^{-3}
best single source†	+0.125	+0.093 [+0.043, +0.144]	1.2×10^{-6}
precision weighting	+0.275	+0.243 [+0.130, +0.356]	1.2×10^{-10}
fixed mixture	+0.437	+0.406 [+0.314, +0.498]	1.5×10^{-18}

Region membership is less stable than region size or outcome. The fraction of consensus-extrapolation region windows in which a vasopressor was actually administered is $12.1 \pm 0.9\%$ across seeds of the weighted source and 7.9% under the naive one, which is 2.9 points below the lowest seed. The regions are thus reproducible in size and in mortality while differing somewhat in membership, and the difference appears on the vasopressor dimension, where the guideline is silent and behaviour is determined by the optimised source alone.

Table A10: **Sensitivity of the four regions to the partition threshold.** Each row applies the same threshold θ to both axes. The region shares move with θ by arithmetic necessity, since both axes are empirical CDF positions, so the shares carry no evidence. What the table shows is that the ordering of 90-day mortality across the four regions is unchanged throughout, and in particular that the consensus extrapolation region carries higher mortality than the known controversy region at every threshold tested. The row at $\theta = 0.50$ recomputes the partition from the stored signals rather than reading the gate probe output, so two of its entries differ from Table A8 by 0.1 point through rounding at the region boundary.

		90-day mortality (%), share in brackets			
θ	n	routine	known contr.	consensus extr.	true edge
0.30	40,000	14.9 [10%]	19.7 [21%]	33.1 [19%]	41.7 [50%]
0.40	40,000	17.8 [18%]	23.2 [23%]	34.7 [22%]	44.7 [37%]
0.50	40,000	20.7 [28%]	25.9 [22%]	36.5 [22%]	47.8 [27%]
0.60	40,000	23.1 [40%]	29.3 [20%]	38.4 [21%]	50.8 [19%]
0.70	40,000	25.5 [53%]	32.2 [17%]	39.8 [18%]	55.7 [12%]
Ordering preserved at every threshold: yes					

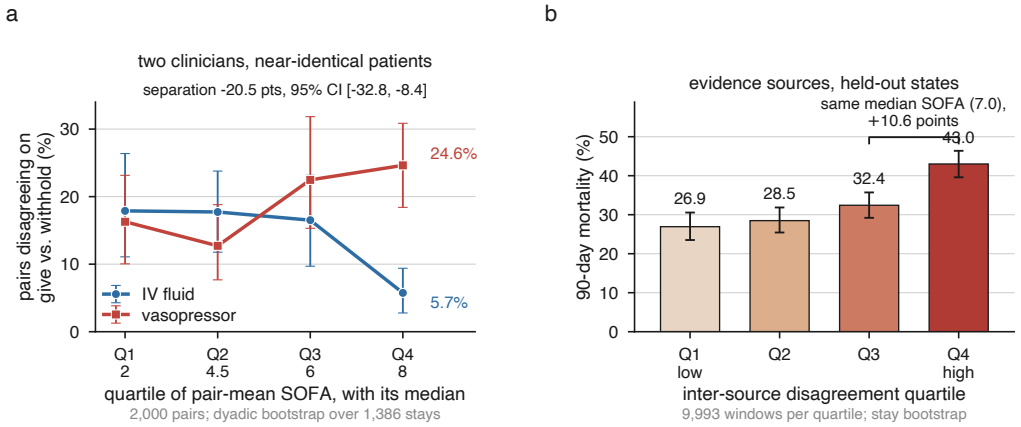


Figure A1: **Disagreement is structured, and informative beyond severity. (a)** Among near-matched patients, disagreement about fluid decreases with severity while disagreement about vasopressor increases; intervals use a stay-level dyadic bootstrap. **(b)** Ninety-day mortality increases across quartiles of inter-source disagreement, while the upper two quartiles share a median SOFA of 7.0; mortality intervals are stay-clustered.

F.6 Are the two disagreement dimensions related?

Within a matched pair, disagreement about fluid and disagreement about vasopressor are close to independent. Both dimensions were contested in 47 of 2,000 pairs against 53.3 expected under independence, giving a ϕ coefficient of -0.023 with a dyadic interval of $[-0.100, +0.065]$. The statistic is computed on the stored pair indices rather than on a rebuilt matching, and its interval uses the same node bootstrap as the rates themselves.

This is a stronger statement than opposite severity gradients. Gradients of opposite sign are still compatible with a single latent difficulty driving both dimensions with loadings of opposite sign, and a null ϕ is not. It is the empirical counterpart of representing silence per action dimension rather than per source: if how contested a decision is depends on which action is at issue, then so must the record of which sources spoke.

F.7 Does the disagreement signal depend on the guideline, and where do the rules fire?

A compiled rule that fires places its target near the action ceiling and far from routine practice, so firing almost always produces disagreement. Our cohort is built from circulatory failure rather than from an infection criterion, so it contains patients for whom the fluid recommendation does not apply and may be contraindicated. On those states the recorded disagreement could reflect a rule applied outside its coverage rather than a genuine dispute between evidence sources. This is a question about what $b(s)$ measures, and the descriptive stance of Section 5.1 does not answer it, so we bound where the concern applies.

The signal does not depend on the guideline. A rule can distort only the states on which it fires. Of the $40{,}000$ held-out windows, $30{,}939$ (77.4%) are never fired on, including 54.6% of the high-disagreement half. Restricted to those states, the difference in mortality between the highest and lowest disagreement quartile is $+7.9$ percentage points (stay-clustered 95% CI $[+3.3, +12.2]$), against $+16.1$ ($[+11.8, +20.0]$) on all states. Quartiles are recut within each subset, since reusing the full-sample cut points would place never-fired states in the low quartiles by construction. The gradient therefore survives without the guideline, at roughly half the magnitude.

Part of that reduction reflects selection rather than contamination. The rules fire on sicker patients by construction, since the compiled predicate requires lactate above 2: fired states have median SOFA 7.0 and 42.8% mortality against 6.0 and 29.8% where no rule fires. Firing is thus correlated with severity and severity with outcome, without any misuse being required to explain it.

The main finding lies outside the objection's reach. No compiled rule fires on any of the $11{,}188$ routine states or on any of the $8{,}860$ consensus-extrapolation states. The composition of the consensus-extrapolation region, its treatment rates, its mortality, and the 1.76-fold ratio between it and the routine region are therefore unaffected by whether a rule was ever applied outside its coverage. This is structural rather than statistical: a mechanism that operates only through firing has nothing to act on where the firing count is zero.

Inside the two regions where rules do fire. The known-controversy region fires on 38.4% of its states, and the two subsets reach the same outcome: 24.9% mortality ($[21.0, 28.9]$) where a rule fires against 26.3% ($[23.0, 29.6]$) where none does, a difference of -1.4 points with interval $[-5.9, +2.8]$. The fired subset also has the *lower* median SOFA (6.0 against 7.0) and the higher median lactate (2.50 against 1.30): raised lactate with less organ failure, which is the population the recommendation is written for. The name given to that region applies to the whole of it.

The true-edge region fires on 51.2% of its states and behaves differently. The two subsets share a median SOFA of 8.0, but the fired subset has higher lactate (2.80 against 1.40), receives fluid less often (79.2% against 95.3%) and vasopressors less often (41.2% against 51.0%), and dies 13.0 percentage points more frequently ($[+8.3, +17.7]$). Two readings are consistent with that pattern and the available data do not separate them: the rules may be firing on patients for whom fluid is contraindicated, or they may be correctly marking a group of severely ill patients treated less intensively overall who fare worse. The first would ordinarily show withheld fluid alongside *greater* vasopressor use as a substitute, whereas both fall together here, which favours the second; but without an infection adjudication or a record of goals of care we report the ambiguity rather than resolving it with a proxy we could not defend.

F.8 Transport against mixture in the symmetric setting

Figure A2 reports the within patient dispersion, which is the quantity the argument concerns. Pooling draws across patients instead would give standard deviations of 1.63 and 1.29, but those numbers are dominated by variation in where each patient's optimum lies and understate the contrast by a factor of four.

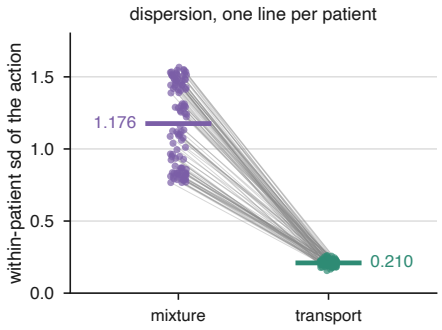


Figure A2: **Dispersion of the sampled action, within patient.** One line per patient, over the 100 patients of the symmetric setting. Because the two sources sit on opposite sides of the optimum, the mixture is bimodal and its draws scatter across both modes, while transport keeps them in one place that a source supports. Transport is narrower on every patient. The horizontal bars are the means, 1.176 and 0.210, which are the values reported in Section 5.5.

F.9 Constraint projection

This experiment runs in the PK/PD simulation and not on MIMIC-IV, so the binding fraction of 45.5% is a property of that environment. Over 400 patients, hard projection gives a regret of 0.112 with no violations, a soft penalty gives 0.145 with a violation rate of 35.6%, and no constraint gives 0.148 with 35.9%.

The per region breakdown contains between 21 and 218 patients, so we do not read an ordering from it. The stable observation is that the region in which the unconstrained arm performs worst is consensus extrapolation, with a regret of 0.242 against 0.036 to 0.064 elsewhere, which is also where the context-dependent violation weight $c(s)$ is highest.

The constraint-only arm of Table A9, in which the rule source forbids actions without ever proposing a dose, loses to the arbitrated decision by 0.78 to 0.94 in every regime. A filter that names no dose cannot move the decision toward the optimum; it can only prevent one class of error, which is the division of labour the projection of Equation 9 rests on.

The same comparison bounds what the violation term of Equation 10 can be asked to do. Subtracting it from the arbitration logits moves the violation rate from 35.9% to 35.6%, which is within noise, while projection removes violations entirely. The soft penalty is therefore a preference among feasible routes and an ingredient of attribution, not a safety mechanism; treating it as one would misplace the guarantee that Equation 9 actually provides.

F.10 Displacement from routine practice

Mean absolute displacement from routine practice grows across the regions, from 21 mL in the routine region to 185 mL at the true edge on the fluid dimension, and from 0.0019 to 0.0211 mcg/kg/min on the vasopressor dimension. Part of this growth is structural, since the permitted departure $a(s)\kappa_j$ increases with atypicality. The comparisons that are not definitional are those within an atypicality level, where displacement still grows with disagreement, from 21 to 56 mL among typical states and from 63 to 185 mL among atypical ones.

F.11 Attribution under intervention

The two interventions of Section 5.6 are shown in Figure A3. The correlation between a source's weight and the movement caused by removing it is $+0.79$ for the guideline and -0.55 for the optimised source, so the two are not summarised by a single statement that weight predicts movement. The negative correlation is the same asymmetry seen in the prediction error: states where the opti-

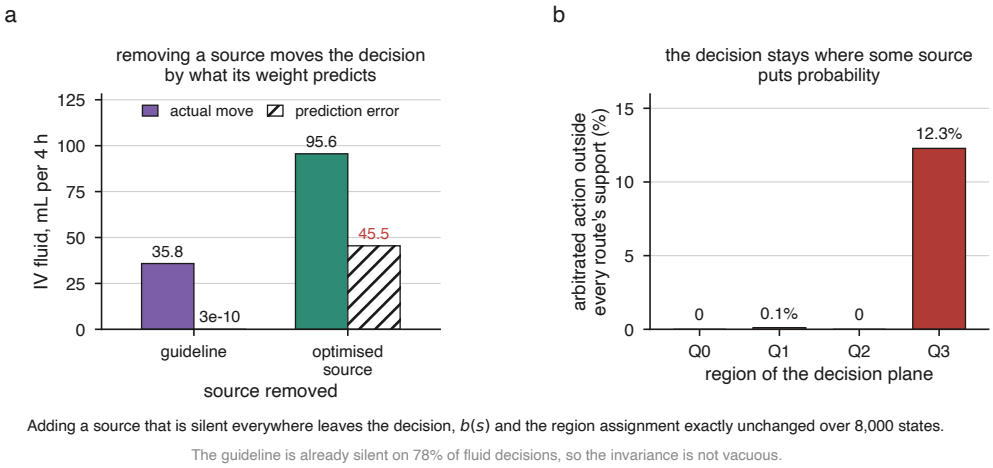


Figure A3: **Two interventions on the representation. (a)** Source removal exposes both influence and coverage. Removing the guideline moves the fluid decision by 35.8 mL on average, which its recorded weight predicts to within 3×10^{-10}. Removing the optimised source moves the decision by 95.6 mL, with a weight-only prediction error of 45.5 mL because the remaining guideline is silent on most states and the decision therefore falls back to the anchor. **(b)** The arbitrated action lies more than two nats below every route's density (a factor of about 7.4) on 3.3% of states, concentrated in the region where both sources speak often and disagree.

mised source carries most weight are largely states where the guideline is silent, and removing the optimised source there leaves nowhere for the decision to go.

F.12 Worked cases

Selection of the traced cases is not neutral. Ordering by displacement favours states where the guideline fired in those regions where it fires at all, and in the two low firing regions it cannot favour them, because there are almost no fired states to select. Every case is recorded with its row index in the test split, its firing state, and its within region displacement rank, so an alternative can be substituted and the figure regenerated.

Two features of Figure A4 are worth stating so the arithmetic can be followed. The millilitre values printed on each panel are rounded from the continuous quantities the decision is computed in, so the anchor and the weighted travel distances do not sum exactly to the printed arbitrated dose; the residual is rounding alone, and the exact reconstruction reported in Section 5.6 is to 2×10^{-16}. In panel (b) the silent guideline still shows a small non-zero travel: its route is constructed from a fallback to π_{behav}, but the anchor and the fallback destination are independent draws from that distribution, so the route has a short length even though its two ends share a distribution. Its arbitration weight is exactly zero, so that length does not enter the decision.

G Reproducibility

Code, the frozen guideline source with its hash, the compiled rule set, the annotation workbook, the context reference indices, and the result files behind every figure will be released upon publication.

The starting point of each traced decision is a sample from routine practice rather than its mean. The configured seed reaches each evidence source's own generator, so repeated runs place the same state at the same anchor; an earlier version seeded only the arbitrator while the policies held unseeded generators, which left the millilitre values in Figure A4 irreproducible across runs. Separately, analyses draw their own subsample of the test split, so counts should be compared within an analysis and not across them.

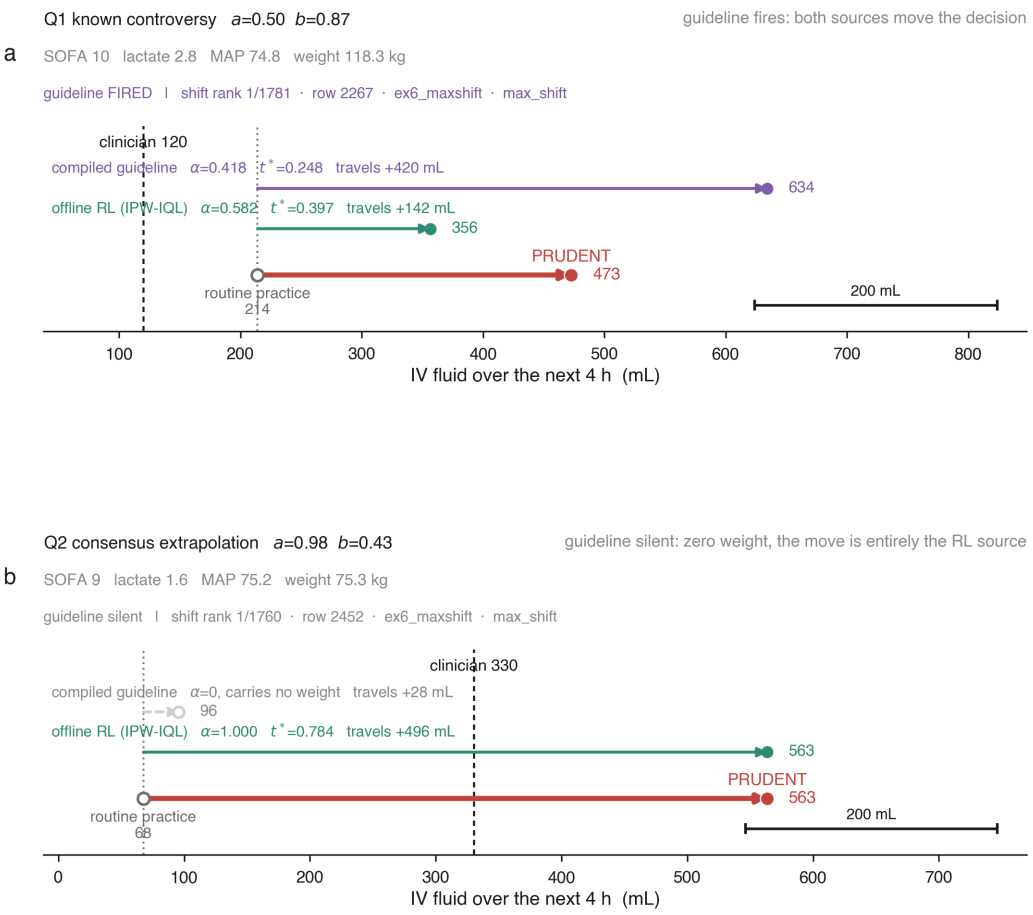


Figure A4: **Two decisions traced end to end.** Each panel is a fluid-dose number line in mL per 4 h. The decision starts at routine practice (π_{behav}, open circle); each evidence source proposes a destination reached by transporting a distance t^* from that anchor; PRUDENT is the α-weighted combination. The clinician's actual dose is shown for reference, not as a target. Note the scale bars: each panel is scaled to its own route geometry, so arrow lengths are not comparable across panels. **(a)** A contested state in the known controversy region. The guideline fires, its route travels $+420$ mL and the offline RL route $+142$ mL, and the decision moves from 214 to 473 mL. **(b)** A rare state in the consensus-extrapolation region. The guideline is silent, so it carries weight exactly zero and its route travels $+28$ mL while the RL route travels $+496$ mL: the entire departure from routine practice, 68 to 563 mL, comes from one source. Both cases are the largest-displacement state in their region; the row index and rank are printed in the figure so each can be checked against the stored result files.