TwinPath AI: Longitudinal Prediction Care Alignment for Clinical AI Governance

Team Members:
Student: Satyajit Movidi (UWF)
Student: Maria Ramirez (UWF)
Faculty: Dr. Stephen Russell (UWF)

Community Advisors:
Heather Anderson (MS DS, UWF Alumni)
Venessa Goodnow (VP, Jackson Health System)
Fabio Montes Suros (AI Engineer, Jackson Health System)

Abstract

Background: Modern healthcare AI creates a governance problem that model accuracy alone cannot answer. A predictive model may identify elevated risk correctly while subsequent care is delayed, not administered, differently documented, or otherwise no longer congruent with the condition that produced the prediction. Current healthcare policy and accreditation increasingly emphasize lifecycle governance, local oversight, transparency, and ongoing monitoring of deployed AI. TwinPath AI addresses a specific operational gap within that larger problem: whether documented care remains aligned with model identified risk over time.
Methods: TwinPath AI tracks two longitudinal paths for each hospitalization: daily VTE risk and documented prevention behavior. A LightGBM model produces the risk path. A deterministic Governance Engine combines risk category with medication order, administration, delay, and therapeutic anticoagulation evidence to assign one of six auditable patient day states. A local large language model (LLM) Governance Agent does not assign those states; it makes the governance layer usable by translating natural language questions into SQL, retrieving the supporting evidence, and interpreting alignment conditions across individual encounters and populations. The framework was developed on 541,469 MIMIC-IV admissions and 2,701,574 patient day observations. Four local LLMs were evaluated on 50 deterministic retrieval questions, 10 single reference interpretation questions, and 10 acceptable range governance questions.
Results: The predictive model achieved calibrated AUROC 0.8299, AUPRC 0.2332, and Brier score 0.0341, providing a useful longitudinal risk signal for governance. Qwen3.6:27b ranked first among the four tested agent models on all three reported benchmarks, with deterministic SQL accuracy of 1.00, Complex Single interpretation of 0.69, Complex Distribution acceptability of 0.86, and the lowest reported Complex Single semantic frameshift at 0.108. Across deterministic retrieval, fact-grounded interpretation, contradiction analysis, repeated-sample stability, and reference-relative semantic measures, the evaluation provides quantitative evidence that the agent can perform the core information tasks required for longitudinal governance in the tested retrospective environment.
Conclusions: TwinPath AI shifts the center of evaluation from prediction alone to prediction-care alignment. Its contribution is an auditable governance architecture in which deterministic rules define the high consequence state and an LLM provides evidence access, synthesis, and explanation. The multi-axis agent evaluation supports a governance capability claim within the defined evidence basis: the agent can retrieve governed evidence, interpret alignment conditions, and remain measurably faithful to bounded governance references. Prospective clinical benefit, clinician adjudication, lead time, and patient outcome effects remain future validation targets.

1 Problem Background and Significance

Healthcare is moving from isolated AI model evaluation toward organizational governance of AI in use. The Office of the National Coordinator for Health Information Technology established algorithm transparency requirements for predictive decision support interventions in certified health IT, which supports care in more than 96 percent of U.S. hospitals [1]. The U.S. Food and Drug Administration has framed AI enabled medical device oversight around total product lifecycle risk management rather than a single development time assessment [2]. In June 2026, Joint Commission launched its Responsible Use of AI in Healthcare certification for healthcare organizations, with standards covering governance, data management, risk and bias reduction, monitoring and validation of safety and effectiveness, transparency, and education [3]. These developments make the provider side question increasingly concrete: a health system must be able to observe how an AI system behaves after deployment and how that behavior interacts with care delivery.

Prediction-care alignment is one part of that governance obligation that conventional model metrics do not expose. AUROC can describe discrimination and calibration can describe agreement between predicted and observed risk, but neither shows whether a high risk patient subsequently received the documented prevention pathway expected by the organization. A model can perform as designed while the operational pathway around it becomes delayed, incomplete, inconsistently documented, or difficult to audit. In that setting, the governance problem is not simply whether the model is correct. It is whether risk recognition and care execution remain coherent over time. VTE prevention provides a useful application because the relevant signals are distributed across the EHR and unfold longitudinally. Prior research has demonstrated predictive signal for VTE in structured clinical data [4, 5], and AI enabled VTE surveillance has been framed as a human and AI teaming problem [6]. Prevention, however, is not represented by a single field. Risk, prophylaxis orders, medication administrations, delays, missed administrations, and therapeutic anticoagulation appear as separate

events. TwinPath AI converts those fragments into an observable governance trajectory and makes the relationship between predicted risk and documented care available for patient level and population level review.

The central contribution is therefore not another VTE risk model. The predictive model is necessary because it supplies the first path, but the research question begins after prediction: when a model identifies elevated risk, can a healthcare organization determine whether the documented care pathway remains aligned, identify where it diverges, and interrogate that divergence in an auditable way? TwinPath AI addresses that question by separating deterministic governance state assignment from LLM based evidence retrieval and interpretation.

2 Data Source

TwinPath AI was developed using the publicly available MIMIC-IV version 3.1 database [7, 8]. Multiple MIMIC-IV tables were integrated into a longitudinal SQL representation of each hospitalization, including patient demographics, hospital admissions, diagnosis codes, laboratory measurements, medication orders and administration records, ICU stays, hospital service assignments, and transfer events. These sources support two different functions within the framework. Demographic, admission, comorbidity, and longitudinal laboratory data provide the information used to estimate daily VTE risk, while medication orders, medication administrations, and therapeutic anticoagulation evidence provide the structured care evidence used by the Governance Engine and Governance Agent.

VTE labeled admissions were identified from ICD-9-CM and ICD-10-CM diagnosis codes. Admissions with VTE in the primary diagnosis position were excluded, leaving 8,795 admissions with VTE recorded in a nonprimary diagnosis position among 541,469 total admissions, or 1.62 percent of the cohort. The resulting admissions were transformed into 2,701,574 patient day observations with a maximum observation window of 30 hospital days. Because MIMIC-IV diagnosis ordering does not provide diagnosis onset time, the nonprimary VTE label is used as an encounter level coding proxy and should not be interpreted as proof that a new VTE occurred after hospital admission. This distinction bounds the predictive model and prevents the present study from making incident event or lead time claims that the source data cannot support.

Daily prediction features were generated using only information available through each hospital day. The final predictive representation contained 83 engineered longitudinal features. Medication administration evidence was intentionally excluded from the prediction feature set and retained for governance analysis so that evidence used to evaluate the care pathway was not also used to construct the risk signal being governed. All patient day observations belonging to the same hospital admission were assigned to the same training, validation, or testing partition, preventing future days from the same admission from crossing those partitions. The reported partitioning is admission based rather than explicitly patient based, so repeated admissions from the same patient may cross partitions. Table 1 summarizes the data used by the predictive and governance components and clarifies the scale at which the framework was evaluated.

Table 1: The MIMIC-IV cohort provides longitudinal prediction features and structured care evidence for TwinPath AI.

Characteristic	Description
Data source	MIMIC-IV version 3.1, a publicly available deidentified EHR database.
Hospital admissions	541,469 admissions represented in the analytic cohort.
VTE labeled admissions	8,795 admissions with VTE in a nonprimary diagnosis position, corresponding to 1.62 percent of admissions.
Patient day observations	2,701,574 longitudinal observations generated from the admission cohort.
Prediction features	83 engineered features derived from patient, admission, comorbidity, laboratory utilization, and longitudinal laboratory data.
Governance inputs	Daily risk predictions, prophylaxis medication orders, medication administrations, delay evidence, and therapeutic anticoagulation evidence.
Maximum observation window	30 hospital days.

3 AI Methodology and Tools Used

TwinPath AI creates governance observability by pairing two paths that are usually evaluated separately. The first is the model generated risk trajectory. The second is the documented prevention trajectory. Their relationship is converted into a deterministic governance state that can be queried and explained without allowing the LLM to define the state itself.

3.1 TwinPath AI Framework

A pooled longitudinal LightGBM model produces a daily VTE risk score from clinical information accumulated through the current hospital day [9]. The implementation groups scores by percentile. The top 10 percent risk category, including the top 5 percent and top 1 percent subcategories, is treated as high risk for governance. The deterministic Governance Engine then evaluates the risk category together with structured order, administration, delay, and therapeutic anticoagulation

evidence. For patient i on day t, the governance state can be represented as

$$G_{it} = f(R_{it}, O_{it}, A_{it}, D_{it}, T_{it}), \tag{1}$$

where R is the model risk category, O is qualifying prophylaxis order evidence, A is administration evidence, D is delay evidence, and T is therapeutic anticoagulation evidence. The function f is deterministic and auditable. Table 2 defines the six states used in the evaluated database. The labels describe the relationship visible in structured data and do not independently determine clinical appropriateness.

Table 2: The six deterministic patient day states operationalize prediction care alignment from structured evidence.

Governance State	Operational Definition
Not High Risk	The model score was below the configured high risk threshold for that patient day.
Potential Gap / Needs Review	The patient day was high risk and the structured record did not establish qualifying prophylaxis or therapeutic anticoagulation. The state identifies an unresolved condition for review, not a confirmed care error.
Ordered but Not Administered	Qualifying prophylaxis was ordered, but no corresponding administration was documented for that patient day.
Aligned	Qualifying prophylaxis administration was documented for a high risk patient day under the deterministic rules.
Delayed	Qualifying prophylaxis was documented, but administration timing met the rule based definition of delay.
Therapeutic Anticoagulation	Therapeutic anticoagulation was documented, so additional prophylactic anticoagulation was not expected under the deterministic rules.

The Governance Agent operates over this auditable substrate. It translates a governance question into SQL, executes that query against the governance database, and produces a narrative grounded in the returned records. This division of labor is deliberate. The deterministic layer answers, "What governance state is present under the defined rules?" The agent answers, "What evidence supports that state, how does it evolve across time, and what pattern appears across patients or populations?" The resulting architecture allows a hospital quality or AI governance function to move from isolated risk scores to reviewable questions about persistence, recurrence, transitions, and concentrations of misalignment without requiring manual database analysis for each investigation.

3.2 Computational Framework

The prototype uses Python, LangGraph, LangChain, SQLite, Ollama, and locally hosted LLMs. Database schema information is supplied to the agent so generated SQL can target the governance tables and source evidence. The prototype presentation layer exposes the risk path and documented care path together with governance states, supporting records, and aggregate patterns while remaining separate from the governance logic. This makes the alignment layer visible as a patient trajectory and as a population level review surface rather than leaving it as a hidden database classification. Dashboard usability and review efficiency are not evaluated in the current study. Local inference supports institutional data locality by design, although production privacy and security would also require access control, logging, network safeguards, and institution specific deployment controls.

4 Results and Evaluation Metrics

The predictive model and the Governance Agent are evaluated separately because they serve different roles. The predictive model establishes the risk path on which alignment is conditioned. The agent evaluation tests the functions required to make the governance layer operational: retrieving the correct evidence, interpreting it faithfully, and remaining within acceptable semantic bounds when multiple interpretations are valid.

4.1 Predictive ML Model Evaluation

Model development evaluated Logistic Regression, Random Forest, XGBoost, and LightGBM together with multiple feature and temporal representations. The final pooled LightGBM model used 83 engineered features covering patient characteristics, admission context, chronic comorbidities, laboratory utilization, and longitudinal laboratory summaries. Medication administration variables were excluded from prediction and reserved for governance, preserving a separation between the signal used to identify risk and the evidence used to evaluate subsequent prevention behavior.

Table 3 reports the final risk model results. Calibration changed discrimination minimally while reducing the Brier score from 0.1481 to 0.0341. The calibrated AUROC of 0.8299 and AUPRC of 0.2332 were obtained in a test population with 1.61 percent VTE coded admissions, providing a useful risk ranking signal for the governance layer rather than a stand alone treatment rule.

Table 4 places the TwinPath AI model beside four recent MIMIC-IV VTE prediction studies. The TwinPath AI test set is substantially larger and has lower prevalence than the comparison cohorts. Direct ranking by accuracy or

Table 3: The final longitudinal risk model preserved discrimination after calibration while reducing Brier score.

Metric	Raw Model	Calibrated Model
AUROC	0.8302	0.8299
AUPRC	0.2397	0.2332
Brier Score	0.1481	0.0341

precision would be inappropriate because prevalence, population, outcome construction, feature timing, and threshold selection differ across studies. The comparison nevertheless shows that the TwinPath risk model provides meaningful discrimination in a broader and more imbalanced cohort, which is sufficient for its role as the first path in the governance framework. A supplemental report entitled *Reference ML Model Literature* examines these comparability constraints and the methodological limitations of the four published benchmarks in greater detail.

Table 4: TwinPath AI is shown with representative MIMIC-IV VTE models for context, not direct ranking.

Study	Test N	VTE N	VTE %	AUROC	Accuracy	Precision	Recall
Zhang et al. [10]	7,560	268	3.54	0.956	0.940	0.360	0.888
He et al. [11]	448	68	15.18	0.778	0.864	0.798	0.647
Qi et al. [12]	312	20	6.41	0.723	0.907	0.154	0.100
Yang et al. [13]	568	75	13.20	0.890	0.824	0.413	0.787
TwinPath AI	**109,206**	**1,759**	**1.61**	**0.830**	**0.880**	**0.084**	**0.651**

The predictive model is therefore a necessary enabling component but not the endpoint of the study. Its purpose is to establish a longitudinal risk path against which prevention behavior can be governed. The research contribution begins where most prediction studies stop: after risk has been estimated and the organization must determine whether the documented care pathway remains congruent with it.

4.2 TwinPath AI Governance Agent

The Governance Agent is a multi stage pipeline consisting of natural language question interpretation, SQL generation, database execution, and narrative synthesis. A governance answer can fail because the wrong records were retrieved, because the correct records were interpreted incorrectly, or because repeated answers drift across materially different meanings. The evaluation therefore separates retrieval from interpretation rather than reducing performance to a single LLM score. Table 5 summarizes the three test sets. The deterministic SQL set tests whether the agent can reach the correct governance evidence. The Complex Single set tests whether retrieved facts are interpreted against one bounded reference answer. The Complex Distribution set tests whether generated responses remain within a predefined range of acceptable interpretations when no single narrative is canonical. The latter two reference sets were author developed from computed facts and governance definitions and are not clinician adjudicated clinical ground truth.

Table 5: The agent tests isolate retrieval, bounded interpretation, and acceptable distributional reasoning.

Test	Reference Basis	Questions	Governance Capability
Deterministic SQL	Canonical SQL result set	50	Retrieve the correct structured evidence for a governance question.
Complex Single	One reference answer plus computed facts	10	Interpret the retrieved evidence accurately and consistently.
Complex Distribution	Set of acceptable reference answers	10	Keep generated interpretations within the acceptable semantic range.

4.2.1 Evaluation Metrics Selection

The validation harness evaluates the agent against the form of ground truth available for each governance task rather than relying on a single judge or a single semantic score. The deterministic SQL tier pairs each natural language governance question with canonical SQL, executes both the gold query and the agent generated query against the same database, and compares the returned result sets. Matching is insensitive to row order, column order, and column names, and numeric strings are normalized before comparison. Accuracy therefore measures retrieval of the correct governance evidence rather than similarity between SQL strings. Executable rate, retry count, and latency are retained separately so syntactic failure, incorrect retrieval, and recovery burden remain distinguishable.

The Complex Single tier tests interpretation after retrieval. Each question contains a canonical reference answer and independently verifiable facts whose expected values are recomputed from gold SQL at evaluation time. Numeric facts

are matched as complete numeric tokens, with optional tolerance for rounding, and categorical facts are matched case insensitively. This gives the narrative response an objective factual backbone that embedding similarity cannot substitute for. A fixed rubric judge then scores interpretation, appropriate governance caveats, and faithfulness while decomposing the answer into supported, unsupported, and contradicted claims. Numeric correctness is deliberately excluded from that judge because the fact matcher already measures it directly. Each question is also resampled five times at temperature 0.7. Responses are embedded and grouped into semantic clusters using cosine similarity with a default threshold of 0.85. Stability is summarized by semantic entropy,

$$H = -\sum_{c} p_c \log_2 p_c, \tag{2}$$

where p_c is the proportion of repeated responses in semantic cluster c. Low entropy indicates that repeated runs retain the same underlying interpretation; higher entropy indicates movement among different interpretations. Pairwise embedding dispersion is retained as a continuous companion measure.

The Complex Distribution tier addresses governance questions for which several responses can be acceptable. The agent is resampled eight times, and the candidate responses are evaluated against the complete set of acceptable references rather than a single canonical answer. Reference and agent responses are embedded in a shared semantic space and clustered over the same support. Jensen-Shannon divergence compares the resulting reference and agent distributions,

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2} D_{\mathrm{KL}}(P \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M), \qquad M = \frac{1}{2}(P + Q), \tag{3}$$

with lower values indicating closer distributional agreement. Semantic coverage records the fraction of reference modes reached by the agent. A separate membership judge scores each sampled answer against the full acceptable set, and contradictions with the reference consensus are counted independently. When a distribution question includes computed facts, claim-level faithfulness is evaluated against those facts as an additional check.

Stored responses are also scored in a common reference-relative semantic geometry. The evaluator records nearest-reference similarity, translational distance, self-neighbor similarity, neighborhood cohesion, neighborhood entropy, and neighborhood rewiring. Frameshift in Figure 3 combines translational distance with neighborhood rewiring, so low frameshift requires a response to remain close to an acceptable reference while also preserving its surrounding semantic neighborhood. The code additionally computes lexical reframing, bidirectional Kullback-Leibler divergence, maximum mean discrepancy, and energy distance as diagnostic measures. These measures are complementary rather than interchangeable: retrieval accuracy tests access to the correct evidence, fact coverage tests numerical grounding, rubric and contradiction measures test interpretive faithfulness, repeated-sample entropy tests stability, and reference-relative geometry tests whether meaning moves away from the acceptable governance region.

4.2.2 Results

Figure 1 shows that Qwen3.6:27b ranked first on all three reported test sets. It achieved 1.00 deterministic SQL accuracy, 0.69 Complex Single interpretation, and 0.86 Complex Distribution acceptability. The 1.00 result establishes that the tested Qwen configuration retrieved the correct result set for all 50 deterministic governance questions. The interpretive tiers then test a different requirement: whether the agent can turn the retrieved evidence into a faithful governance explanation when one answer is bounded by computed facts and when several answers are legitimately acceptable. The separation matters because governance requires both evidentiary correctness and interpretive control.

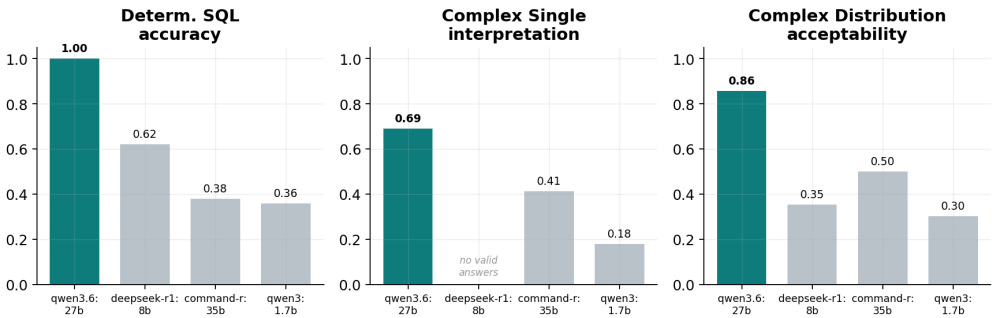

Figure 1: Agent model performance across the three reported governance test sets.

Figure 2 reports a separate failure mode. Qwen3.6:27b had the lowest observed contradiction burden among models with valid responses, while command-r:35b and qwen3:1.7b contradicted references more often. For a governance agent, contradiction matters because a fluent answer that conflicts with the evidence can misdirect review even when the underlying database is correct. The result supports model selection within the tested benchmark, but it does not establish clinical safety because the interpretive references were not clinician adjudicated.

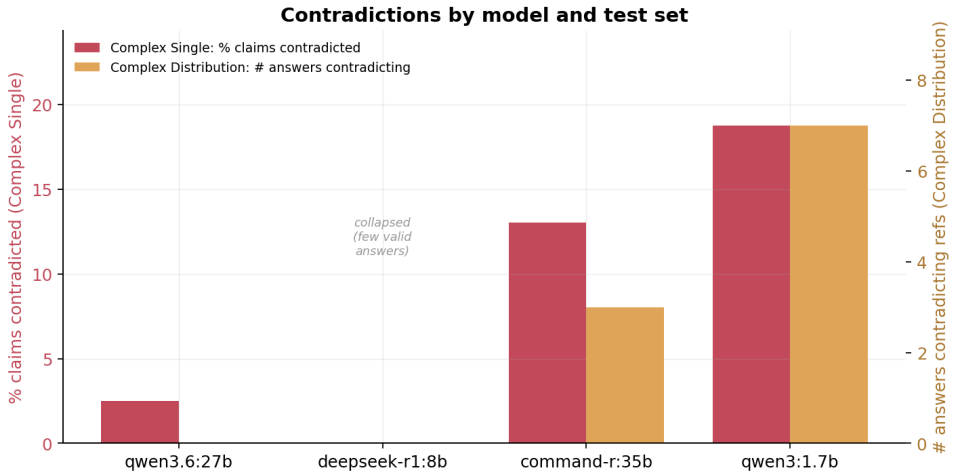

Figure 2: Contradiction burden differs materially across agent models and test sets.

Figure 3 shows semantic geometry for the Complex Single responses. Qwen3.6:27b had the lowest reported frameshift from the references at 0.108, compared with 0.230 for DeepSeek-r1:8b, 0.142 for command-r:35b, and 0.117 for qwen3:1.7b. DeepSeek-r1:8b answered only 3 of 10 Complex Single questions in the underlying results, so its geometry does not represent equivalent coverage. Together, the three figures show that local models are not interchangeable for governance use: retrieval success, bounded interpretation, contradiction behavior, and semantic stability change with the model at the core of the same agent architecture.

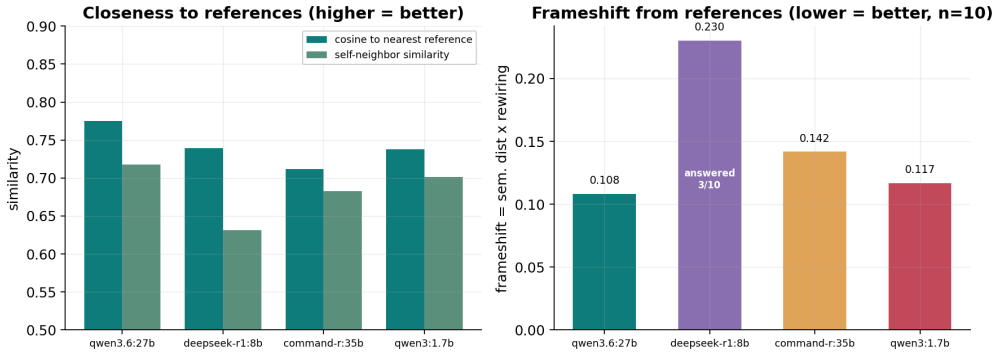

Figure 3: Semantic closeness and frameshift quantify how Complex Single responses relate to their references.

Taken together, the three test tiers support the governance capability of the Qwen3.6:27b based TwinPath AI agent in the evaluated setting. The agent retrieved the deterministic evidence correctly across the full SQL test set, led the tested models on bounded single-reference interpretation and acceptable-range reasoning, produced fewer reference contradictions, and remained closest to the reference semantic region by the reported frameshift measure. These are complementary tests of the functions required to govern prediction care alignment: reaching the correct evidence, preserving its factual content, reasoning within defined governance bounds, and remaining stable enough for repeated analytic use. The demonstrated capability is retrospective governance analysis over the TwinPath evidence substrate rather than autonomous clinical decision making. The distinction preserves the deterministic origin of the governance state while establishing that the agent can make that state and its supporting evidence practically investigable across individual patients and populations.

5 Ethical Considerations

TwinPath AI is designed for governance review rather than autonomous clinical decision making. Potential Gap / Needs Review is intentionally a review state rather than a claim that care was inappropriate. Structured EHR data can omit bedside actions, clinical rationale, device use, and documentation outside the queried tables, so the system cannot infer that missing structured evidence equals missing care. The agent does not prescribe medication, determine clinical contraindications beyond encoded evidence, or override clinician judgment. The architecture can support subgroup governance because prediction metrics, governance state frequencies, persistence, and review burden can be computed separately across age, sex, race and ethnicity, admission type, ICU type, and medical or surgical populations. All reported experiments used de-identified MIMIC-IV data. A provider deployment would require minimum-necessary data use, role-based access, encrypted storage and transmission, audit logging, retention limits, and compliance with applicable organizational and legal requirements. TwinPath AI does not require unrestricted access to the complete clinical record.

The governance engine can operate on a bounded evidence packet containing only the variables necessary for the risk calculation and care-pathway state. Local inference can reduce external data transfer, but production deployment would still require institutional controls for access, logging, security, privacy, and model governance. TwinPath AI is a clinical-governance and quality-review system, not an autonomous treatment system. It does not diagnose VTE, prescribe prophylaxis, label care as negligent, or override clinicians. Its outputs are intended to be reviewed by authorized clinical or quality personnel.

6 Limitations and Future Work

The study is retrospective and uses one public EHR dataset. The VTE outcome is a nonprimary diagnosis coding proxy rather than a time stamped incident event, so the present study does not establish lead time before confirmed hospital acquired VTE. The development split keeps patient days from one admission together but does not establish patient level separation across repeated admissions. The calibration procedure and the threshold used for the reported binary accuracy, precision, and recall also require fuller documentation for independent reproduction. These limitations primarily bound claims about the predictive input and downstream clinical effects; they do not change the observed agent benchmark results against the defined governance database and reference sets.

The governance layer is bounded by structured evidence. A review state indicates that the observed record does not satisfy the deterministic alignment rule; it does not prove clinical error. Complex Single and Complex Distribution references were author developed from governance definitions and computed database facts rather than independently clinician adjudicated. The validation is nevertheless more rigorous than a single reference-answer comparison: deterministic questions are checked against gold SQL result sets, numerical facts are recomputed from gold SQL, interpretation and contradictions are scored separately, repeated runs are evaluated for stability, and reference-relative semantic measures test movement within and away from the acceptable answer space. The current results therefore support governance capability within the defined retrospective environment, while clinician adjudication is required to extend that claim to clinical validity. Dashboard usability, alert burden, subgroup effects, review time, and prospective patient outcomes were not evaluated. The present study validates TwinPath AI as a retrospective governance architecture but does not yet establish the clinical validity or operational utility of its alignment states. Governance classifications are assigned deterministically, while the LLM provides evidence retrieval, synthesis, and interrogation rather than the underlying governance decision. The current high-risk threshold and governance-state definitions have not been independently clinically validated, and the MIMIC-IV VTE label does not provide reliable event-onset timing, limiting causal or lead-time interpretations of the longitudinal risk path.

Future work should retain deterministic state assignment while engaging clinicians and quality-improvement personnel to adjudicate a stratified sample spanning all six governance states, refine clinically meaningful risk thresholds and exception criteria, and determine whether identified states correspond to conditions that warrant review in real clinical workflows. Multi-institution replication should test whether the alignment rules, thresholds, and agent benchmark transfer across different EHR implementations and documentation cultures. Planned evaluation should first observe the system in real clinical settings without influencing care, measuring clinician agreement, review burden, unresolved gaps, and differences across patient groups. A true lead-time analysis will require a data source with reliable VTE onset timing and clearer temporal separation between risk assessment and subsequent care. The same governance architecture can subsequently be evaluated in care pathways such as sepsis response, pressure injury prevention, glycemic management, and surgical site infection prevention.

Conclusion

TwinPath AI addresses a growing problem in modern healthcare AI: deployment creates an obligation to govern not only the model, but also the relationship between model identified risk and the care process that follows. The framework makes that relationship observable by pairing a longitudinal risk path with a longitudinal documented care path, assigning deterministic alignment states, and placing an LLM agent over the resulting evidence for retrieval, synthesis, and investigation. The predictive model provides a credible foundation at substantial scale, with 541,469 admissions, more than 2.7 million patient days, calibrated AUROC 0.8299, and AUPRC 0.2332. The centerpiece is the governance layer and the agent that makes it operational. Qwen3.6:27b retrieved the correct evidence for all 50 deterministic SQL questions and led the tested models on the two interpretive benchmarks. Its evaluation did not depend on one LLM judge: gold SQL result matching, database-computed fact checks, claim-level contradiction analysis, repeated-sample stability, acceptable-range membership, and reference-relative semantic geometry examine different ways a governance answer can fail. The combined results support the claim that the agent delivers a measurable capability for accessing, interrogating, and interpreting an auditable governance representation. It can interrogate auditable longitudinal states, recover their supporting evidence, and synthesize that evidence while remaining measurably bounded by the governance references. This separation provides a practical basis for continuous oversight while preserving human review and a deterministic source of truth for the high consequence governance state.

AI Usage Statement: Generative AI tools, including ChatGPT, Claude, Gemini, and Gamma AI, were used to support code development and debugging, technical documentation, writing and presentation refinement, image generation, and slide development. Palantir Foundry was used for dashboard development, and ChatGPT supported development of the interactive demonstration. All AI assisted content, code, analyses, and presentation materials were reviewed and validated by the project team, which retains responsibility for the final work.

References

[1] Office of the National Coordinator for Health Information Technology, "Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing Final Rule," 2024. Online resource.

[2] U.S. Food and Drug Administration, "Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations," Draft Guidance, Jan. 2025. Online resource.

[3] Joint Commission, "Responsible Use of AI in Healthcare Certification," Jun. 2026. Online resource.

[4] A. Pandor et al., "Risk assessment models for venous thromboembolism in hospitalised adult patients: A systematic review," *BMJ Open*, vol. 11, no. 7, e045672, 2021.

[5] A. Liu, K. D. Varathan, V. Ramoo, and T. Gunarathne, "Machine learning prediction models for deep vein thrombosis in hospitalized patients: A systematic review and meta-analysis," *Frontiers in Medicine*, vol. 13, Art. no. 1849096, 2026.

[6] S. Russell, F. Montes Suros, V. Goodnow, and J. Zeng, "VTE Surveillance: AI Enabled Human Machine Teaming for Venous Thromboembolism," in *Proc. 31st Americas Conf. Inf. Syst. (AMCIS)*, Montreal, QC, Canada, 2025, Paper 1196.

[7] A. Johnson, L. Bulgarelli, T. Pollard, B. Gow, B. Moody, S. Horng, L. A. Celi, and R. Mark, "MIMIC-IV," PhysioNet, Version 3.1, Oct. 2024. Available: `https://doi.org/10.13026/kpb9-mt58`.

[8] A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, pp. 1–13, 2023.

[9] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[10] Y. Zhang, X. Ren, L. Liu, J. Zha, Y. Gu, and H. Ye, "Machine Learning Algorithms to Predict Venous Thromboembolism in Patients With Sepsis in the Intensive Care Unit: Multicenter Retrospective Study," *JMIR Medical Informatics*, vol. 14, Art. no. e80969, 2026. doi: 10.2196/80969.

[11] M. He, W. Liu, Z. Lu, Y. Lv, Q. Zhang, X. Jin, and P. Han, "Development of an Interpretable Machine Learning Model for Predicting Venous Thromboembolism in Intensive Care Unit Patients With Intracerebral Hemorrhage," *Frontiers in Neurology*, vol. 16, Art. no. 1691549, 2026. doi: 10.3389/fneur.2025.1691549.

[12] H. Qi, L. Li, J. Fang, T. Pei, A. Li, Z. Ding, and T. Chen, "Development and Validation of an Interpretable Machine Learning Model for Predicting Venous Thromboembolism in ICU Patients With Traumatic Brain Injury: A Multicenter Study," *World Neurosurgery*, vol. 202, Art. no. 124399, 2025. doi: 10.1016/j.wneu.2025.124399.

[13] C. Yang, H. Ma, X. Zeng, et al., "CatBoost Machine Learning Model for Thrombosis Risk Prediction in Critically Ill Cancer Patients: A MIMIC-IV Database Study," *Clinical and Applied Thrombosis/Hemostasis*, vol. 32, 2026. doi: 10.1177/10760296251408357.

[14] S. Russell, "When workflows stay stable but meaning moves in agentic analyst pipelines," in *Proc. 17th Int. Conf. Applied Human Factors and Ergonomics*, Istanbul, Turkiye, Jul. 2026, doi: 10.54941/ahfe1007675.

[15] S. Russell, "Semantic substrate theory: An operator theoretic framework for geometric semantic drift," arXiv preprint arXiv:2602.18699, Feb. 2026, doi: 10.48550/arXiv.2602.18699.