EPIDEMIOLOGY AND HEALTH,
ISSN / eISSN 2092-7193
Publisher KOREAN SOC EPIDEMIOLOGY, DEPT SOCIAL &
PREVENTIVE MEDICINE,Korean Red cross Hospital ,
SUNGKYUNKWAN UNIV. SCH MEDICINE,
Address : 50, Hyeoksin-ro, Wonju-si,Suwon ,16419, Republic of Korea
Volume 26-Issue 1-(2026)
DOI:https://doi-xx.org/1050/17703058999634

Metrological Reconstruction of VTE Risk Scales in Neurology: Reliability, Validity and SITraceable Interpretation

Ermeng Meng[1], Min Yu[1], Tianyu Xue[1,*]

[1] Department of Neurology, Dalian Third People's Hospital Affiliated to Dalian University of Technology, Dalian 116033, China

First author: Ermeng Meng, Email: mengerming@163.com

* Corresponding author: Tianyu Xue, Email: mailto:987460397@qq.com

Abstract

Neurology inpatients, particularly those with stroke, are routinely screened for venous thromboembolism (VTE) using risk assessment models (RAMs) such as Padua, Caprini, and IMPROVEDD. Differences in model structure, thresholds, and intended populations limit comparability across sites and hinder transparent quality improvement. This paper presents a patient-data-free metrological reconstruction framework that treats the "score-to-probability" mapping as a measurable quantity with a defined traceability chain and an explicit uncertainty expression. Parameters and risk strata are extracted from peerreviewed publications and used to rebuild a transparent mapping from RAM scores to event probability. Uncertainty is summarized following GUM concepts, distinguishing sampling error in published parameters from modelchoice and applicability sources, and is propagated via repeated sampling and sensitivity analyses. Cross-model calibration is performed on a virtual ward distribution derived from published summaries. Outputs include plainlanguage mappings, uncertainty ranges, and a minimal reporting dataset to support reproducibility and comparability. The framework requires no access to individuallevel data, questionnaires, or clinical records and is suitable for neurology quality improvement and methods reporting.

Keywords: venous thromboembolism; risk assessment model; metrology; uncertainty; traceability; neurology; calibration

1. Introduction

Preventing venous thromboembolism (VTE) is a core objective in neurology units, yet practice remains heterogeneous because hospitals deploy different risk assessment models (RAMs), adopt non-uniform cutoffs, and serve distinct case mixes; as a consequence, an identical clinical presentation may be labeled “high risk” in one ward and “intermediate risk” in another, complicating benchmarking, undermining protocol harmonization, and obscuring cross-site learning. A metrological lens offers a principled remedy. In measurement science, the quantity that guides action—the measurand—must be explicitly defined, linked to reference information through a traceability chain, and accompanied by an uncertainty statement. Applied to VTE RAMs, the actionable output is not the raw score but the implied probability of VTE over a specified observation window, which should be presented as a measurement result with documented provenance and uncertainty rather than a deceptively precise point value.

This paper introduces a patient-data-free framework that reconstructs score-to-probability mappings from peer-reviewed sources, expresses uncertainty in terms aligned with widely adopted metrology guidance (separating sampling variability from structural and applicability components), and calibrates RAMs against one another on a virtual neurology-ward distribution derived from published summaries. The framework yields plain-language probability ranges, cross-model equivalences that clarify when different RAMs convey comparable risk information, and a minimal reporting dataset that records model versions, parameter provenance, uncertainty ranges, and threshold sensitivity in a compact, audit-ready form. By design, the approach avoids collection or handling of individual-level information, requires no questionnaires or access to clinical records, and relies on transparent steps that can be independently rerun and versioned, thereby lowering adoption barriers while preserving rigor, reproducibility, and SI-traceable interpretation. Positioned as a methods contribution rather than a clinical study, it provides a common measurement language for ward-level decision support, facilitates safe migration between RAMs, and enables quality programs to compare performance across departments or institutions without exchanging patient records, ultimately supporting more consistent and defensible prophylaxis policies in neurology care.

2. Metrological Problem Statement

Measurand. The quantity of interest is the probability that a neurology inpatient experiences venous thromboembolism (VTE) within a clearly defined observation window—typically the duration of hospitalization or a fixed post-admission horizon such as 90 days—reported in percent. The measurand is conceived as a ward-level decision aid rather than an individual prognosis tool and is presented as an interval with an accompanying uncertainty description. Events covered by the measurand follow the definitions used in the source publications (for example, symptomatic deep vein thrombosis or pulmonary embolism confirmed by standard diagnostics); when sources differ, applicability notes explicitly state the operational definition used in the reconstruction.

Scope and users. The framework addresses stroke units and general neurology wards where RAM outputs influence prophylaxis policies, escalation pathways, and quality indicators. Intended users include bedside clinicians, nurse leaders, anticoagulation stewardship teams, hospital quality programs, and methods-oriented researchers.The framework is not positioned for outpatient, peri-procedural, obstetric, or surgical populations, nor for predicting treatment effects. It serves as a common, auditable measurement layer that can be embedded into existing governance processes without altering clinical accountability or local protocols.

Inputs. Reconstructions rely solely on publication-level inputs: (i) RAM structures and variable definitions; (ii) point-based rules or regression coefficients; (iii) baseline risks for the development cohort or validation cohorts; and (iv) event rates reported for score bands or risk strata. Preference is given to peer-reviewed derivation papers and high-quality external validations with transparent reporting of precision (for example, confidence intervals). When multiple credible sources exist, they are catalogued, differences are noted, and alternative mappings are retained rather than collapsed, so that structural uncertainty is preserved and later communicated.

Traceability. Published parameters are treated as reference inputs and assembled into a unified, versioned measurement workflow composed of: (1) source selection with inclusion rationale; (2) parameter extraction into a structured table that records citation, page/table location, and any caveats; (3) reconstruction, which turns scores into probability ranges using direct alignment (when stratum rates are available) or simple monotone fitting (when only point systems and baseline risks are reported); and (4) reporting, which produces probability intervals, uncertainty descriptions, and cross-model equivalences. Each artifact (table, script, and output) carries a version identifier and change log so another site can reproduce the same result from the same public sources.

Reporting conventions. Probabilities are reported in percent with consistent rounding (for example, to one decimal place for intervals when helpful). Uncertainty is communicated as ranges or as expanded uncertainty with a plain-language coverage statement (for instance, “a range that is expected to contain most plausible values under the published inputs”). We avoid specialized symbols and formulas so that tables and narrative statements remain accessible to multidisciplinary clinical teams. When inputs from different papers imply different time windows or event definitions, these differences are labeled clearly, and cross-window comparisons are avoided unless justified in text.

Assumptions and boundaries. The method assumes that published parameters are a reasonable proxy for the local ward context or, when they are not, that differences can be modeled as an uncertainty source. It does not: discover new predictors; refit models to local data; estimate treatment effects; or replace external validation on real-world outcomes. Instead, it provides a transparent baseline for interpretation and comparison that can later be linked to local audits if available. When published information is too sparse or inconsistent to support a defensible mapping (for example, conflicting strata with no precision estimates), the framework withholds a numerical mapping and reports the gap explicitly.

Uncertainty taxonomy and propagation. To keep interpretation practical, uncertainty sources are grouped into (a) Type A components, tied to sampling variability reported in the literature, and (b) Type B components, tied to structural choices (direct vs. indirect mapping, monotonicity enforcement), applicability to neurology inpatients (transportability), threshold definitions, and incomplete reporting. Propagation is conducted through repeated sampling and scenario analysis over plausible ranges implied by the sources, yielding probability intervals along with a qualitative breakdown (high/medium/low) of dominant contributors.

Quality controls and acceptance checks. Reconstructions are subjected to simple, auditable checks: monotonicity of the score-to-probability curve; plausibility relative to reported baseline risks; stability of the interval under small perturbations of inputs; and agreement with external validation summaries where available. Failure of these checks triggers re-examination of sources and, if necessary, labeling of the mapping as provisional.

Outputs and minimal reporting dataset (MRD). The section’s outputs are measurement-ready statements: (i) a score-to-probability mapping expressed as intervals with coverage language; (ii) an uncertainty summary that distinguishes Type A and Type B components; (iii) cross-model equivalences on a virtual neurology-ward scenario set; and (iv) an MRD capturing model version, provenance, observation window, uncertainty ranges, recommended threshold intervals with sensitivity notes, and version identifiers for the parameter table and scripts. All outputs are free of patient-level information, enabling adoption without ethics review while maintaining rigor, reproducibility, and SI-traceable interpretation.

3. Methods

3.1 Literature parameter extraction

We conduct a structured search of peer-reviewed literature to identify VTE RAM derivation studies and high-quality external validations, screening titles/abstracts and full texts in duplicate and resolving any disagreements by discussion. For each eligible RAM, we extract a standardized set of items—model name and version, target population, variable definitions, scoring rules or coefficients, baseline risk, and event rates for pre-specified score bands or risk strata—together with any reported precision measures (for example, standard errors or confidence intervals). Extraction is performed independently by two reviewers using a shared codebook; inter-reviewer discrepancies are logged and reconciled, and every data element is traceable to a specific page, table, figure, or appendix so that the resulting parameter-extraction table functions as the first artifact in the traceability chain. As part of quality screening, we record study setting, inclusion and exclusion criteria, sample size, analytic choices, and whether neurology inpatients (including stroke subtypes) were explicitly represented; we also capture author-reported limitations—such as handling of missing data, outcome definitions, censoring, and assumptions that could affect transportability to neurology wards—and flag potential sources of bias. Data management follows auditable conventions: each row in the table carries versioning metadata (date, extractor initials), full provenance (citation with page/figure), observation window, and concise applicability notes (e.g., exclusion of hemorrhagic stroke, peri-operative cohorts, or ICU-only samples). This structure supports straightforward updates as new publications appear, enables clear precedence rules when multiple reports exist for the same RAM (preferring final peer-reviewed versions and larger validations), and provides a stable foundation for downstream reconstruction, uncertainty analysis, and cross-model calibration.

3.2 Score-to-probability reconstruction

We reconstruct the relationship between each RAM's total score and the probability of VTE using only information reported in peer-reviewed sources. When a publication provides event rates for predefined score bands or risk strata, we align those bands to the stated score ranges and transcribe the mapping verbatim, adding brief notes on the observation window and outcome definition so users can see exactly what the probabilities refer to. When only weights or point rules are available, we pair them with published baseline risks and stratum summaries to fit a simple monotone score-to-probability curve—favoring transparent procedures (for example, a smooth monotone fit or isotonic smoothing) and verifying plausibility against the reported baseline and known clinical gradients. Each reconstruction undergoes basic acceptance checks: monotonicity across the full score range, coherence with any reported stratum rates, and stability under small perturbations of inputs. If multiple plausible mappings arise—because strata overlap, summaries conflict, or alternative baseline risks are reasonable—we preserve them as parallel "alternative mappings" and carry them forward explicitly as a source of structural uncertainty rather than forcing a single specification. We do not mix time horizons or outcome definitions; instead, mappings are labeled with their observation window (e.g., in-hospital vs. 90-day) and event definition (e.g., symptomatic DVT/PE) and are compared only within like-for-like settings unless clearly justified. Every mapping is versioned, linked to its source set in the extraction table, and accompanied by a concise change log so that another site can reproduce the same result from the same public inputs.

3.3 Uncertainty evaluation

We classify and communicate uncertainty in a way that is both faithful to the literature and practical for ward-level decisions. First, we separate Type A components—sampling variability reported by the source studies, visible in the width of confidence intervals for event rates or coefficients—from Type B components, which arise from model-choice effects (for example, using a direct stratum-to-score mapping versus an indirect fit), transportability gaps between the derivation cohort and a typical neurology ward, threshold definitions, and incomplete reporting. To propagate uncertainty, we perform repeated sampling over the plausible ranges implied by the publications: baseline risks are drawn within their reported intervals; stratum rates and coefficients are varied within their documented precision; and, where multiple credible mappings exist, we sample across those alternatives as well, treating them as parallel specifications rather than forcing a single choice. Each repetition yields a probability for a given score; aggregating many repetitions produces a distribution that we summarize as a confidence range and, where helpful, as an expanded uncertainty with a plain-language coverage statement (for example, "a range expected to contain most plausible values given the published inputs"). We check convergence heuristically (stability of the range as repetitions increase) and document any remaining instability as part of the result. We then provide sensitivity views that isolate drivers of variation: one-at-a-time changes to individual inputs (e.g., baseline risk, a high-weight item, or the mapping choice) and scenario analyses that combine inputs to mirror salient subgroups such as immobile stroke presentations—without invoking any patient-level data. Finally, we rank contributors to spread using simple labels (high/medium/low) and note directional effects (widening vs. upward shift), enabling clinical teams to see which assumptions matter most, how robust their thresholds are to reasonable variation, and where additional published evidence would most efficiently reduce uncertainty.

3.4 Virtual neurology-ward distribution

To compare RAMs without accessing patient records, we construct a scenario set that approximates a typical neurology ward using only published summaries—such as age distribution, the proportion with ischemic versus hemorrhagic stroke, prevalence of immobility, and common comorbidities (for example, diabetes, heart failure). The goal is not to represent every site but to provide a common yardstick for method comparison. Each scenario is a brief, non-identifiable description (e.g., "older adult with acute ischemic stroke and immobility," "middle-aged patient with transient ischemic symptoms and preserved mobility") paired with an observation window (in-hospital or 90-day) and a short note on inclusion assumptions. Scenarios are selected to span routine ward heterogeneity—age bands, mobility status, stroke subtype, and a small set of coexisting conditions—while avoiding excessive granularity that would imply individual records. We document the provenance of each assumption (citation and page), version the scenario catalog, and include applicability notes when a RAM explicitly excludes a subgroup. The scenario set is then used to apply each reconstructed RAM and generate probability ranges and risk classifications, enabling cross-model calibration and plain-language equivalence statements. We periodically review scenarios as new summaries are published, adding or retiring items through a simple change log. This construct is strictly methodological: it supports comparison, sensitivity analysis, and threshold planning, but it is not a forecast for any specific patient or institution and therefore requires no consent, data-sharing agreements, or de-identification procedures.

3.5 Cross-model calibration and comparison

We apply each reconstructed RAM to the full scenario set and, for every scenario–score combination, compute the corresponding probability range and the resulting risk classification under the RAM's native thresholds. Using these outputs, we create plain-language equivalences that help users translate between models—for example, "within the acute ischemic-stroke with immobility scenarios, Padua score bands that trigger a high-risk label yield probability intervals that substantially overlap those produced by the corresponding IMPROVE-DD high-risk bands (see Table 4)". Equivalences are generated only after aligning observation windows and outcome definitions and are expressed as intervals rather than single numbers to preserve uncertainty. We document how the equivalence was determined (e.g., overlap of the central portions of probability intervals across multiple scenarios with consistent classification) and include concise notes on limits of interchangeability.

3.6 Threshold sensitivity

Because prophylaxis policies depend on thresholds, we perform a targeted threshold-sensitivity analysis to test how small, realistic shifts in scores or cutoffs alter the share of scenarios labeled high risk, using only the reconstructed mappings and the virtual ward scenarios. For each RAM and observation window, we first record the baseline high-risk classification rate across the scenario set, then perturb either the decision threshold (e.g., move the cutoff up or down by one point) or the effective score (e.g., emulate routine variation from documentation differences or inter-rater judgment) and recompute the classification rate under identical conditions. The comparison yields a direction (increase or decrease in high-risk flags) and an approximate magnitude (percentage-point change), which we summarize with uncertainty ranges derived from the same repeated-sampling procedure used in the reconstruction step. We pay particular attention to "steep zones" of the score-to-probability curve, where small numerical changes translate into large swings in classification, and we note any scenarios in which probability intervals straddle the threshold—conditions under which policy outcomes are most fragile. Results are reported in plain language -for example, "a one-point decrease in the cutoff increases the share of scenarios flagged as high risk by a small but policy-relevant margin (see the sensitivity summary in Table 3 and the scenario-level appendix)”.These outputs help teams gauge whether local policies possess a practical margin of safety against routine measurement variation and inform simple mitigations—such as adopting threshold *bands* instead of single cutoffs, adding brief confirmation steps when scores fall within a narrow buffer around the decision point, or scheduling more frequent mapping reviews in steep zones—without relying on patient records or complex statistical machinery.

3.7 Quality assurance and reproducibility

To promote full reproducibility and auditability, we make publicly available the parameter-extraction table and the accompanying, human-readable scripts that carry out reconstruction, uncertainty propagation, cross-model comparison, and threshold-sensitivity analyses. The repository includes a concise README describing inputs, expected outputs, and step-by-step execution, plus lightweight tests that verify monotonicity checks and consistency with cited baseline risks. A second researcher, independent of the initial extraction, re-runs the entire workflow from the released materials to confirm that the same tables and narrative statements are generated from the same published inputs; any discrepancies trigger a documented adjudication process and, if needed, corrective commits. All artifacts—tables, scenarios, scripts, and rendered outputs—are tracked under version control with semantic version tags and a brief change log that records what changed, why, and which sources were affected. Releases pin citation DOIs and page locations to guard against drift. When the literature is updated (for example, a new external validation reports revised strata), we update only the relevant rows in the extraction table, re-run the workflow, and issue a new minor or patch version; material alterations to model definitions or observation windows increment the major version. This disciplined approach ensures that other sites can reproduce results byte-for-byte, trace conclusions to specific passages in the source papers, and adopt or adapt the workflow with clear provenance and minimal friction—without touching patient-level data at any stage.

4. Results

4.1 Inventory of included RAMs

An inventory table provides a structured overview of each included RAM—such as Padua, Caprini, and IMPROVE-DD—capturing the model name, acronym, referenced version and year, primary citation (with DOI), intended population and setting (general medical vs. neurology inpatients; ward vs. ICU), observation window (in-hospital or fixed-day horizon), and outcome definition used in the source (e.g., symptomatic DVT/PE). For reconstruction planning it records the input type available—full coefficients and a baseline risk; point system with stratum-level event rates; or mixed/other—and whether precision measures (confidence intervals, standard errors) are reported. The table flags neurology-specific representation (ischemic and/or hemorrhagic stroke cohorts in derivation or validation), lists native thresholds, and notes explicit exclusions (e.g., peri-operative patients). Quality and portability fields summarize sample size, reporting completeness, handling of missing data, and transportability notes relevant to neurology wards. A reconstruction status column indicates whether a direct mapping is feasible (verbatim stratum rates), an indirect monotone fit is required (weights plus baseline risk), or alternative mappings must be retained due to overlapping strata or inconsistent summaries; mismatched time windows or event definitions are highlighted to prevent inappropriate cross-window comparisons. Gaps that preclude defensible mapping (for example, absent baseline risk or undefined strata) are tagged as provisional with a short rationale. Together, these fields identify where reconstruction is straightforward and where assumptions are unavoidable, prioritize models for analysis, and provide a transparent starting point for uncertainty budgeting and cross-model calibration.The inventory of included models and parameter sources is summarized in Table 1.

Table 1. Included VTE risk assessment models and parameter sources

Model	Primary citation (year)	Intended population	Observation window	Native thresholds	Reported stratum event rates	Notes
Padua Prediction Score (PPS)	Barbar et al., 2010	Hospitalized medical inpatients	Up to 90 days post-admission	Low: <4; High: ≥4	Low: 0.3%; High: 2.2% (with prophylaxis) / 11% (without prophylaxis)	Derivation cohort was general medical; add applicability note for neurology wards.
IMPROVE VTE Risk Score	Spyropoulos et al., 2011 (IMPROVE)	Acutely ill medical inpatients	In-hospital to ~90 days	Low: 0–1; Moderate: 2–3; High: ≥4	Score 0: 0.5%; Score 1: 1.0%; 2–3: 1.9%; ≥4: 5.0%	Ternary scheme; baseline risk varies by cohort.
IMPROVE-DD (IMPROVE + D-dimer)	Spyropoulos et al., 2020 (TH Open)	Acutely ill medical inpatients with D-dimer integration	42 days and 77 days (per study)	Low: 0–1; At-risk: ≥2 with elevated D-dimer	42-day: Low 0.39% vs At-risk 1.11%; 77-day: Low 0.91% vs At-risk 2.22%	Definition of 'at-risk' depends on D-dimer threshold and assay; align to source.
Caprini Risk Assessment Model (surgical)	Caprini 2005; Pannucci 2017 meta-analyses	Surgical inpatients	30–60 days (varies)	Common groupings: 3–4; 5–6; 7–8; ≥9 (very high)	3–4: 0.7%; 5–6: 1.8%; 7–8: 4.0%; ≥9: 10.7%	Designed for surgical cohorts; use as scale reference in neurology.

4.2 Reconstructed score-to-probability mappings

For each RAM, we present a compact, bedside-ready mapping that links clearly defined score bands to probability ranges rather than single point estimates, accompanied by brief applicability notes that state the observation window, outcome definition, and any population caveats (e.g., "derived primarily from general medical inpatients; interpret with caution in hemorrhagic stroke”). When the literature supports more than one credible reconstruction—for example, because strata overlap or baseline risks differ across sources—we display parallel mappings side-by-side, label them unambiguously, and explain in a sentence what distinguishes them. Each range includes a plain-language coverage statement (such as “expected to contain most plausible values given published inputs”), and the table avoids specialized symbols so it can be dropped into protocols without additional explanation or calculation. To prevent misuse, we align mappings only within the same time horizon and event definition, highlight native thresholds used in the source studies, and flag “steep zones” where small score changes can shift classification. Footnotes capture exclusions, transportability notes, and precision limitations, while a version tag ties the row back to the parameter-extraction table and change log. The result is a readable, auditable tool that clinicians can consult during rounds or embed in policy documents to translate scores into comparable probability intervals with their key assumptions in view.The reconstructed score-to-probability mappings for all included RAMs are presented in Table 2.

Table 2. Reconstructed score-to-probability mappings by score band and observation window.

Model	Score band / stratum	Observation window	Probability (%)	Source / notes
Padua	<4 (Low)	≤90 days	0.3	Barbar 2010 (low-risk event rate)
Padua	≥4 (High), with prophylaxis	≤90 days	2.2	Barbar 2010 (on-prophylaxis subgroup)
Padua	≥4 (High), without prophylaxis	≤90 days	11.0	Barbar 2010 (no-prophylaxis subgroup)
IMPROVE	0	≤90 days	0.5	Spyropoulos 2011 / guidance summaries
IMPROVE	1	≤90 days	1.0	Spyropoulos 2011 / guidance summaries
IMPROVE	2–3	≤90 days	1.9	Spyropoulos

				2011 / guidance summaries
IMPROVE	≥4	≤90 days	5.0	Spyropoulos 2011 / guidance summaries
IMPROVE-DD	0–1 (Low)	42 days	0.39	Spyropoulos 2020 (TH Open)
IMPROVE-DD	≥2 + elevated D-dimer (At-risk)	42 days	1.11	Spyropoulos 2020 (TH Open)
IMPROVE-DD	0–1 (Low)	77 days	0.91	Spyropoulos 2020 (TH Open)
IMPROVE-DD	≥2 + elevated D-dimer (At-risk)	77 days	2.22	Spyropoulos 2020 (TH Open)
Caprini (surgical)	3–4	30–60 days	0.7	Pannucci 2017 meta-analysis
Caprini (surgical)	5–6	30–60 days	1.8	Pannucci 2017 meta-analysis
Caprini (surgical)	7–8	30–60 days	4.0	Pannucci 2017 meta-analysis
Caprini (surgical)	≥9	30–60 days	10.7	Pannucci 2017 meta-analysis

4.3 Uncertainty summary

A summary table consolidates the main uncertainty sources for each RAM. Typical patterns include moderate Type A uncertainty due to limited event counts in published strata and notable Type B uncertainty when the development cohort differs from neurology wards. For each source we indicate whether it primarily widens the interval, shifts it upward or downward, or interacts with specific subgroups. This table functions as a concise risk register for measurement-related uncertainty.A consolidated summary of Type A and Type B uncertainty sources is shown in Table 3.

Table 3. Uncertainty summary: Type A (sampling) and Type B (structural/applicability)

Model	Type A (sampling)—typical sources	Type B (structural/applicability)—typical sources

Padua	Limited events per stratum; published CI widths.	General medical derivation; threshold ≥4; effect of prophylaxis; transportability to neurology wards.
IMPROVE	Registry-derived estimates; moderate discrimination.	Ternary thresholds; baseline risk and length-of-stay differences; neurology transportability.
IMPROVE-DD	Rates depend on D-dimer assay and cutoffs.	Combined score+biomarker rule; 42/77-day windows; cohort differences.
Caprini (surgical)	Heterogeneity across surgical specialties.	Surgical-only design; limited direct applicability to neurology inpatients.

4.4 Cross-model equivalences

A cross-model table summarizes qualitative and semi-quantitative equivalences between RAMs applied to the virtual ward scenarios, translating score bands in one model to probability ranges in another after aligning observation windows and outcome definitions. For each scenario family (e.g., acute ischemic stroke with immobility; non-stroke neurology with preserved mobility) we record the proportion of scenarios in which the probability intervals overlap and classify equivalence as strong (overlap across most scenarios with consistent risk classification), conditional (overlap confined to certain subgroups or only within midrange scores), or weak (intervals rarely overlap or native thresholds imply divergent classifications). Equivalence entries include brief notes on why alignment holds or fails—such as differences in baseline risk, excluded populations, or “steep zones” where small score shifts change classification—and indicate any non-interchangeable regions where uncertainty bands do not intersect. Where relevant, we add threshold translations (e.g., “Padua ≥X corresponds to Caprini ≥Y in these scenarios”) expressed as intervals with plain-language coverage statements. Each row carries version tags and links to the parameter-extraction table so sites can audit provenance and update as new publications appear. These statements give organizations a practical map for migrating between RAMs or harmonizing protocols across departments while keeping assumptions and limits explicit.Cross-model equivalence statements on the virtual neurology-ward scenarios are reported in Table 4.

Table 4. Cross-model equivalences on virtual neurology-ward scenarios.

Scenario family	Equivalence statement	Equivalence strength	Notes
General medical inpatient (non-neurology)	Padua ≥4 (2.2% with prophylaxis; 11% without) approximates IMPROVE ≥2 (1.9% for 2–3; 5.0% for ≥4) in the ~2–5% range when prophylaxis is typical.	Conditional	Overlap stronger when comparing Padua ≥4 to IMPROVE ≥4 in non-prophylaxis contexts; apply caution for neurology transportability.
Acutely ill medical with elevated D-dimer	IMPROVE-DD at-risk (≥2 + elevated D-dimer) corresponds to upper IMPROVE strata (≥4), with observed rates ~1–2.22% at 42–77 days.	Conditional	Depends on D-dimer threshold and assay; align observation windows.
Surgical inpatient (scale reference)	Caprini 7–8 (≈4%) or ≥9 (≈10.7%) generally exceeds typical medical-model risk bands under routine prophylaxis.	Weak	Caprini is not intended for medical/neurology inpatients; shown for scale only.

4.5 Minimal reporting dataset

We define a minimal reporting dataset (MRD) to standardize communication and enable independent replication without relying on patient-level data. The MRD records the exact model version and source (citation, DOI, observation window, outcome definition) and includes a concise narrative description of the reconstructed score-to-probability mapping, with a direct pointer to the parameter-extraction table that lists page/table locations for every input. It summarizes the main uncertainty sources and ranges, separating sampling variability from structural or applicability components and providing plain-language coverage statements. It also lists recommended threshold intervals with sensitivity notes, indicating how small score or cutoff shifts (±1–2 points) change high-risk classification rates in the virtual ward scenarios and highlighting any "steep zones" where decisions are fragile. Finally, the MRD links to the scripts and version information, including a change log that documents what changed and why, so another site can reproduce the same tables and narrative statements from the same public sources. Compact enough to sit in a methods appendix yet complete enough for replication and audit, the MRD offers a uniform, SI-aligned record of provenance, assumptions, uncertainty, and threshold guidance that wards can embed directly into protocols, dashboards, and quality reports.

4.6 Implementation vignette

To illustrate use, consider a neurology ward that currently reports high-risk rates based on Padua. Using the reconstructed mapping, the team converts scores to probability intervals and adds uncertainty bands to protocol documents. They compare these intervals with those produced by IMPROVE-DD for the same scenarios and observe substantial overlap for common patient types, enabling harmonized thresholds across two hospital sites that use different RAMs. The MRD is appended to their protocol, and quality dashboards begin to display probability intervals rather than binary labels. No patient records were accessed to accomplish these steps.

5. Discussion

This work reframes VTE risk scoring as a measurement problem and applies metrological principles—explicit definition of the measurand, traceability to reference information, and transparent uncertainty—to a domain that is often treated as purely statistical. By converting score outputs into probability intervals with documented provenance, the framework establishes a common, audit-ready scale that can be used across neurology wards regardless of the specific RAM in use.

Transparency. The approach distinguishes variability arising from sampling error in the literature from variability introduced by structural choices and applicability assumptions. Clinicians and quality teams can see how much each component contributes to the final probability range, reducing reliance on single point estimates whose precision is unclear and clarifying why two sites may report different "risk rates" even when their case mix appears similar.

Consistency. Threshold policies become more robust when cutoffs are interpreted alongside uncertainty bands. Recognizing when a decision point lies within a "steep zone" of the score-to-probability curve helps teams introduce practical safeguards—such as buffer zones, confirmation steps, or scheduled reviews—so that small variations in scoring, documentation, or patient mix do not produce outsized swings in classification rates.

Comparability. Converting scores to probability intervals enables like-for-like reporting across sites using different RAMs. Cross-model equivalence statements provide a plain-language map for protocol harmonization and migration planning, while the minimal reporting dataset (MRD) standardizes documentation of versions, assumptions, uncertainty ranges, and threshold sensitivity. Together these elements support more defensible benchmarking and clearer communication with governance bodies.

Ethical and privacy considerations. All inputs are drawn from published literature, and all analyses are scenario-based. The framework does not access or generate patient-level data and therefore requires no consent procedures, data-sharing agreements, or de-identification pipelines. This design lowers adoption barriers for smaller hospitals and facilitates rapid uptake in resource-constrained settings without compromising rigor.

Limitations. Reconstruction quality is bounded by the completeness and clarity of the source literature. Some RAMs provide limited detail, necessitating wider intervals or parallel “alternative mappings.” External validations may pool populations that differ from neurology wards, affecting transportability. The virtual ward scenarios are intentionally simplified and may not capture complex multimorbidity or rare conditions. Finally, the framework does not replace external validation on real-world outcomes; instead, it supplies a transparent baseline that can later be linked to local audits when available.

Future work. The parameter table can be updated on a regular cadence as new derivations and validations appear, with versioning to preserve provenance. Sites that obtain aggregated outcome summaries can use the same structure to refine baseline risks while retaining the reconstruction and uncertainty apparatus. Packaging the workflow as lightweight software and embedding MRD fields into institutional templates and dashboards would further streamline use, improve consistency of reporting, and support periodic re-calibration without introducing patient-level data flows.

6. Conclusion

In summary, we present a metrology-oriented, patient-data-free framework for reconstructing and comparing VTE risk assessment models in neurology that treats the score-to-probability relationship as a measurable quantity with defined provenance and uncertainty. By explicitly defining the measurand (probability of

VTE over a stated window), documenting a transparent traceability chain from published parameters to reconstructed mappings, and reporting uncertainty as clear ranges with coverage statements, the approach places different RAMs on a common, audit-ready scale. The accompanying minimal reporting dataset standardizes how model versions, assumptions, thresholds, and sensitivity findings are communicated, enabling independent replication and straightforward updates as new literature appears. Applied to virtual ward scenarios, the framework supports cross-model calibration and plain-language equivalences, helping services harmonize protocols, plan migrations between RAMs, and interpret quality metrics consistently across departments and institutions. Because it relies solely on publication-level inputs and scenario analyses, the method avoids ethical and privacy risks associated with patient-level data, lowers barriers to adoption, and remains suitable for diverse settings, including resource-constrained hospitals. Taken together, these features enhance comparability, transparency, and reproducibility, providing neurology teams with a rigorous yet practical foundation for risk communication, threshold policy design, and continuous quality improvement.

7. Ethics statement

This methodological study uses only parameters, baseline risks, and risk-stratum rates reported in peer-reviewed publications and applies simulation-based analyses to synthetic scenarios. No recruitment, intervention, linkage, or access to medical records occurred; no individual-level data, identifiable information, or protected health information were collected, analyzed, or shared. Accordingly, the work qualifies as not human subjects research and does not require institutional ethics committee/IRB approval or informed consent. If requested by the journal or host institution, an administrative determination of exemption or NHSR status can be provided. No animal research was performed.

References

1.Barbar S, Noventa F, Rossetto V, et al. A risk assessment model for the identification of hospitalized medical patients at risk for venous thromboembolism: the Padua Prediction Score. *J Thromb Haemost.* 2010;8(11):2450–2457. doi:10.1111/j.1538-7836.2010.04044.x

2.Caprini JA. Thrombosis risk assessment as a guide to quality patient care. *Dis Mon.* 2005;51(2–3):70–78. doi:10.1016/j.disamonth.2005.02.003

3.Spyropoulos AC, Anderson FA Jr, FitzGerald G, et al; IMPROVE Investigators. Predictive and associative models to identify hospitalized medical patients at risk for venous thromboembolism. *Chest.* 2011;140(3):706–714. doi:10.1378/chest.10-1944

4.Spyropoulos AC, Lipardi C, Xu J, et al. Modified IMPROVE VTE risk score and elevated D-dimer identify a high VTE risk in acutely ill medical patients for extended thromboprophylaxis. *TH Open.* 2020;4(1):e59–e65. doi:10.1055/s-0040-1705137

5.Rosenberg D, Eichorn A, Alarcon M, et al. External validation of the IMPROVE VTE risk assessment model for medical patients in a tertiary health system. *J Am Heart Assoc.* 2014;3(6):e001152. doi:10.1161/JAHA.114.001152

6.Dennis M, Sandercock PA, Reid J, et al. European Stroke Organisation guidelines for prophylaxis for venous thromboembolism in immobile patients with acute ischaemic stroke. *Eur Stroke J.* 2016;1(1):6–19. doi:10.1177/2396987315621007

7.Powers WJ, Rabinstein AA, Ackerson T, et al. 2018 guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the AHA/ASA. *Stroke.* 2018;49(3):e46–e110. doi:10.1161/STR.0000000000000158

8.Joint Committee for Guides in Metrology (JCGM). *Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement (GUM).* JCGM 100:2008.

9.Joint Committee for Guides in Metrology (JCGM). *Evaluation of Measurement Data—Supplement 1 to the GUM: Propagation of Distributions Using a Monte Carlo Method.* JCGM 101:2008.

10.Joint Committee for Guides in Metrology (JCGM). *International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM), 3rd ed.* JCGM 200:2012.

11.Moumneh T, Kahn SR, Abou-Nassar K, et al. Validation of risk assessment models predicting venous thromboembolism in hospitalized medical patients. *J Thromb Haemost.* 2020;18(10):2585–2595. doi:10.1111/jth.14796

12.Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med.* 2015;162(1):55–63. doi:10.7326/M14-0697

13.Pannucci CJ, Swistun L, MacDonald JK, Henke PK, Brooke BS. Individualized Venous Thromboembolism Risk Stratification Using the 2005 Caprini Score to Identify the Benefits and Harms of Chemoprophylaxis in Surgical Patients: A Meta-analysis. Ann Surg. 2017;265(6):1094-1103. doi:10.1097/SLA.0000000000002126.