

Original Research Article

Development and Internal Technical Evaluation of a Transparent Dashboard for Indonesian Medical Malpractice Disputes: A Pilot Study

Dedi Kusmanto¹, Hendri Safitri¹, Firzan Dahlan^{2*}, Fachrudin Sani³¹Department of Forensic Medicine and Medicolegal, CMHC Research Center, Palembang, Indonesia²Department of Educational Science, Enigma Institute, Palembang, Indonesia³Department of Intensive Therapy, ANC Medical Center, Jakarta, Indonesia

ARTICLE INFO

Article history

Received: March 30, 2026

Revised: April 20, 2026

Accepted: May 13, 2026

Available online: May 16, 2026

Published: May 30, 2026

Keywords

Medical malpractice

Medicolegal informatics

Natural language processing

Explainable artificial intelligence

Dashboard

Corresponding author

Firzan Dahlan

E-mail: firzan.dahlan@enigma.or.id

ORCID: 0009-0003-1193-4431

DOI

10.59345/sjfm.v4i1.372

ABSTRACT

Background. Judicial assessment of suspected medical malpractice requires simultaneous consideration of clinical facts, documentation, professional standards, causation, organizational responsibility, and procedural proof.

Objective. To develop and technically evaluate a transparent local dashboard that structures these elements for education and research without substituting for expert or judicial judgment.

Methods. Sixteen de-identified Indonesian disputes (2006–2026) were purposively assembled, one record per dispute. Eight prespecified concepts were manually coded and compared with a regular-expression dictionary, free of outcome terms, applied to evidence notes (dictionary-based natural language processing [NLP]). Five features entered a ridge logistic model ($\lambda = 1$) evaluated by leave-one-dispute-out cross-validation. Six families of deterministic software tests were run.

Results. The corpus contained six granted and ten rejected outcomes. Across 128 concept decisions, the dictionary produced 33 true positives, 2 false positives, and 13 false negatives (micro-precision 94.3%, micro-recall 71.7%, micro-F1 81.5%; macro-F1 87.9%). Severe injury was positive in 15 of 16 disputes and had the lowest recall (0.40). Out-of-fold sensitivity was 5/6 and specificity 10/10; the Brier score was 0.050 with manual features and 0.061 with NLP-derived features. All 49 software tests passed.

Conclusion. The prototype provides traceable exploration of a small medicolegal corpus. Because one abstractor produced the labels and notes, and the predictors restate the courts' own grounds, the near-perfect internal separation is not evidence of prediction or external validity. Use should be limited to education, hypothesis generation, and abstraction quality assurance.

Copyright. © 2026 The Author(s). Published by Phlox Institute: Indonesian Medical Research Organization, on behalf of the Sriwijaya Journal of Forensic and Medicolegal. The authors retain copyright of their work.

Access licence. This is an open access article distributed under the terms of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International Licence (CC BY-NC-SA 4.0), which permits non-commercial use, sharing, adaptation and reproduction in any medium, provided the original author and source are credited and any derivative work is distributed under the same licence.

1. Introduction

Suspected medical malpractice is not reducible to the severity of an adverse clinical outcome. Review may consider whether a specific duty or standard was breached, whether the breach caused the alleged harm, whether informed consent and documentation were adequate, whether an organization contributed to the event, and whether procedural or evidentiary requirements were met. Claims research indicates that the defensibility of care and evidence of error are more closely associated with compensation than injury severity alone.¹ Registry and litigation studies from several countries also show substantial variation in claim populations, compensation structures, allegation patterns, and outcome definitions, which makes direct transfer of a single severity-based risk model inappropriate.^{2–5} Specialty-specific analyses add that communication, diagnosis,

procedural context, and care setting shape the pathway from an adverse event to a claim.^{6–11}

In Indonesia, professional accountability has been regulated by the Medical Practice Act (Law No. 29 of 2004)¹² and the Health Law (Law No. 36 of 2009).¹³ Law No. 17 of 2023 on Health consolidated and replaced these and related statutes.¹⁴ Alleged malpractice may therefore be pursued through civil claims for damages, criminal proceedings, administrative sanctions, and professional disciplinary review, and decisions from the courts are published in the Supreme Court Decision Directory (Direktori Putusan Mahkamah Agung).¹⁵ A corpus spanning 2006–2026 consequently crosses more than one statutory regime.

Public court decisions show how these questions are argued, but they are written to explain adjudication, not to serve as research datasets. A single clinical conflict may generate decisions at several judicial levels, terminology

varies across specialties and procedures, key facts may appear in the factual narrative, expert evidence, reasoning, or disposition, and public websites differ in accessibility. A conventional spreadsheet preserves citations but does not provide transparent concept extraction, subset-level summaries, case-level audit trails, or safeguards against overinterpretation. Recent legal natural language processing (NLP) work shows that document structure,¹⁶ rationale annotation,^{17,18} and the distance between model explanations and expert reasoning¹⁹ must be treated as empirical design questions.

Interactive dashboards can make such a corpus navigable, but user-centered studies emphasize navigability and task fit,^{20,21} evaluation with intended users,²² and continuing monitoring of machine-learning-based decision support²³ over visual polish. Explanations can also alter trust and reliance in different directions depending on explanation quality, workflow, and user expertise.^{24–28} Domain-specific legal language differs from general text,²⁹ and rationale-aware approaches may better support human review.¹⁷

A rule-based baseline suits a small corpus because every positive result can retain the phrase that activated the rule, changes can be versioned, and errors can be discussed concept by concept. Legal NLP experiments reinforce the need to test whether textual explanations are grounded in legally relevant reasons,³⁰ how large language models behave on statute and judgment tasks,³¹ whether privacy and provenance are preserved,³² how multinational legal language models are adapted,³³ and how far annotation can be trusted at the edge of human agreement.³⁴ Transparency in this setting has four dimensions: data transparency (a traceable link from record to source), measurement transparency (explicit concept definitions), computational transparency (visible rules and testable software invariants), and interpretive transparency (a clear statement of what the output can and cannot support).

This study integrates four elements that are often reported separately: dispute-level source provenance, a transparent concept-extraction layer, case-level decomposition of a regularized multivariable model, and deterministic testing of dashboard functions. The primary aim was to develop and technically evaluate a transparent dashboard for structured exploration of a pilot corpus of Indonesian medical malpractice disputes. Secondary objectives were to quantify concept-level concordance between dictionary-based NLP and manual coding, to describe the direction of five prespecified model contributions, to describe out-of-fold model behavior, and to check the deterministic functions of the dashboard. The dashboard is intended for education, hypothesis generation, and abstraction quality assurance; it is not a tool for adjudication or provider scoring.

2. Methods

2.1 Study design and reporting

This retrospective system-development pilot used a purposive corpus of de-identified Indonesian medical malpractice disputes dated 2006–2026. The unit of analysis was an independent dispute rather than a court document: decisions in a related appeal, cassation, or review sequence were treated as one record when they concerned the same clinical conflict, because counting each document would overrepresent disputes that progressed through more levels. The study was a technical and exploratory pilot, not a prevalence survey, a representative

sample of Indonesian litigation, or an evaluation of a tool for active cases.

The legal disposition recorded in the corpus (“granted” or “rejected”) was treated as the observed binary outcome for descriptive modeling; it was not treated as a clinical endpoint or as a measure of care quality. Reporting followed the applicable transparency principles of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD+AI) statement,³⁵ used as a transparency framework because the outcome was a legal disposition rather than a clinical prognosis.

2.2 Sources, corpus formation, and abstraction

Source material comprised the Supreme Court Decision Directory (Direktori Putusan Mahkamah Agung; putusan3.mahkamahagung.go.id),¹⁵ government legal-document repositories, decision indexes, and linked secondary summaries retained in the corpus manifest. Inclusion depended on source accessibility and on sufficient factual and procedural detail to identify the underlying clinical conflict, link related proceedings, assign a disposition, and complete the prespecified abstraction fields. The resulting corpus of 16 disputes is a purposive sample, not a census or probability sample.

Each record stored a study identifier, decision number, year, judicial level, clinical domain, binary disposition, coding confidence, a concise evidence note, eight binary concepts, and source URLs. Coding confidence summarized how directly the accessible material supported the abstraction: high when the information was explicit, moderate when coding required synthesis across less direct descriptions, and low when documentary detail was limited. It was displayed as a filter and descriptive attribute and was not used to delete records. Party names, addresses, contact details, and clinician identifiers were excluded from the analytical dataset and dashboard; study identifiers were used in the interface.

2.3 Operational concepts and manual reference coding

Eight prespecified binary concepts were coded at the dispute level: (1) an adverse professional, disciplinary, or ethical finding; (2) a specific breach accompanied by an asserted causal link; (3) a material informed-consent issue; (4) a documentation or medical-record problem; (5) organizational or supervisory failure; (6) an accepted defense based on standard operating procedure (SOP) or recognized medical risk; (7) a formal or evidentiary barrier; and (8) severe injury, including death or disability. A positive label indicated that the concept was materially represented in the evidence note and abstraction record; it did not quantify intensity or decide legal sufficiency. The framework is a research ontology for corpus navigation, not a substitute for full legal and clinical interpretation.

One abstractor created the evidence notes and manual labels within the same workflow, and no second independent annotator was available. Agreement between the dictionary and manual coding is therefore reported as internal abstraction concordance, not as validation against independently annotated judicial text.

2.4 Dictionary-based NLP

A visible regular-expression dictionary was applied only to the evidence-note field. Rules were built around phrases that express each operational concept, and outcome words such as “granted” and “rejected” were excluded to reduce direct outcome leakage. Because the notes were written

with knowledge of the disposition, the dictionary is described as free of outcome terms rather than outcome-blind. Each positive extraction retained its triggering text. The dictionary was written and scored on the same 16 notes, so its performance is a resubstitution result; no thresholds were tuned, and no held-out set was available.

Precision was the proportion of dictionary-positive decisions that were also manually positive; recall, the proportion of manually positive decisions detected; specificity, the proportion of manually negative decisions left negative; and F1, the harmonic mean of precision and recall. Macro metrics are the unweighted means of the eight concept-level values; micro metrics pool all 128 decisions (16 disputes $\times$ 8 concepts). Exact (Clopper-Pearson) 95% confidence intervals for proportions were calculated from the reported counts.

2.5 Exploratory regularized model

A ridge logistic model used five prespecified binary features: specific breach plus causation, adverse professional finding, organizational failure, accepted SOP/risk defense, and formal/evidentiary barrier. Severe injury, consent, and documentation were left for descriptive exploration. For dispute i with feature values x_{ij} , the linear predictor and fitted probability were

$$\text{logit}(\pi_i) = \eta_i = \beta_0 + \sum_{j=1}^5 \beta_j x_{ij}$$

$$\pi_i = \frac{1}{1 + \exp(-\eta_i)}$$

and coefficients were estimated by maximizing the log-likelihood minus a ridge penalty $\lambda \Sigma \beta_j^2$ with λ fixed at 1.

$$\ell_\lambda(\beta) = \sum_{i=1}^{16} [y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i)] - \lambda \sum_{j=1}^5 \beta_j^2, \lambda = 1$$

The fixed penalty limits coefficient divergence under sparse data and near-separation,³⁶ but five predictors with

six events (about 1.2 events per predictor) fall far below accepted sample-size guidance for model development,³⁷ and the penalty does not remedy this.

Internal behavior was evaluated by leave-one-dispute-out cross-validation: in each iteration 15 disputes were used for fitting and the omitted dispute received an out-of-fold probability. Performance measures were the area under the receiver operating characteristic curve (AUROC), average precision (AP), Brier score, accuracy, sensitivity, and specificity at a threshold of 0.50. A sensitivity analysis replaced manually coded predictors with dictionary-derived predictors. Coefficients and exponentiated coefficients were interpreted only as directions and relative contributions within this corpus, not as causal effects or transportable odds ratios. Because the predictors were chosen after case review and describe grounds that courts themselves rely on, they are close to components of the outcome, and the model is presented as a descriptive display, not as a predictor.

2.6 Dashboard functions and software tests

The local dashboard offered search by decision number and clinical domain; filters for disposition, judicial level, coding confidence, and each concept; concept prevalence recalculated for the active subset; a global coefficient-contribution view; and a de-identified dispute explorer showing the triggering text (Figure 1). The interface stated that model directions were not causal and that the system was not designed to determine malpractice.

Deterministic tests covered six families: corpus reconciliation and identifier uniqueness (4 tests), concept filters (8), judicial-level filters (4), case-level additivity of feature contributions to the model logit (16), probability bounds for all 16 disputes (16), and coefficient completeness (1), for 49 tests in total. Search, disposition and confidence filters, prevalence recalculation, and the case explorer were not covered by automated tests. These checks verify software behavior and internal arithmetic only; they do not assess clinical, legal, temporal, or external validity, and no usability evaluation with intended users was performed.

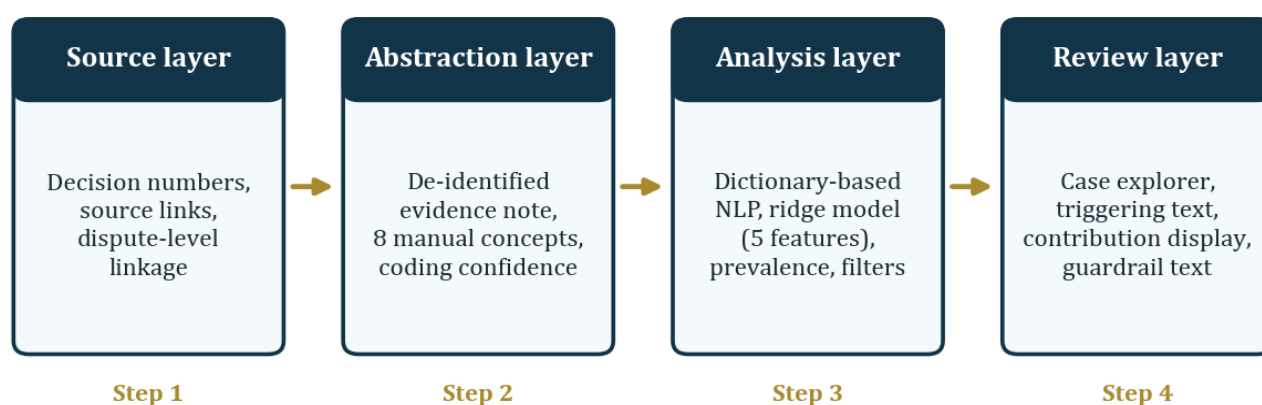


Figure 1. Provenance-to-review workflow of the dashboard: the source layer retains decision identifiers and links, the abstraction layer holds de-identified evidence notes and coded concepts, the analysis layer generates dictionary output, model contributions, filters, and prevalence, and the review layer returns the user to case-level evidence under visible guardrail text.

2.7 Ethics and data handling

Ethical approval was granted by the Enigma Institute Ethics Committee (No. 012/2026). The study analyzed public judicial documents and a de-identified research dataset without participant contact. Public availability of a

judgment was not treated as permission to redistribute unnecessary personal details.

Before submission, aggregate quantities were cross-checked: macro averages from the concept-level values, micro precision and recall from integer confusion counts, classification metrics against the six granted and ten rejected outcomes, exponentiated coefficients against the

reported multipliers, and the six test families against the total of 49.

3. Results

3.1 Corpus characteristics

The corpus comprised 16 independent disputes decided between 2006 and 2026: six (37.5%) with a granted

outcome and ten (62.5%) with a rejected outcome. Coding confidence was high for nine records, moderate for six, and low for one; no dispute was excluded on this basis. Because inclusion depended on access and detail, these proportions do not estimate the national frequency of malpractice findings or claimant success. Table 1 summarizes the corpus and analysis profile.

Table 1. Corpus, analysis, and software-test profile.

Characteristic	Value	Comment
Unit of analysis	16 independent disputes, 2006–2026	One record per dispute, not per document
Disposition	6 granted; 10 rejected	Not a national success rate
Coding confidence	9 high; 6 moderate; 1 low	No independent second annotator
Concept decisions	128 (16 disputes × 8 concepts)	Manual coding as reference
Exploratory model	5 predictors; ridge $\lambda = 1$	No causal interpretation
Internal evaluation	Leave-one-dispute-out cross-validation	No temporal or external validation
Software tests	49 of 49 passed	Software checks only

Notes: The 49 tests comprised 4 reconciliation/uniqueness, 8 concept-filter, 4 judicial-level-filter, 16 contribution-additivity, 16 probability-bound, and 1 coefficient-completeness test.

3.2 Dictionary concordance with manual coding

Macro precision was 95.7%, macro recall 83.7%, and macro-F1 87.9%. Pooled across 128 concept decisions, there were 33 true positives, 2 false positives, 13 false negatives, and 80 true negatives, giving micro-precision 94.3% (95% CI 80.8–99.3%), micro-recall 71.7% (95% CI 56.5–84.0%), and micro-F1 81.5%. Four concepts (specific breach plus causation, informed-consent issue, organizational failure, and formal/evidentiary barrier) reached F1 = 1.00. Documentation problem had precision 1.00, recall 0.75, and F1 0.86; adverse professional finding had precision and recall of 0.80 and F1 0.80; accepted

SOP/risk defense had precision 0.86, recall 0.75, and F1 0.80 (Table 2, Figure 2).

Severe injury had precision and specificity of 1.00 but recall of 0.40 (F1 0.57). Reconstructing the integer counts that are consistent with Table 2 and the pooled totals gives a single solution in which severe injury was positive in 15 of the 16 disputes: six were detected and nine were missed, and the specificity of 1.00 rests on a single negative record. The concept therefore varied little across the corpus and carried almost no discriminating information. The error profile was asymmetric: positive dictionary outputs were usually supported by manual coding, but a missing match could not be taken as evidence that a concept was absent.

Table 2. Dictionary-based NLP performance by concept against manual reference coding (N = 16 disputes per concept).

Concept	Manual positives	TP	FP	FN	TN	Precision	Recall	Specificity	F1
Specific breach + causation	n.r. ¹	n.r.	n.r.	n.r.	n.r.	1.00	1.00	1.00	1.00
Informed-consent issue	n.r. ¹	n.r.	n.r.	n.r.	n.r.	1.00	1.00	1.00	1.00
Organizational failure	n.r. ¹	n.r.	n.r.	n.r.	n.r.	1.00	1.00	1.00	1.00
Formal/evidentiary barrier	n.r. ¹	n.r.	n.r.	n.r.	n.r.	1.00	1.00	1.00	1.00
Four concepts combined	14	14	0	0	50	1.00	1.00	1.00	1.00
Documentation problem	4	3	0	1	12	1.00	0.75	1.00	0.86
Adverse professional finding	5	4	1	1	10	0.80	0.80	0.91	0.80
Accepted SOP/risk defense	8	6	1	2	7	0.86	0.75	0.88	0.80
Severe injury	15	6	0	9	1	1.00	0.40	1.00	0.57
Macro average²						0.957	0.837	0.974	0.879
Pooled (micro)	46	33	2	13	80	0.943	0.717	—	0.815

Notes: TP, true positives; FP, false positives; FN, false negatives; TN, true negatives; NLP, natural language processing; SOP, standard operating procedure; n.r., not reported. Counts for documentation, adverse professional finding, accepted SOP/risk defense, and severe injury were reconstructed from the reported ratios and pooled totals. ¹Individual counts for the four concepts with F1 = 1.00 were not recoverable; their combined counts are shown. ²Unweighted means of the concept-level values; recomputing from integer counts gives macro specificity 0.973.

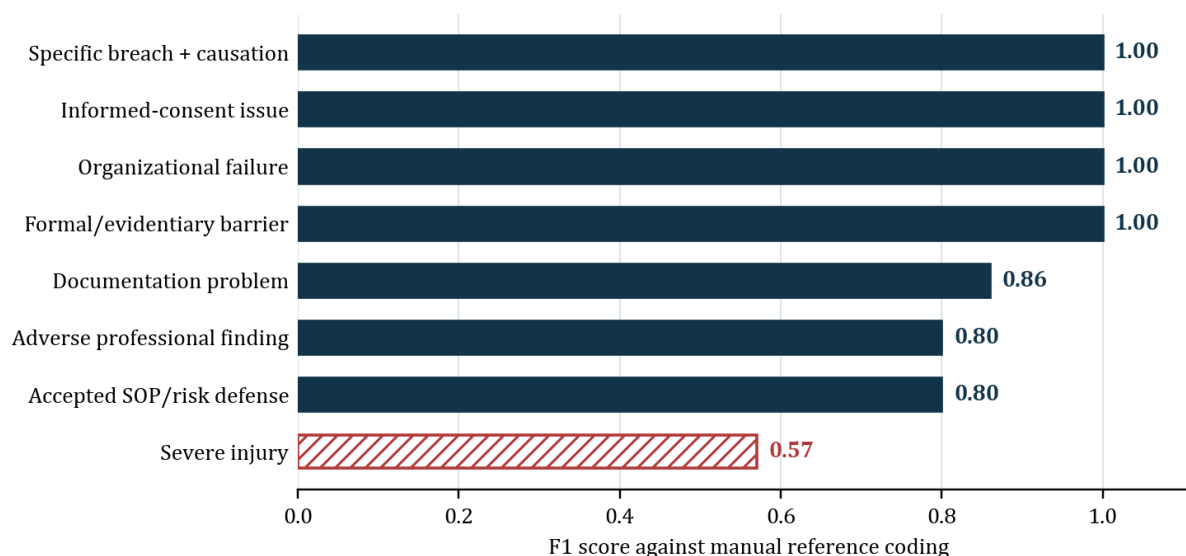


Figure 2. F1 of the dictionary-based NLP for the eight prespecified concepts against manual reference coding (128 binary concept decisions). The hatched bar marks severe injury, which had the lowest F1 (0.57) because the evidence-note field did not consistently repeat death or disability recorded elsewhere in the abstraction.

3.3 Exploratory multivariable pattern

Three ridge coefficients were positive and two negative (Table 3, Figure 3). The largest positive coefficient was for specific breach plus causation (1.30; multiplier 3.66) and the largest negative for accepted SOP/risk defense (−0.93; multiplier 0.39). Adverse professional finding (0.86),

organizational failure (0.61), and formal/evidentiary barrier (−0.45) completed the pattern. These values describe the fitted corpus and are not causal or jurisdiction-wide effects. No intervals could be reported for the coefficients because row-level data were not available to this analysis.

Table 3. Ridge logistic coefficients ($\lambda = 1$) fitted to the 16 disputes.

Predictor	Coefficient (β)	Odds multiplier, $\exp(\beta)$	Direction
Intercept (β_0)	not reported	—	—
Specific breach + causation	1.30	3.66	Positive (largest)
Adverse professional finding	0.86	2.36	Positive
Organizational failure	0.61	1.84	Positive
Accepted SOP/risk defense	−0.93	0.39	Negative (largest)
Formal/evidentiary barrier	−0.45	0.64	Negative

Notes: Multipliers were calculated from unrounded coefficients (for example, $\exp[1.30]$ gives 3.67 from the rounded value). Coefficients describe the fitted corpus only. SOP, standard operating procedure.

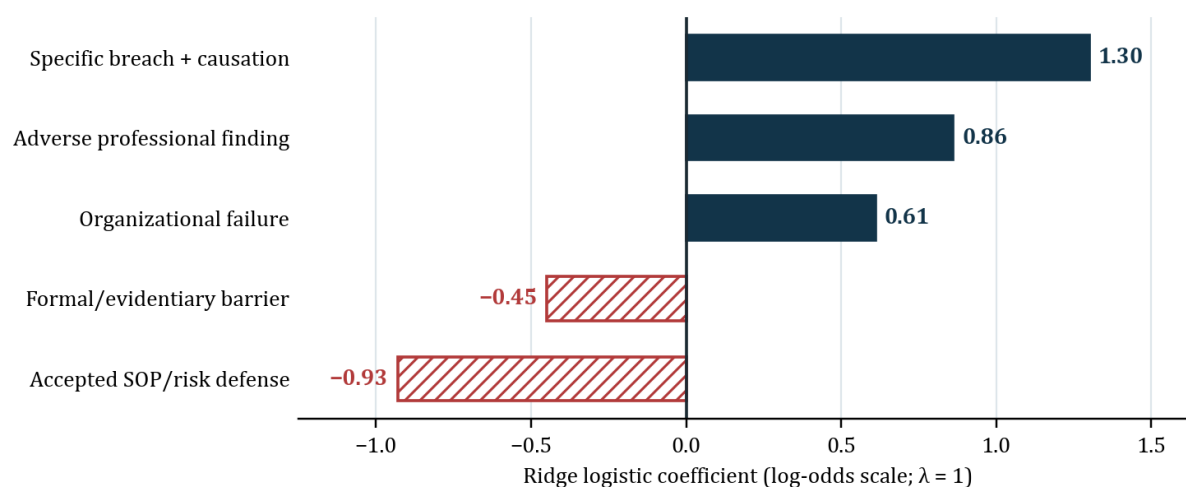


Figure 3. Direction and magnitude of the five ridge logistic coefficients (log-odds scale). Solid bars increase and hatched bars decrease the fitted corpus logit. Coefficients describe this pilot corpus and are not causal or transportable effect estimates; no intervals are shown because they could not be computed.

With manually coded features, the out-of-fold AUROC and AP were both 1.00, the Brier score was 0.050, and 15 of 16 disputes (93.8%; 95% CI 69.8–99.8%) were classified correctly at the 0.50 threshold. Sensitivity was 5/6 (83.3%; 95% CI 35.9–99.6%) and specificity 10/10 (100%; 95% CI 69.2–100%). The single misclassified dispute was one of the six granted outcomes. Replacing manual features with

dictionary-derived features gave the same AUROC, AP, accuracy, sensitivity, and specificity, with a Brier score of 0.061 (Table 4). With six events, a purposive corpus, predictors chosen after case review, and predictors that restate the grounds on which courts decide, the separation is expected and should not be read as a finding about model quality.

Table 4. Out-of-fold performance of the ridge model (leave-one-dispute-out cross-validation, threshold 0.50).

Feature set	AUROC	AP	Brier	Accuracy	Sensitivity	Specificity
Manual coding	1.00	1.00	0.050	15/16 (93.8%)	5/6 (83.3%)	10/10 (100%)
Dictionary-derived (sensitivity analysis)	1.00	1.00	0.061	15/16 (93.8%)	5/6 (83.3%)	10/10 (100%)

Notes: AUROC, area under the receiver operating characteristic curve; AP, average precision. Exact 95% confidence intervals: accuracy 69.8–99.8%, sensitivity 35.9–99.6%, specificity 69.2–100%. Intervals for AUROC, AP, and Brier score could not be computed without row-level data.

3.4 Software verification

All 49 deterministic tests passed: 4 reconciliation/uniqueness, 8 concept-filter, 4 judicial-level-filter, 16 contribution-additivity, 16 probability-bound, and 1 coefficient-completeness test. The additivity tests confirmed for every dispute that the intercept and displayed contributions reconstructed the fitted logit, and the probability-bound tests confirmed finite outputs between zero and one. Passing these tests shows that the implemented functions behaved as specified on this snapshot; it does not validate legal conclusions, the usefulness of the explanations, or generalization to new disputes.

4. Discussion

4.1 Principal findings

In this pilot, a provenance-centered dashboard organized 16 Indonesian malpractice disputes without collapsing clinical outcome, professional conduct, causation, organizational responsibility, and procedural proof into a single label. The dictionary reached high precision and lower recall, the ridge model displayed corpus-specific directions, and all deterministic software tests passed. These findings illustrate technical feasibility of an auditable research and educational workflow; they do not show that the dashboard is useful, valid, or generalizable. Keeping severe injury as a separate concept is consistent with claims research showing that defensibility of care and evidence of error relate more closely to claim resolution than severity.¹ Litigation studies across pelvic surgery,³⁸ pediatric emergency ultrasound,³⁹ procedural sedation,⁴⁰ intracranial hemorrhage,⁴¹ and Achilles tendon rupture⁴² likewise show that allegations and dispositions depend on clinical and procedural context.

4.2 Dictionary error structure

Only two false positives occurred across 128 decisions, but 13 manually positive concepts were missed. For corpus exploration these errors differ in consequence: a false positive sends a reviewer to an irrelevant case, whereas a false negative can remove a relevant case from a filtered subset. Dictionary-positive results can therefore prioritize inspection, but dictionary-negative records should not be excluded. The severe-injury result shows that the problem was not a missing synonym but a mismatch between the location of the information and the extraction input; the dictionary could have been applied to the structured attribute that held death or disability, and this explanation was identified after the evaluation. Because severe injury was positive in 15 of 16 disputes, the concept carries little information for filtering. The four concepts with F1 = 1.00

reflect phrases in a small set of notes that the rules were written around, not linguistic stability. Full-text, section-aware extraction and annotation by at least two annotators are needed before model comparisons become informative.^{29,43}

4.3 Model interpretation

The ridge model was designed to make contributions visible, not to establish a predictor. Its positive directions for breach with causation, adverse professional finding, and organizational failure and its negative directions for accepted-risk defense and evidentiary barrier are conceptually coherent, but coherence is not estimation precision: the coefficient of –0.45 for the evidentiary barrier is probably indistinguishable from zero in six events. The AUROC of 1.00 is a ranking result from 16 cross-validation fits, and the predictors encode the grounds on which courts decide; the circularity is an inherent limit of this design and not only a small-sample effect. The increase in Brier score from 0.050 to 0.061 when dictionary-derived features replaced manual features suggests that extraction error can alter probability magnitude while rank and classification summaries appear unchanged, but this rests on 16 observations.

4.4 Intended uses and governance

For education, the dashboard can help learners compare disputes with similar injuries but different breach or causation profiles and distinguish a statistical pattern from a legal conclusion. For research, concept filters, source triggers, and confidence filters can support abstraction quality assurance. For clinicians, the concept framework may inform documentation and consent practice and expert-witness preparation, without constituting legal advice. None of these uses was tested with users. Governance must stay visible because polished probabilities or coefficient charts can create automation bias; the World Health Organization guidance on artificial intelligence for health emphasizes transparency, accountability, safety, and continuing evaluation,⁴⁴ and a recent reader study found marked clinician-level variation in how explanations affected confidence and reliance.⁴⁵ Applied here, these principles support de-identification, role-based access, versioned rules and source manifests, visible uncertainty, and explicit prohibitions on provider scoring or active-case adjudication.

4.5 Strengths and limitations

Strengths are a dispute-level unit that avoids double counting appellate sequences, preserved provenance, prespecified concepts, a fixed penalty, out-of-fold values

for every dispute, and a sensitivity analysis linking extraction error to model output.

Limitations are substantial. First, the corpus is small and purposive; selection depended on accessibility and detail, which may correlate with outcome and court level, and the period crosses more than one statutory regime. Second, one abstractor produced both the notes and the labels, with no second annotator or agreement statistic. Third, the dictionary was written and scored on the same notes, and the outcome construct (granted or rejected) simplifies mixed findings. Fourth, the predictors restate the courts' reasoning, feature selection followed case review before cross-validation, and the model has about 1.2 events per predictor; risk-of-bias appraisal with PROBAST+AI⁴⁶ was not performed. Fifth, no confidence intervals could be given for AUROC, AP, Brier score, or coefficients, and the intercept is not reported. Sixth, no usability evaluation was performed, and the 49 software tests do not assess whether the explanations are understood or legally correct. Seventh, the dashboard was not compared with a spreadsheet, a baseline model, or an existing legal-analytics tool, and no Indonesian court-decision study was available for comparison. Finally, row-level data and code are not yet publicly deposited.

A next-stage study should define a sampling frame and outcome, apply a locked codebook with at least two blinded annotators and report agreement, freeze the dictionary and test it on a temporally separate set, evaluate the dashboard with forensic physicians, lawyers, and students, report descriptive cross-tabulations with intervals alongside or instead of predictive modeling, and deposit the corpus manifest, dictionary, code, and tests.

5. Conclusion

This pilot shows that a dispute-level, provenance-linked dashboard with a transparent dictionary, a descriptive ridge display, and deterministic software tests can be built for a small corpus of Indonesian medical malpractice disputes. The dictionary had high precision and lower recall, and the model's near-perfect internal separation reflects its design and sample, not predictive validity. The dashboard is suited to education, hypothesis generation, and abstraction quality assurance. Independent annotation, a defined outcome, a larger and temporally separate corpus, user evaluation, and public deposit of materials should precede any further claim.

Declarations

Use of generative artificial intelligence. None. The authors declare that no generative artificial intelligence (AI) tools were used in the preparation, drafting, analysis, or editing of this manuscript.

Author contributions. DK: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft. HS: Investigation, Data curation, Validation, Writing – review & editing. FD: Conceptualization, Supervision, Project administration, Writing – review & editing. FS: Resources, Validation, Writing – review & editing.

Funding. None. The authors received no external funding for this study.

Conflict of interest. The authors declare no conflict of interest.

Ethical approval. Ethical approval was granted by the Enigma Institute Ethics Committee Indonesia (No. 012/2026).

Consent to participate. Not applicable. The study used public judicial documents and de-identified records without direct participant contact.

Data availability. Source decisions are publicly available in the Supreme Court Decision Directory. The de-identified corpus

manifest (decision numbers and links), dictionary, analysis code, and test files are held by the authors.

Reporting guideline. TRIPOD+AI, used as a transparency framework.

Acknowledgements. None.

References

1. Studdert DM, Mello MM, Gawande AA, et al. Claims, errors, and compensation payments in medical malpractice litigation. *N Engl J Med.* 2006;354(19):2024-2033. doi:10.1056/NEJMsa054479
2. Welling M, Takala A. Patterns of malpractice claims and compensation after surgical procedures: a retrospective analysis of 8,901 claims from the Finnish patient insurance registry. *Patient Saf Surg.* 2023;17(1):3. doi:10.1186/s13037-023-00353-0
3. Klemann D, Mertens H, van Merode F. Trends and developments in medical liability claims in the Netherlands. *Healthcare.* 2022;10(10):1929. doi:10.3390/healthcare10101929
4. Lian OM, Randsborg PH, Jakobsen RB, et al. Prevalence of malpractice claims after arthroscopic shoulder surgery: analysis of 69,097 procedures from a national registry in Norway. *Patient Saf Surg.* 2023;17(1):25. doi:10.1186/s13037-023-00378-5
5. Li N, Wang Z. Medical malpractice compensation in China: a retrospective study of litigation cases from 2010 to 2015. *Risk Manag Healthc Policy.* 2026;19:1-12. doi:10.2147/RMHP.S565182
6. Humphrey KE, Sundberg M, Milliren CE, et al. Frequency and nature of communication and handoff failures in medical malpractice claims. *J Patient Saf.* 2022;18(2):130-137. doi:10.1097/PTS.0000000000000937
7. Sundel MH, Blackburn KW, Seyoum N, et al. Lessons in liability: examining medical malpractice suits against general surgeons in Maryland. *Am J Surg.* 2023;225(4):748-752. doi:10.1016/j.amjsurg.2022.11.008
8. Myers LC, Einbinder J, Camargo CA Jr, et al. Characteristics of medical malpractice claims involving emergency medicine physicians. *J Healthc Risk Manag.* 2021;41(1):9-15. doi:10.1002/jhrm.21450
9. Rosenkrantz AB, Siegal D, Skillings JA, et al. Oncologic errors in diagnostic radiology: a 10-year analysis based on medical malpractice claims. *J Am Coll Radiol.* 2021;18(9):1310-1316. doi:10.1016/j.jacr.2021.05.001
10. Morton JM, Khoury H, Brethauer SA, et al. First report from the American Society for Metabolic and Bariatric Surgery closed-claims registry: prevalence, causes, and lessons learned from bariatric surgery medical malpractice claims. *Surg Obes Relat Dis.* 2022;18(7):943-947. doi:10.1016/j.soard.2022.04.004
11. Quraishi NA, Shetaiwi A, D'Aquino D, et al. Malpractice litigation and spinal surgery in the National Health Service: a single tertiary-level centre perspective over 12 years. *Ann R Coll Surg Engl.* 2021;103(8):548-552. doi:10.1308/rcsann.2021.0102
12. Republic of Indonesia. Law No. 29 of 2004 on Medical Practice [Undang-Undang Nomor 29 Tahun 2004 tentang Praktik Kedokteran]. State Gazette of the Republic of Indonesia 2004 No. 116; Supplement No. 4431. Jakarta: Government of the Republic of Indonesia; 2004.
13. Republic of Indonesia. Law No. 36 of 2009 on Health [Undang-Undang Nomor 36 Tahun 2009 tentang Kesehatan]. State Gazette of the Republic of Indonesia 2009 No. 144; Supplement No. 5063. Jakarta: Government of the Republic of Indonesia; 2009.
14. Republic of Indonesia. Law No. 17 of 2023 on Health [Undang-Undang Nomor 17 Tahun 2023 tentang Kesehatan]. State Gazette of the Republic of Indonesia 2023 No. 105; Supplement No. 6887. Jakarta: Government of the Republic of Indonesia; 2023.
15. Mahkamah Agung Republik Indonesia. Direktori Putusan Mahkamah Agung [Supreme Court Decision Directory]. Available from: <https://putusan3.mahkamahagung.go.id> [accessed January 22, 2026].
16. Malik V, Sanjay R, Nigam SK, et al. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and

- Explanation. *Proc 59th Annu Meet Assoc Comput Linguist*. 2021;4046-4062. doi:10.18653/v1/2021.acl-long.313
17. Wu Y, Liu Y, Lu W, et al. Towards interactivity and interpretability: a rationale-based legal judgment prediction framework. *Proc 2022 Conf Empir Methods Nat Lang Process*. 2022;4787-4799. doi:10.18653/v1/2022.emnlp-main.316
 18. Feng Y, Li C, Ng V. Legal judgment prediction via event extraction with constraints. *Proc 60th Annu Meet Assoc Comput Linguist*. 2022;648-664. doi:10.18653/v1/2022.acl-long.48
 19. Valvoda J, Cotterell R. Towards explainability in legal outcome prediction models. *Proc 2024 Conf North Am Chapter Assoc Comput Linguist*. 2024;7269-7289. doi:10.18653/v1/2024.naacl-long.404
 20. Yarahuan JKW, Lo HY, Bass L, et al. Design, usability, and acceptability of a needs-based, automated dashboard to provide individualized patient-care data to pediatric residents. *Appl Clin Inform*. 2022;13(2):380-390. doi:10.1055/s-0042-1744388
 21. Tsangaris E, Edelen M, Means J, et al. User-centered design and agile development of a novel mobile health application and clinician dashboard to support the collection and reporting of patient-reported outcomes for breast cancer care. *BMJ Surg Interv Health Technol*. 2022;4(1):e000119. doi:10.1136/bmjst-2021-000119
 22. ter Beek E, Kos M, Streppel M, et al. Assessing the usability of a tumor dashboard for multidisciplinary care teams for first time users: first exploration of a comparative participatory cognitive walkthrough to show mismatches in cognitive models. *Stud Health Technol Inform*. 2021;286:99-106. doi:10.3233/SHTI210646
 23. Hekman DJ, Barton HJ, Maru AP, et al. Dashboarding to monitor machine-learning-based clinical decision support interventions. *Appl Clin Inform*. 2024;15(1):164-169. doi:10.1055/a-2219-5175
 24. Goel K, Sindhgatta R, Kalra S, et al. The effect of machine learning explanations on user trust for automated diagnosis of COVID-19. *Comput Biol Med*. 2022;146:105587. doi:10.1016/j.combiomed.2022.105587
 25. Alabi RO, Almangush A, Elmusrati M, et al. Measuring the usability and quality of explanations of a machine learning web-based tool for oral tongue cancer prognostication. *Int J Environ Res Public Health*. 2022;19(14):8366. doi:10.3390/ijerph19148366
 26. Schwartz JM, George M, Rossetti SC, et al. Factors influencing clinician trust in predictive clinical decision support systems for in-hospital deterioration: qualitative descriptive study. *JMIR Hum Factors*. 2022;9(2):e33960. doi:10.2196/33960
 27. Salwei ME, Hoonakker P, Carayon P, et al. Usability of a human factors-based clinical decision support in the emergency department: lessons learned for design and implementation. *Hum Factors*. 2024;66(3):647-657. doi:10.1177/00187208221078625
 28. Gombolay GY, Silva A, Schrum M, et al. Effects of explainable artificial intelligence in neurology decision support. *Ann Clin Transl Neurol*. 2024;11(5):1224-1235. doi:10.1002/acn3.52036
 29. Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: the muppets straight out of law school. *Findings Assoc Comput Linguist EMNLP*. 2020;2898-2904. doi:10.18653/v1/2020.findings-emnlp.261
 30. Staliunaite I, Valvoda J, Satoh K. Comparative study of explainability methods for legal outcome prediction. *Proc Nat Legal Lang Process Workshop (NLLP)*. 2024;243-258. doi:10.18653/v1/2024.nllp-1.20
 31. Vats S, Zope A, De S, et al. LLMs – the good, the bad or the indispensable? A use case on legal statute prediction and legal judgment prediction on Indian court cases. *Findings Assoc Comput Linguist EMNLP*. 2023;12451-12474. doi:10.18653/v1/2023.findings-emnlp.831
 32. Zhang Z, Hu X, Zhang J, et al. FEDLEGAL: the first real-world federated learning benchmark for legal NLP. *Proc 61st Annu Meet Assoc Comput Linguist*. 2023;3492-3507. doi:10.18653/v1/2023.acl-long.193
 33. Chalkidis I, Garneau N, Goanta C, et al. LeXFiles and LegalLAMA: facilitating English multinational legal language model development. *Proc 61st Annu Meet Assoc Comput Linguist*. 2023;15513-15535. doi:10.18653/v1/2023.acl-long.865
 34. Thalken R, Stiglitz E, Mimno D, et al. Modeling legal reasoning: LM annotation at the edge of human agreement. *Proc 2023 Conf Empir Methods Nat Lang Process*. 2023;9252-9265. doi:10.18653/v1/2023.emnlp-main.575
 35. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378
 36. Sinkovec H, Heinze G, Blagus R, et al. To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *BMC Med Res Methodol*. 2021;21(1):199. doi:10.1186/s12874-021-01374-y
 37. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. doi:10.1136/bmj.m441
 38. Yao B, Slopnick E, Sheyn D, et al. Medical malpractice litigation in non-mesh-related pelvic organ prolapse surgery: an analysis of 91 cases. *Female Pelvic Med Reconstr Surg*. 2021;27(4):255-259. doi:10.1097/SPV.0000000000000795
 39. Solomon L, Emma M, Gibbons LM, et al. Current risk landscape of point-of-care ultrasound in pediatric emergency medicine in medical malpractice litigation. *Am J Emerg Med*. 2022;58:16-21. doi:10.1016/j.ajem.2022.05.010
 40. Eltorai A. Procedural sedation by non-anesthesiologists: a review of malpractice litigation. *J Leg Med*. 2022;42(1-2):67-74. doi:10.1080/01947648.2023.2174768
 41. Pecorari IL, Flaquer I, Bergemann R, et al. Medical malpractice and intracranial hemorrhages in the United States: an analysis of 121 cases over 35 years. *Heliyon*. 2023;9(4):e14885. doi:10.1016/j.heliyon.2023.e14885
 42. Arif HA, Devkumar D, Sultan A, et al. Retrospective analysis of malpractice claims following Achilles tendon rupture. *Foot Ankle Orthop*. 2025;10(4):24730114251397974. doi:10.1177/24730114251397974
 43. Fei Z, Shen X, Zhu D, et al. LawBench: benchmarking legal knowledge of large language models. *Proc 2024 Conf Empir Methods Nat Lang Process*. 2024;7933-7962. doi:10.18653/v1/2024.emnlp-main.452
 44. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. ISBN 978-92-4-002920-0.
 45. Nicolson A, Bradburn E, Gal Y, et al. The human factor in explainable artificial intelligence: clinician variability in trust, reliance, and performance. *NPJ Digit Med*. 2025;8(1):658. doi:10.1038/s41746-025-02023-0
 46. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. doi:10.1136/bmj-2024-082505