Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cross-institutional calibration and representation stability of language- model risk estimates in digestive surgery: a retrospective two-database study

shi xiao
First Affiliated Hospital of Soochow University
han ye
First Affiliated Hospital of Soochow University
zhi qiaoming
First Affiliated Hospital of Soochow University
shao dong hao
First Affiliated Hospital of Soochow University
kuang yuting
Hongwens@126.com
First Affiliated Hospital of Soochow University

Research Article

Keywords: Large language models, Clinical prediction, Calibration, External validation, Digestive surgery, Representation stability, Public databases

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11262838/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Cross-institutional calibration and representation stability of language-model risk estimates in digestive surgery: a retrospective two-database study

Shi Xiao[1], Han Ye[1]*, Zhi Qiao Ming[1], Shao Dong Hao[1], Yuting Kuang[1]*

[1] Department of General Surgery, The First Affiliated Hospital of Soochow University, Suzhou 215006, Jiangsu, P.R. China

*Correspondence: Han Ye, hanyeor@163.com; Yuting Kuang, Hongwens@126.com.

Article type: Research article

Abstract

Background

Large language models can report numerical surgical risks, but discrimination, probability accuracy and stability under equivalent input representations may differ. We evaluated these properties across two institutions using locally deployed language models and common clinical predictors.

Methods

This retrospective study included 1,909 digestive resection operations in VitalDB and 1,791 scheduled-resection candidates in MOVER. The outcome was in-hospital death or discharge more than 14 days after anesthesia end. Qwen3-14B and Gemma3-12B, both quantized to Q4_K_M, generated probabilities from ten fields at anesthesia end. Logistic regression and gradient boosting used the same clinical fields. Source evaluation used patient-grouped folds; external evaluation transferred models and source-only logistic calibration maps without fitting to MOVER outcomes. Paired tests compared identical-text replay, reversed field order and equivalent hemoglobin/albumin unit conversion. Uncertainty was estimated with 1,000 patient-cluster bootstrap samples conditional on fitted models.

Results

The outcome occurred in 220/1,909 VitalDB operations (11.5%) and 140/1,791 MOVER candidates (7.8%). External areas under the receiver operating characteristic curve were 0.788 for logistic regression, 0.782 for gradient boosting, 0.656 for Qwen and 0.722 for Gemma. Raw Brier scores were 0.0760, 0.0723, 0.0699 and 0.0689, respectively. Source calibration increased external Brier scores by 0.01065 for Qwen (95% interval 0.00791–0.01320) and 0.00200 for Gemma (0.00015–0.00377). Among 429 MOVER operations with a convertible laboratory value, unit conversion changed raw risk by an average absolute 0.93 percentage points for Qwen (0.75–1.14) and 6.49 for Gemma (5.77–7.20). The corresponding changes after fixed source calibration were 1.20 and 2.66 percentage points. Some representation changes improved outcome error, whereas others worsened it.

Conclusions

In these two quantized models, lower aggregate probability error coexisted with weaker discrimination, unsuccessful calibration transfer and representation-dependent risks. External outcome evaluation and paired representation tests provide complementary evidence when assessing numerical language-model risk estimates.

Keywords: Large language models; Clinical prediction; Calibration; External validation; Digestive surgery; Representation stability; Public databases

Background

Numerical risk estimates can inform postoperative monitoring and communication only when their meaning is sufficiently stable across patients and settings. Large language models (LLMs) offer a convenient interface: a short clinical record can be converted directly into a probability without fitting a new task-specific model. Perioperative studies have evaluated this approach for postoperative outcomes and prognosis [1,2]. Its convenience raises a practical question for medical

informatics: what evidence is needed before a number generated from a clinical prompt can be treated as an individual risk estimate?

Discrimination and probability accuracy answer different questions. A model may rank patients poorly yet obtain a relatively low average squared probability error when most outcomes are absent. Conversely, useful ranking can coexist with probabilities that are too extreme. Calibration assesses agreement between predicted and observed risks and can change when a model moves between institutions [3]. For LLMs, an additional distinction concerns how a probability is obtained. Token probabilities, verbal confidence and directly generated numerical risks are different quantities. Prior clinical work has examined their relationship to diagnostic uncertainty and calibration [4].

Input representation adds another potential source of variation. Changing the order of clinical fields or converting hemoglobin from g/dL to g/L preserves clinical information when the units are stated correctly. A risk estimate that changes under these transformations introduces uncertainty unrelated to a change in the patient. Perioperative cross-national evaluations have already examined prompt format and calibration [2]. The remaining issue addressed here is whether medically equivalent representations affect directly generated numerical risks and their outcome error after a fixed source calibration map is transferred.

We studied digestive tract, hepatic and pancreatobiliary resection contexts in two independently collected surgical databases. The study combined three evaluations: comparison with conventional models using common clinical fields; transfer of source-only calibration; and paired representation tests with an identical-text replay control. The objective was to characterize the reliability of reported probabilities under a defined workflow, rather than optimize a new clinical prediction tool.

Methods

Study design and data sources

This was an exploratory retrospective analysis of VitalDB, a surgical database from South Korea [5], and the Medical Informatics Operating Room Vitals and Events Repository (MOVER), a US operating-room database [6]. VitalDB provided the source cohort for conventional model fitting and calibration; MOVER provided the external cohort. The analysis used structured clinical and administrative data. Waveforms were not required. Six selected MOVER source files were received and checked against the provider's version-2 MD5 listings.

The analysis was not prospectively preregistered. Cohort expansion and the 14-day endpoint threshold were informed by inspection of available data. Common-input inference rules were recorded before formal LLM responses were evaluated, but after cohort characteristics and conventional baseline results had been inspected. Some additional comparisons were introduced during the analysis; complete prospective timing cannot be claimed. Reporting locations and these deviations are described in Additional file 1, with reference to TRIPOD+AI principles [7].

Participants and outcome

VitalDB screening used an explicit operation-name mapping for adult non-transplant digestive tract, liver and pancreatobiliary resections. Simple cholecystectomy, donor or transplant procedures, isolated splenectomy, and drainage or bypass without a specified resection were excluded. Records required observed outcome information and the ordering operation start ≤ operation end ≤ anesthesia end ≤ discharge. Of 1,942 candidates, 1,929 met eligibility criteria; 20 pilot operations were excluded, leaving 1,909 operations. Repeated operations were retained and patient identity was used in validation and uncertainty estimation.

MOVER screening used a separate exact mapping of scheduled procedure names. Adults required a unique official operation-to-patient crosswalk, an inpatient encounter, observed discharge disposition and the ordering admission ≤ anesthesia start ≤ anesthesia stop ≤ discharge. Exact duplicates were removed and conflicting operation records were excluded. The discharge disposition "Expired" defined death; records with "Coroner" disposition were excluded pending clarification. The resulting 1,791 operations represent scheduled-resection candidates. No independent review established that every planned resection was performed. A sensitivity analysis required corroborating anesthesia events linked by patient identity and start/stop times within five minutes.

The binary endpoint was in-hospital death or a time from anesthesia end to hospital discharge exceeding 14 days. Discharge information therefore defined a completed admission outcome. This endpoint captures mortality and prolonged hospitalization; it does not measure adjudicated postoperative complications. No 30-day follow-up outcome was inferred. Prediction was positioned at anesthesia end because anesthesia duration was an input.

Common predictors and harmonization

Ten clinical fields were used: age, body mass index (BMI), American Society of Anesthesiologists (ASA) physical status, preoperative hemoglobin, albumin and creatinine, anesthesia duration, sex, anatomical domain and operative approach. Anatomical domains were upper gastrointestinal, intestinal, hepatic and pancreatobiliary. Sex was harmonized to M/F, and laparoscopic procedures were mapped to the shared Videoscopic approach category. Specific procedure names, identifiers and outcome labels were excluded from LLM prompts. Other clinical variables were not added selectively to one model family.

MOVER laboratory extraction linked patients across encounters using the corrected patient crosswalk. We selected the latest numeric result with an explicit supported unit collected strictly before anesthesia start and within 90 days. Conflicting values were treated as missing. The extraction

scanned 29,079,344 laboratory rows. Collection before anesthesia does not establish that the result had become available to clinicians at prediction time. Age was derived from birth date; height and weight were converted from the source feet/inches and ounces to metric units for BMI.

For the LLMs, missing values were represented as MISSING and explicitly described as unknown rather than normal. Numeric values were serialized to four decimal places. Conventional models retained source floating-point precision; the maximum audited serialization difference was 0.00004994 in the respective field's units. Thus, the models shared clinical fields but used different computational representations. Additional file 1 provides field definitions, units, mapping rules and the exact generic prompt.

Language models and conventional comparators

Qwen3-14B and Gemma3-12B-it were run locally using Q4_K_M quantization. Qwen's thinking mode was disabled. The runtime was llama.cpp build b11385 with CUDA 12.4 on an NVIDIA RTX 4070 SUPER with 12 GB memory. Temperature was 0, seed 20261004, top-p 1, top-k 0 and min-p 0; the context limit was 4,096 tokens and the output limit 192 tokens. A constrained JSON response contained a probability between 0 and 1 and up to three evidence-field names. These field names were not evaluated as faithful explanations of model behavior. No fine-tuning was performed. Model and software source links are provided in references [8–10].

The prompt specified the prediction time, endpoint, field meanings and units. Each model produced one baseline prediction for each of 3,700 operations across the two databases. Schema, range, coverage, duplicate-key and truncation checks were applied to saved responses. Gemma baseline inference was interrupted after 1,782 responses because of a timeout; the remaining cases were resumed with the same model and generation settings but a longer timeout. This runtime discontinuity was retained in the audit record.

Conventional comparators were logistic regression and histogram-based gradient boosting. Numeric preprocessing comprised median imputation, missingness indicators and standardization. Categorical preprocessing used modal imputation and one-hot encoding with unknown categories ignored. All preprocessing was fitted within the applicable training partition. Logistic regression used C = 1 and a maximum of 2,000 iterations. Gradient boosting used a learning rate of 0.05, 200 iterations, seven maximum leaf nodes, 30 minimum samples per leaf, L2 regularization of 1 and no early stopping. Parameters were fixed without a hyperparameter search. Implementations used scikit-learn [11].

Validation and calibration

Source discrimination and error for conventional models were estimated using five fixed patient-grouped folds. An individual patient's operations could not appear in both training and evaluation partitions. For calibrated conventional predictions, four inner patient-grouped folds within each outer training set generated predictions for fitting a logistic calibration map. The outer training model and its calibration map were then evaluated on the held-out outer fold.

LLM predictions were fixed outputs from pretrained models. Their source calibration maps were fitted to patients outside each held-out fold. For external evaluation, conventional models were fitted on the full VitalDB cohort. Their calibration maps used VitalDB out-of-fold probabilities, whereas the LLM maps used all source LLM probabilities. Every external map was fixed before application to MOVER; MOVER outcomes were not used for model fitting or recalibration. Moving a conventional calibration map from out-of-fold to full-source-fit predictions can introduce training-size-related differences.

Calibration used logistic regression of the outcome on logit-transformed probabilities, clipped to [0.000001, 0.999999], with C = 1,000,000 and 2,000 maximum iterations. Raw and calibrated stages were both retained. Pooled internal discrimination can change because different folds use different maps; a single increasing external map preserves ranking.

Paired representation experiments

Each model was evaluated under three additional conditions: identical-text replay, reversal of field order, and conversion of hemoglobin and albumin from g/dL to g/L by multiplying available values by ten and updating the unit text. Creatinine units were unchanged. These variants preserved the underlying clinical information and endpoint. Unit conversion was evaluated in the full cohort, in operations with at least one convertible laboratory value, and separately in operations where both values were missing. The last group tests changes in unit wording, not numerical conversion.

There were 11,100 additional responses per model, giving 29,600 saved baseline and variant responses overall. These are repeated observations on 3,700 operations, not independent patients. Comparisons included mean absolute probability change, excess absolute change over identical-text replay, and the paired Brier-score difference between variant and baseline. The proportion changing by at least ten percentage points was descriptive and was not a clinical harm threshold. The same source calibration maps were applied to all variants without refitting.

Statistical analysis

We calculated the area under the receiver operating characteristic curve (AUROC), average precision, Brier score and mean predicted probability. Brier-score difference relative to common-input logistic regression was the primary comparative error measure; the other metrics provided complementary assessments. Calibration intercepts, slopes and fixed-bin summaries were also examined. The Brier score is the mean squared difference between probability and binary outcome; smaller values indicate lower aggregate error within the same cohort. It is affected by event frequency and cannot alone establish superior individual risk prediction.

Paired uncertainty estimates used 1,000 patient-cluster bootstrap samples with seed 20261004. Resampled patients contributed all their operations, preserving paired predictions and within-patient clustering. Percentile 95% intervals were conditional on saved predictions and fitted calibration

maps; models were not refitted in each bootstrap sample. Comparisons were descriptive, with no multiplicity adjustment. Sensitivity analyses retained the first operation per patient, the first operation per admission, or anesthesia-corroborated MOVER cases. Anatomical and missingness strata were described without confirmatory subgroup claims.

Use of generative AI in research preparation

OpenAI Codex assisted with English manuscript drafting and editing. Its assistance was separate from the two locally evaluated prediction models. The author independently reviewed the code, results, references and the final text, and is responsible for the accuracy of the content.

Results

Cohort composition and missingness

The VitalDB cohort contained 1,909 operations in 1,898 patients, with 220 composite events (11.5%) and 11 in-hospital deaths. MOVER contained 1,791 candidates in 1,688 patients, with 140 composite events (7.8%), including 15 deaths and 130 prolonged stays; five operations met both components. Anatomical composition differed substantially: upper gastrointestinal operations accounted for 631/1,909 source operations but 159/1,791 external candidates, whereas intestinal operations accounted for 759/1,909 and 1,127/1,791, respectively (Table 1; Fig. 1).

Laboratory availability also differed. MOVER had observed hemoglobin in 421 operations, albumin in 312 and creatinine in 427; the corresponding source counts were 1,853, 1,849 and 1,849. All three laboratory values were missing in 1,334 MOVER operations, among which 63 events occurred. All three were observed in 303 operations, with 60 events. These groups therefore differed in event frequency as well as missingness. Median age was 63 years in VitalDB and 61 years in MOVER; median anesthesia duration was 249 and 326 minutes, respectively.

Baseline discrimination and probability error

Conventional models discriminated the outcome better than either LLM in both cohorts (Table 2; Fig. 2). In MOVER, raw AUROC was 0.788 for logistic regression, 0.782 for gradient boosting, 0.656 for Qwen and 0.722 for Gemma. Relative to logistic regression, the external AUROC differences were −0.132 for Qwen (95% interval −0.185 to −0.076) and −0.066 for Gemma (−0.110 to −0.022).

The Brier-score comparison gave a different ordering. External raw Brier scores were 0.0760 for logistic regression, 0.0723 for gradient boosting, 0.0699 for Qwen and 0.0689 for Gemma. Relative to logistic regression, the differences were −0.00613 for Qwen (−0.01175 to −0.00069) and −0.00708 for Gemma (−0.01243 to −0.00180). Relative to gradient boosting, the differences were smaller and their intervals included zero: −0.00240 (−0.00684 to 0.00227) and −0.00334 (−0.00707 to 0.00049). Mean raw external risks were 7.6%, 13.7%, 8.1% and 10.7%, respectively, against an observed event frequency of 7.8%.

Source calibration and external transfer

In VitalDB, source calibration reduced Qwen's Brier score from 0.1059 to 0.1000 and Gemma's from 0.1037 to 0.1007. Mean predicted risks moved from 3.9% and 9.0% to approximately 11.5%, matching the source event frequency at the aggregate level.

The fixed source maps increased error in MOVER. Qwen's Brier score rose to 0.0805, a change of 0.01065 (95% interval 0.00791–0.01320); Gemma's rose to 0.0710, a change of 0.00200 (0.00015–0.00377). Their external mean risks increased to 17.7% and 12.5%. Conventional source calibration reduced external Brier scores to 0.0733 for logistic regression and 0.0699 for gradient boosting, with changes of −0.00272 (−0.00353 to −0.00190) and −0.00242 (−0.00388 to −0.00091). These findings demonstrate model-specific calibration transfer in this dataset, rather than a general benefit of source recalibration.

Representation-dependent probabilities

All baseline and variant responses passed the final coverage and schema checks. Identical-text replay produced a mean absolute external change of 0.17 percentage points for Qwen and zero for Gemma. Reversing field order produced changes of 0.96 and 2.52 percentage points, respectively. Across all MOVER candidates, the unit variant produced mean absolute changes of 0.59 and 1.92 percentage points (Fig. 3).

At least one converted laboratory value was available in 429 external operations from 411 patients. Within this subset, Qwen's mean absolute raw-risk change was 0.93 percentage points (95% interval 0.75–1.14), with an excess over replay of 0.70 (0.51–0.91). Gemma's change was 6.49 percentage points (5.77–7.20). Absolute changes reached ten percentage points in 1.4% of Qwen predictions and 44.1% of Gemma predictions. Unit conversion increased Brier score by 0.00381 (0.00163–0.00661) for Qwen and 0.00747 (0.00017–0.01543) for Gemma (Table 3).

Fixed source calibration retained representation dependence. In the same external subset, mean absolute changes were 1.20 percentage points for Qwen (0.99–1.44) and 2.66 for Gemma (2.37–2.94), with Brier increases of 0.00245 (0.00022–0.00514) and 0.00320 (0.00051–0.00615), respectively (Fig. 4). Effects were not uniformly adverse: reversing Gemma's field order reduced its raw external Brier score by 0.00177. The complete paired results, including source-cohort and both-laboratories-missing analyses, are provided in Additional file 2.

Sensitivity analyses

Retaining the first MOVER operation per patient left 1,688 operations and 134 events. Raw AUROCs were 0.789 for logistic regression, 0.781 for gradient boosting, 0.650 for Qwen and 0.715 for Gemma. The anesthesia-corroborated subset contained 1,138 operations and 83 events; AUROCs were 0.803, 0.803, 0.686 and 0.750, respectively. The conventional models' higher discrimination therefore persisted in these restrictions. Anesthesia corroboration verifies event linkage and timing,

not completion of the scheduled resection. Admission-restricted and anatomical/missingness-stratified results are included in Additional file 2.

Discussion

Principal findings

This study identified three distinct limits to the reliability of numerical LLM risks in digestive surgery. Both local models discriminated the composite outcome less well than conventional models using the same clinical fields. Nevertheless, their raw external Brier scores were lower than logistic regression's, and the differences from gradient boosting were uncertain. Source calibration improved aggregate source error but increased external error for both LLMs. Finally, medically equivalent unit and field-order changes altered predicted risks, with effects that persisted after fixed source calibration.

The apparent disagreement between AUROC and Brier score is informative. AUROC evaluates ordering, whereas Brier score combines several properties of a probability forecast and depends on the event distribution. With a 7.8% external event rate, restrained probabilities can avoid large squared errors without separating many event from non-event cases. Matching the overall event frequency also does not establish calibration across the risk range. Reporting the metrics together prevents a favorable result on one axis from being mistaken for overall predictive superiority.

Relation to previous work

Earlier perioperative research established the feasibility of LLM-based risk prediction [1]. Lee and colleagues subsequently compared LLMs with conventional machine learning across INSPIRE and MOVER, examining prompt formats and calibration [2]. Their probability extraction and outcomes differed from the direct numeric output and composite endpoint used here. Gao and colleagues further showed the importance of distinguishing language-model confidence measures from clinical

probabilities [4]. Accordingly, the present contribution is the paired assessment of clinical-equivalence transformations, replay variation, fixed-map calibration transfer and observed outcome error within one common-input evaluation.

Recent work in this journal has evaluated LLM decision support across specialties and case complexity [12]. Our results add a narrower operational question: whether the same structured record produces a stable risk when its representation changes. This complements accuracy-centered evaluation because a model may give similar aggregate performance while individual outputs shift. The effect is especially relevant when data are supplied by different electronic record systems, which can encode laboratory units and field order differently.

Implications for risk communication and model evaluation

The external calibration result cautions against treating source correction as a portable property. In VitalDB, Qwen substantially underpredicted the aggregate event frequency. A map fitted to correct that source pattern increased its external mean risk to more than twice the observed MOVER event frequency. The simultaneous differences in case mix, laboratory availability and endpoint frequency provide plausible context, but this analysis cannot separate their contributions. Local validation should examine the transported probabilities themselves, even when source calibration improves.

Representation testing gives a second, complementary diagnostic. Gemma's unchanged identical-text replay and larger unit-related changes are consistent with sensitivity to representation under this runtime. Qwen exhibited small replay variation despite zero temperature, so its representation effects require comparison with that background variation. Neither result identifies an internal reasoning mechanism or establishes that quantization caused the behavior. The tests assess the stability of the complete model–prompt–runtime combination.

The largest highlighted unit effect occurred in the laboratory-observed external subset. Most MOVER candidates lacked both hemoglobin and albumin, and their unit variant changed wording alone.

Combining these groups diluted the numerical-conversion effect, whereas reporting only the observed subgroup would overstate its prevalence across the full cohort. Both denominators are needed. A change in risk magnitude also needs to be interpreted separately from a change in outcome error: field reversal improved Gemma's external Brier score, while unit conversion worsened it in the numerical subset.

For a clinical decision-support workflow, these findings favor standardized input handling, documented unit conversion and explicit checks of transported calibration and representation stability. They do not identify a treatment threshold or establish benefit from acting on a generated risk. Prospective workflow testing would need to connect prediction changes with actual decisions, clinician interpretation and patient outcomes.

Strengths and limitations

Strengths include common clinical fields across all models, an independently collected external dataset, patient-disjoint source evaluation, calibration restricted to source outcomes, paired replay and representation tests, and reconciliation of saved predictions with final aggregate metrics. The external sensitivity analyses address repeated patients and corroborating anesthesia records. Complete representation results are reported, including changes that improved error.

Several limitations constrain interpretation. First, MOVER eligibility relied on scheduled procedures, and completed resections were not clinically adjudicated. The composite outcome was dominated by prolonged hospitalization, which can reflect discharge practice and health-system factors; mortality counts were too small to support a separate mortality-prediction claim. The data-informed 14-day threshold and exploratory analysis history increase the risk of selective framing, despite retention of the full comparison set.

Second, the cohorts differed in anatomy, event frequency and missingness. MOVER's patient-linked laboratory extraction was temporally restricted, but the recorded collection time did not establish

result availability. Missingness can reflect collection and linkage coverage as well as clinical testing. These limitations affect the fidelity of the intended prediction-time record and the interpretation of transportability.

Third, only two quantized models, one prompt specification and one external institution were evaluated. Pretraining exposure to public data cannot be excluded. The Gemma interruption and Qwen replay variation limit attribution of all changes to representation alone. We did not assess larger models, unquantized versions, clinician comparisons or prospective clinical impact. Evidence-field labels were not treated as verified explanations.

Finally, the bootstrap intervals omit model-refitting uncertainty and were not adjusted for multiple comparisons. Operation-level performance can weight patients with repeated procedures more heavily, although patient-cluster resampling and restricted analyses partly address this issue. Classical calibration transfer also compares out-of-fold source predictions with full-source-fit external predictions. The results support a reliability evaluation in the studied setting, not a deployable risk calculator.

Conclusions

In two digestive-surgery datasets, Qwen3-14B and Gemma3-12B generated probabilities whose discrimination, aggregate error, calibration transfer and representation stability differed. A favorable external Brier score did not ensure stronger discrimination or stable risks under equivalent inputs. Evaluations of numerical LLM risk should report these properties together, with explicit prediction timing, missingness and patient-level validation boundaries.

List of abbreviations

ASA, American Society of Anesthesiologists; AUROC, area under the receiver operating characteristic curve; BMI, body mass index; JSON, JavaScript Object Notation; LLM, large language

model; MOVER, Medical Informatics Operating Room Vitals and Events Repository; SI, International System of Units; TRIPOD+AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus Artificial Intelligence.

Declarations

Ethics approval and consent to participate

This study used deidentified secondary data from VitalDB and access-authorized MOVER files. No participants were recruited or contacted for this analysis.

Consent for publication

Not applicable.

Availability of data and materials

VitalDB and MOVER are available from their respective providers [13,14]. Access to MOVER requires provider authorization and agreement to applicable terms. Patient-level MOVER records and prompts containing such records are not redistributed with this article. Additional file 2 contains aggregate results, and Additional file 3 contains analysis scripts and a reproduction guide. The analysis code is archived in Zenodo [15]. AUTHOR CONFIRM: approve code sharing and the applicable code license; confirm that the proposed supplementary files comply with both data-use agreements.

Competing interests

The authors declare no competing interests.

Funding

Suzhou Young Health Backbone Talent National Mentorship Programme, grant Qngg2024004. The funder had no role in study design, analysis, interpretation, writing, or the decision to submit.

Authors' contributions

S.X. contributed to conceptualization, methodology, formal analysis, and drafting. H.Y. ,S.D.H and Z.Q.M. contributed to investigation and data curation. Y.K. and H.Y. contributed to supervision, project administration, and funding acquisition.

Acknowledgements

We thank the investigators and participants of the originating public datasets and the developers of the open-source methods used in this analysis.

References

1. Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large Language Model Capabilities in Perioperative Risk Prediction and Prognostication. JAMA Surg. 2024;159:928-937. doi:10.1001/jamasurg.2024.1621.

2. Lee J, Kim H, Kim L, Lim L, Lee HC, Lee H. Comparison of large language models and conventional machine learning in postoperative outcome prediction: a retrospective, multi-national development and validation study. Korean J Anesthesiol. 2026;79:332-352. doi:10.4097/kja.25646.

3. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. doi:10.1186/s12916-019-1466-7.

4. Gao Y, Myers S, Chen S, Dligach D, Miller T, Bitterman DS, et al. Uncertainty estimation in diagnosis generation from large language models: next-word probability is not pre-test probability. JAMIA Open. 2025;8:ooae154. doi:10.1093/jamiaopen/ooae154.

5. Lee HC, Park Y, Yoon SB, Yang SM, Park D, Jung CW. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. Scientific Data. 2022;9:279. doi:10.1038/s41597-022-01411-5.

6. Samad M, Angel M, Rinehart J, Kanomata Y, Baldi P, Cannesson M. Medical Informatics Operating Room Vitals and Events Repository (MOVER): a public-access operating room database. JAMIA Open. 2023;6:ooad084. doi:10.1093/jamiaopen/ooad084.

7. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, 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.

8. Qwen Team. Qwen3-14B-GGUF model repository. https://modelscope.cn/models/Qwen/Qwen3-14B-GGUF. Accessed 4 Oct 2026.

9. Unsloth. Gemma-3-12B-it-GGUF model repository. https://modelscope.cn/models/unsloth/gemma-3-12b-it-GGUF. Accessed 4 Oct 2026.

10. ggml-org. llama.cpp, release b11385. https://github.com/ggml-org/llama.cpp/releases/tag/b11385. Accessed 4 Oct 2026.

11. Scikit-learn developers. Scikit-learn: machine learning in Python. https://scikit-learn.org/stable/. Accessed 4 Oct 2026.

12. Esmaeilzadeh P. Evaluating medical Large Language Models (LLMs) for clinical decision support: real-world performance across specialties and complexity levels. BMC Medical Informatics and Decision Making. 2026. doi:10.1186/s12911-026-03807-4.

13. VitalDB. VitalDB open dataset. https://vitaldb.net/dataset/. Accessed 4 Oct 2026.

14. MOVER. Medical Informatics Operating Room Vitals and Events Repository. https://mover.ics.uci.edu/. Accessed 4 Oct 2026.

15. Shi X, Han Y, Kuang Y, Zhi Q, Shao D. Cross-institutional calibration and representation stability of language-model risk estimates in digestive surgery: a retrospective two-database study. Zenodo; 2026. https://doi.org/10.5281/zenodo.23150321.

Tables

Table 1. Cohort characteristics and predictor availability

Characteristic	VitalDB	MOVER candidates
Operations	1,909	1,791
Patients	1,898	1,688
Composite events, n (%)	220 (11.5)	140 (7.8)
In-hospital deaths, n	11	15
Upper gastrointestinal, n	631	159
Intestinal, n	759	1,127
Hepatic, n	250	204
Pancreatobiliary, n	269	301
Age, years	63 (55–70)	61 (49–70)
BMI, kg/m²	22.8 (20.8–25.0)	25.7 (22.4–29.7)
ASA physical status	2 (1–2)	3 (3–3)
Anesthesia duration, min	249 (176–312)	326 (237.5–465.5)
BMI available, n (%)	1,909 (100)	1,406 (78.5)
ASA available, n (%)	1,883 (98.6)	1,791 (100)
Hemoglobin available, n (%)	1,853 (97.1)	421 (23.5)
Albumin available, n (%)	1,849 (96.9)	312 (17.4)
Creatinine available, n (%)	1,849 (96.9)	427 (23.8)

Continuous values are median (interquartile range) among observed records. Anatomical counts refer to operations. MOVER anatomy reflects scheduled procedures; BMI, body mass index; ASA, American Society of Anesthesiologists.

Table 2. Common-input model performance by calibration stage

Model	Stage	Source AUROC	Source Brier	External AUROC	External Brier	External mean risk, %
Logistic regression	Raw	0.759	0.0913	0.788	0.0760	7.6
Logistic regression	Calibrated	0.759	0.0910	0.788	0.0733	7.4
Gradient boosting	Raw	0.741	0.0956	0.782	0.0723	13.7
Gradient boosting	Calibrated	0.741	0.0936	0.782	0.0699	13.5
Qwen3-14B	Raw	0.620	0.1059	0.656	0.0699	8.1
Qwen3-14B	Calibrated	0.612	0.1000	0.656	0.0805	17.7
Gemma3-12B	Raw	0.608	0.1037	0.722	0.0689	10.7
Gemma3-12B	Calibrated	0.595	0.1007	0.722	0.0710	12.5

Source conventional models use patient-grouped out-of-fold evaluation; their calibration is nested. LLM predictions are fixed, with patient-grouped source calibration. External maps use source outcomes only. Values are point estimates; full precision, average precision and available uncertainty intervals are in Additional file 2. External event frequency was 7.8%.

Table 3. External numerical unit-conversion effects

Model	Stage	Mean absolute change, percentage points (95% interval)	Brier difference (95% interval)
Qwen3-14B	Raw	0.93 (0.75 to 1.14)	0.00381 (0.00163 to 0.00661)
Qwen3-14B	Calibrated	1.20 (0.99 to 1.44)	0.00245 (0.00022 to 0.00514)
Gemma3-12B	Raw	6.49 (5.77 to 7.20)	0.00747 (0.00017 to 0.01543)
Gemma3-12B	Calibrated	2.66 (2.37 to 2.94)	0.00320 (0.00050 to 0.00615)

The subset contains 429 operations from 411 patients with at least one observed hemoglobin or albumin value. Differences compare the equivalent-unit prompt with the baseline prompt. Positive Brier differences indicate increased error. Calibration maps were held fixed. Intervals use 1,000 paired patient-cluster bootstrap samples, conditional on fitted models and unadjusted for multiplicity.

Figure legends

Fig. 1. Cohort composition and predictor missingness

Panel A shows missing predictor proportions under the extraction rules used in this study. Panel B shows anatomical-domain proportions; labels give operation counts. VitalDB includes 1,909 operations and MOVER 1,791 scheduled-resection candidates. Laboratory extraction in MOVER used patient-linked measurements collected strictly before anesthesia and within 90 days. Missingness describes available extracted records and does not establish whether testing was performed clinically. MOVER domains derive from scheduled procedure names; actual resection was not independently adjudicated. Counts are operations, with repeated patients retained.

Fig. 2. Discrimination, probability error and mean risk before and after source calibration

Panels show AUROC, Brier score and mean predicted probability in VitalDB and MOVER. Circles denote raw predictions and open squares source-calibrated predictions. Source conventional predictions use patient-grouped outer validation and nested calibration; source LLM calibration holds out patient groups. External models and calibration maps use source data only. The outcome is in-hospital death or discharge more than 14 days after anesthesia end. These are point estimates; paired patient-cluster intervals are supplied in Additional file 2. Overlapping external AUROC markers reflect ranking preserved by a single monotonic map. Brier scores should be compared within cohorts because event frequencies differ.

Fig. 3. Raw-risk changes under replay and equivalent representations

Left panels show mean absolute risk change in percentage points; right panels show the Brier difference from the baseline prompt, where positive values indicate greater error. Points and horizontal lines represent estimates and percentile 95% intervals from 1,000 paired patient-cluster bootstrap samples. Full cohorts contain 1,909 VitalDB operations and 1,791 MOVER candidates.

Numerical unit-conversion subsets contain 1,857 and 429 operations, respectively. The both-laboratories-missing subsets contain 52 and 1,362 operations and test unit wording alone. Intervals condition on saved outputs and exclude training uncertainty and multiplicity adjustment. Identical-text replay characterizes runtime variation. Probability changes are not measurements of clinical harm.

Fig. 4. Representation effects after fixed source calibration

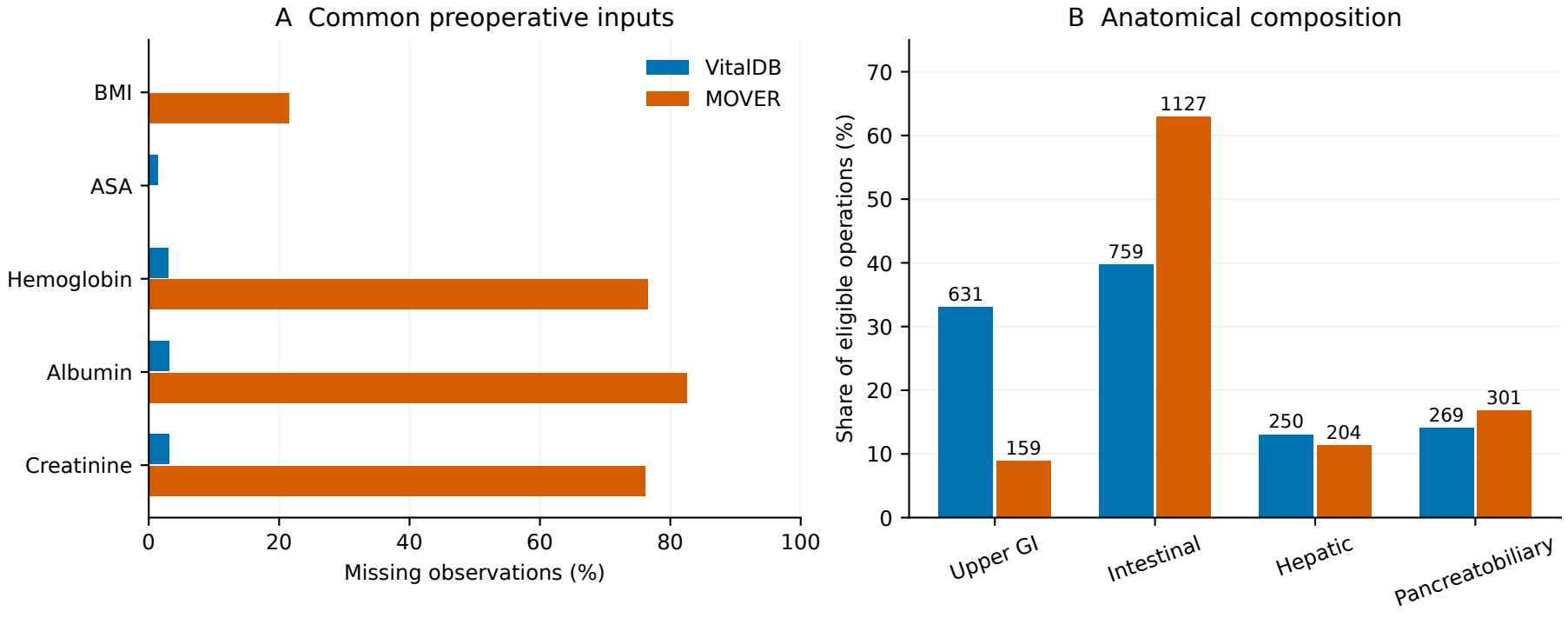
The same source-fitted logistic map is applied to baseline and variant probabilities; maps are not refitted for a variant or using external outcomes. Panel definitions, denominators and paired patient-cluster intervals follow Fig. 3. Mean absolute changes describe the transformed probability scale, whereas Brier differences measure changes in squared outcome error. Calibration can attenuate or amplify representation effects. Comparisons between laboratory-observed and missing subsets are descriptive because their case mix and event frequency differ.

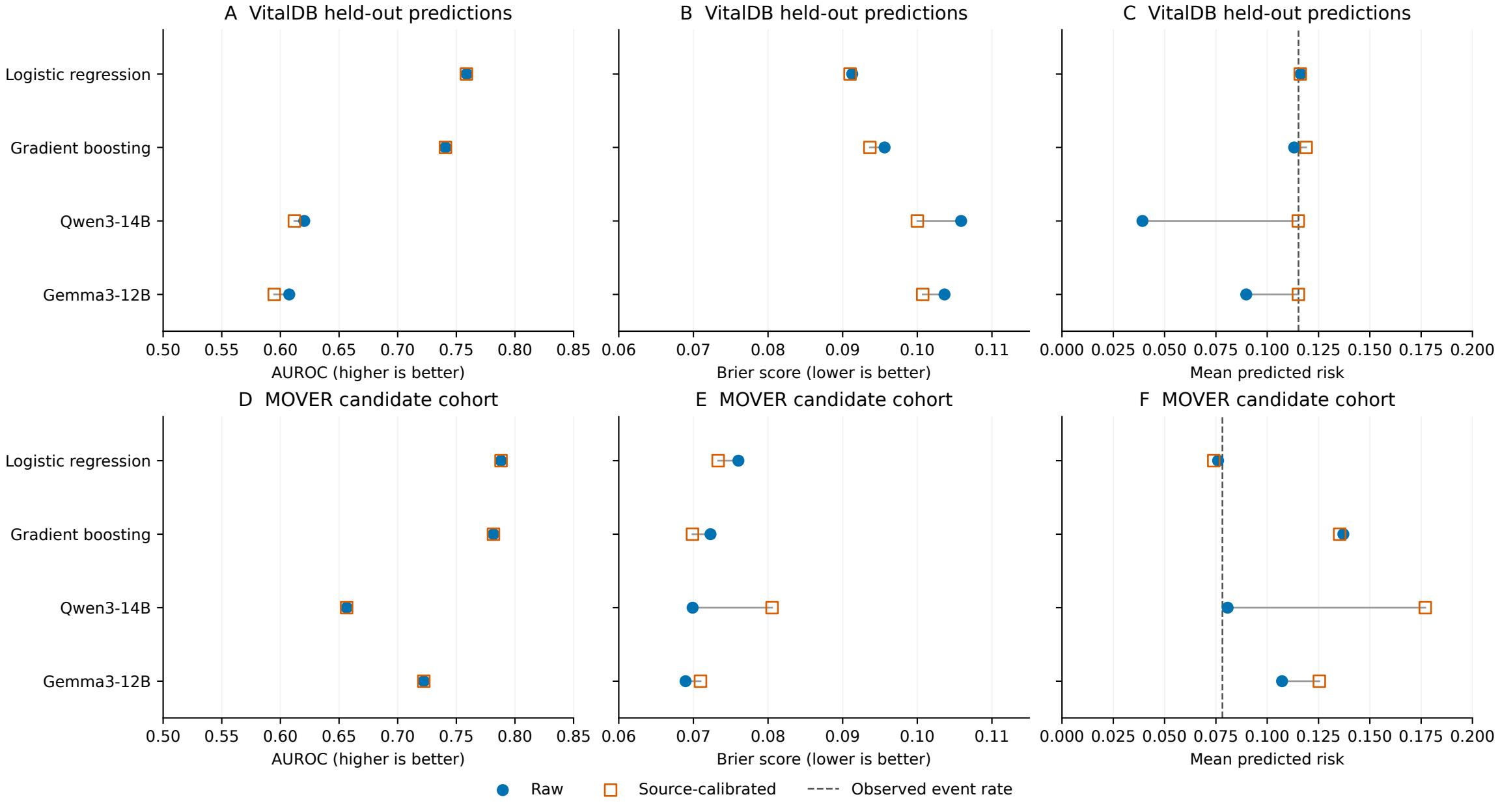
Additional files

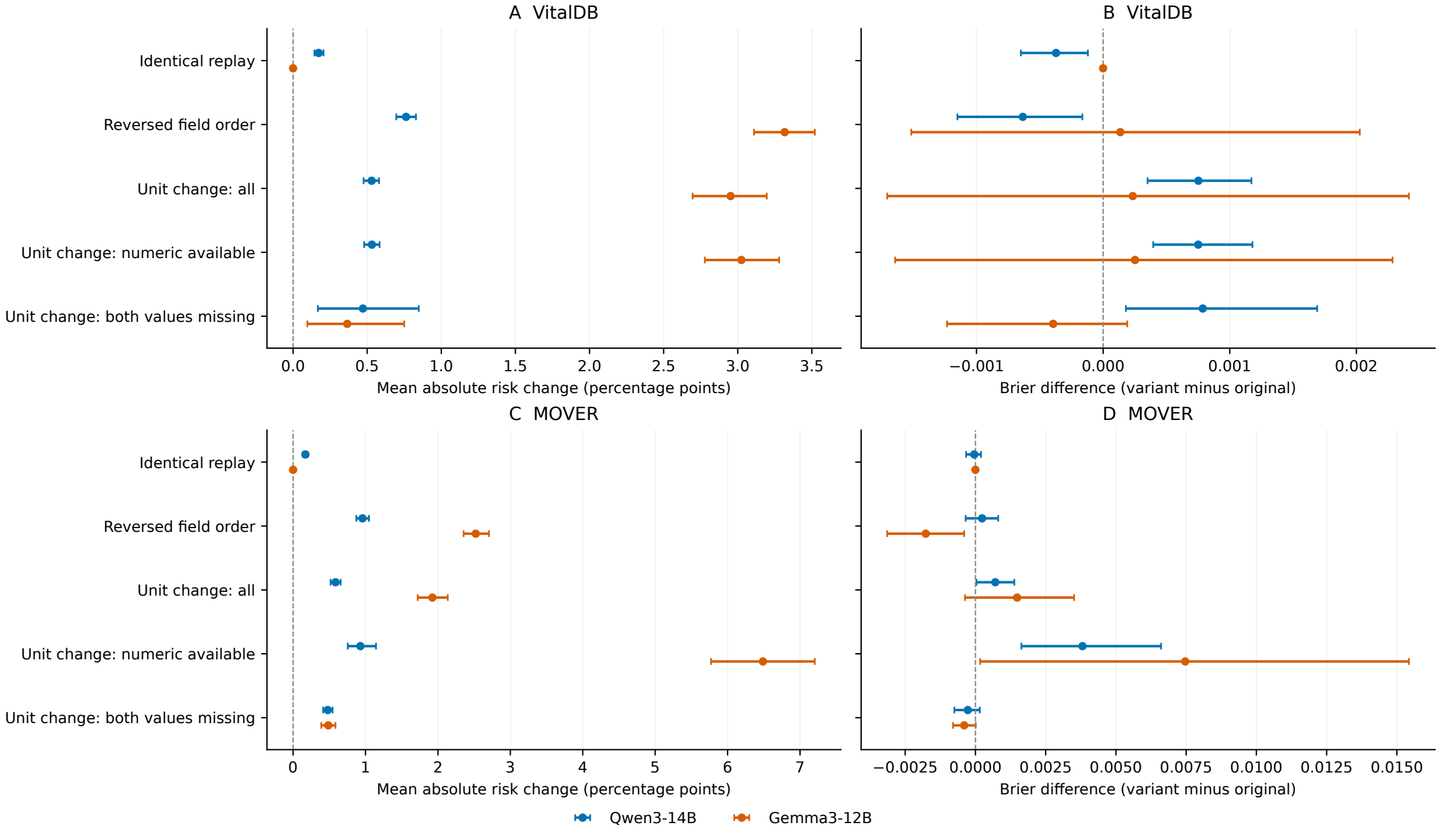
Additional file 1. Supplementary methods and reporting map (DOCX). Cohort definitions, variable dictionary, prompt and representation rules, calibration details, sensitivity interpretation and reporting locations.

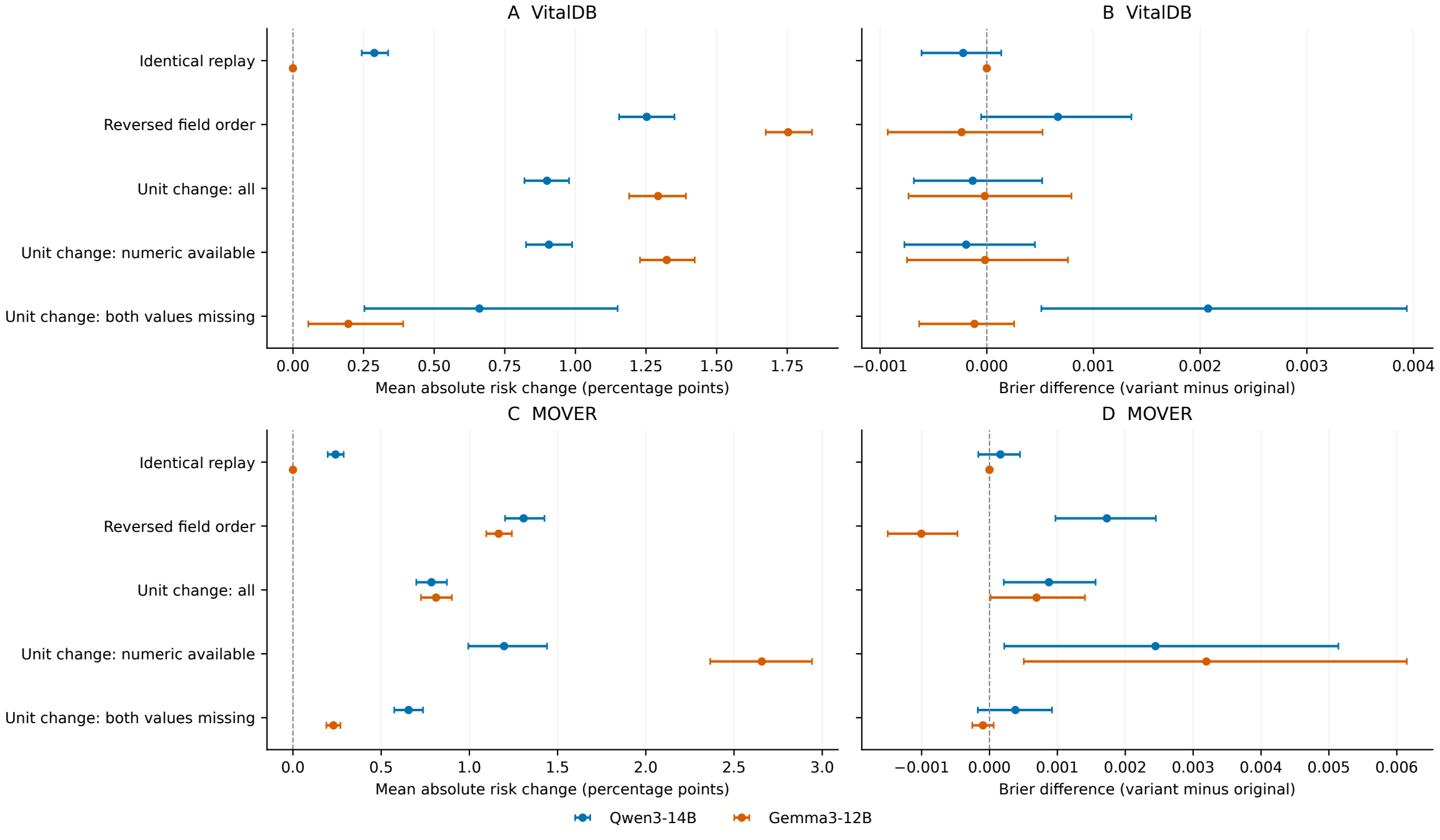
Additional file 2. Aggregate results (ZIP containing CSV files and README). Full baseline comparisons, paired intervals, calibration diagnostics, representation experiments and sensitivity summaries. No patient-level records or prompts are included.

Additional file 3. Analysis code and reproduction guide (ZIP). Audited analysis scripts, procedure dictionaries, generic prompt specification and environment notes. Authorized source data and locally generated case-level predictions are required for full rerunning and are not included.


A Common preoperative inputs
VitalDB
MOVER
BMI
ASA
Hemoglobin
Albumin
Creatinine
0
20
40
60
80
100
Missing observations (%)
B Anatomical composition
Share of eligible operations (%)
0
10
20
30
40
50
60
70
631
159
759
1127
250
204
269
301
Upper GI
Intestinal
Hepatic
Pancreatobiliary


A VitalDB held-out predictions
B VitalDB held-out predictions
C VitalDB held-out predictions
D MOVER candidate cohort
E MOVER candidate cohort
F MOVER candidate cohort
Logistic regression
Gradient boosting
Qwen3-14B
Gemma3-12B
0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85
AUROC (higher is better)
0.06 0.07 0.08 0.09 0.10 0.11
Brier score (lower is better)
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
Mean predicted risk
Raw
Source-calibrated
Observed event rate


A VitalDB
B VitalDB
C MOVER
D MOVER
Identical replay
Reversed field order
Unit change: all
Unit change: numeric available
Unit change: both values missing
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Mean absolute risk change (percentage points)
−0.001
0.000
0.001
0.002
Brier difference (variant minus original)
0
1
2
3
4
5
6
7
−0.0025
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
Qwen3-14B
Gemma3-12B


A VitalDB
B VitalDB
C MOVER
D MOVER
Identical replay
Reversed field order
Unit change: all
Unit change: numeric available
Unit change: both values missing
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
−0.001
0.000
0.001
0.002
0.003
0.004
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.005
0.006
Mean absolute risk change (percentage points)
Brier difference (variant minus original)
Qwen3-14B
Gemma3-12B

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.zip
- Additionalfile3.zip