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.

Longitudinal mock-examination trajectories for early warning of first-attempt failure in the National Medical Licensing Examination among Chinese residency trainees: a retrospective two-cohort study with temporal external validation

Weixian Xie

2026691021@gzhmu.edu.cn

Affiliated Qingyuan Hospital,Guangzhou Medical University,Qingyuan People's Hospital

Jiefang Zhou

Affiliated Qingyuan Hospital,Guangzhou Medical University,Qingyuan People's Hospital

Jiuyang Lin

Affiliated Qingyuan Hospital,Guangzhou Medical University,Qingyuan People's Hospital

Research Article

Keywords: National Medical Licensing Examination, residency training, mock examination, prediction model, machine learning, temporal external validation, medical education

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11404010/v1

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

Additional Declarations: No competing interests reported.

Title page

Title: Longitudinal mock-examination trajectories for early warning of first-attempt failure in the National Medical Licensing Examination among Chinese residency trainees: a retrospective two-cohort study with temporal external validation

Authors: Weixian Xie[1*], Jiefang Zhou[1], Jiuyang Lin[1]

Affiliations:

[1] Affiliated Qingyuan Hospital, Guangzhou Medical University, Qingyuan People's Hospital, Qingyuan 511518, Guangdong, China

Corresponding author: Weixian Xie, MD, Department of Anesthesiology, Affiliated Qingyuan Hospital, Guangzhou Medical University, Qingyuan People's Hospital, No. 35 Yinquan North Road, Qingcheng District, Qingyuan 511518, Guangdong, China. Tel: +86-18826610702. Email: 2026691021@gzhmu.edu.cn. ORCID: 0009-0005-5285-8716.

Abstract

Background. Passing the National Medical Licensing Examination (NMLE) is mandatory for Chinese residency trainees. Mock examinations generate repeated scores, but it is unclear whether their *trajectory* adds value beyond the most recent score. Rather than developing a deployable tool, we performed a methodological validation at one training base to test whether trajectory features carry information beyond the level of performance.

Methods. Retrospective two-cohort study. Cohort A comprised 118 trainees sitting 8 mock examinations before the 2024 NMLE; Cohort B comprised 115 sitting 7 before the 2025 NMLE. The primary quantitative outcome was the continuous first-attempt written score; first-attempt NMLE failure was analysed as secondary corroboration because events were few. Features were derived after within-cohort, within-batch z-standardisation. The prespecified primary model was logistic regression with level features (mean and last score) plus absence/leave/illness/zero-score indicators; random forest, XGBoost and LightGBM served as comparators only. Shape increment was tested with 5-fold cross-validated R^2 and reclassification; bidirectional temporal external validation assessed discrimination and calibration.

Results. After prespecified exclusions, 229 trainees were analysed (116 Cohort A, 113 Cohort B) with 14 first-attempt failures (6.1%); mean written score was 431.9 ± 38.7 (n = 226). Trajectory shape did not help: cross-validated R^2 was 0.3184 for mean-only, 0.3123 for mean + last, 0.2655 for last-only and 0.0594 for mean + last + shape. Only the prespecified logistic model transported stably in both directions (A→B AUC 0.696, 95% CI 0.519–0.874; B→A 0.608, 0.249–0.966); these intervals were wide because only 6–8 failure events occurred in each held-out cohort, and tree-based comparators were direction-dependent (0.196–0.230) with poor calibration. Net benefit exceeded both default strategies for thresholds of 3%–29%; at 7%, 44 trainees (19.2%) would be

flagged (sensitivity 35.7%, specificity 81.9%, positive predictive value 11.4%), the low PPV reflecting the 6.1% event rate. Conclusions were unchanged across 202 time-axis scenarios.

Conclusions. Trajectory shape added no incremental predictive value over the mean of early mock scores. This is a methodological validation of an educational hypothesis, not the development of a deployable tool: only two qualitative conclusions were robust (shape adds no value; tree-based models do not transport); every numerical estimate is an exploratory single-centre value that cannot be extrapolated. Findings should be interpreted with caution given the single-centre design and the low number of failure events; local recalibration is mandatory before any real-world deployment.

Keywords: National Medical Licensing Examination; residency training; mock examination; prediction model; machine learning; temporal external validation; medical education

Background

In China, passing the National Medical Licensing Examination (NMLE) is a statutory requirement for medical graduates to obtain independent practice rights and to continue in residency training [1]. For residency training bases, the first-attempt pass rate is also an important quality metric, and trainees who fail may face delayed graduation, prolonged training and career uncertainty. Regular mock examinations are widely used by training bases to monitor readiness and to allocate remedial teaching. These repeated measurements form a *longitudinal trajectory* that could, in principle, identify high-risk trainees before the actual examination.

Despite the intuitive appeal of trajectory-based early warning, empirical evidence is sparse. Existing studies from China have mostly been cross-sectional, predicting residency completion examination outcomes from a single measurement cycle with logistic regression [2–4]. Other related work has used large language models to generate examination questions [5] and mined text comments from an anaesthesiology rotation [6]. To our knowledge, no previous study has combined (i) longitudinal mock-examination trajectories, (ii) prediction of first-attempt NMLE failure, (iii) machine-learning algorithms with prespecified comparator models, and (iv) temporal external validation across consecutive cohorts.

A central scientific question is whether trajectory *shape* (trend, variability, curvature, late decline) adds predictive information above the most recent mock-examination score. Many clinicians and educators assume that the last score captures current readiness, but a trainee with a stable or improving trajectory may have a different risk from one with a declining trajectory despite the same final score. However, in small cohorts with low event rates, adding shape features may simply increase overfitting.

This study therefore had two aims, both of them methodological: (a) to test, with a prespecified hypothesis, whether trajectory shape improves prediction beyond the mean or the most recent mock-examination score; and (b) to examine, as a methodological demonstration, how well a prespecified model and its comparator algorithms transport across two consecutive cohorts. We state clearly what the study is not: it is not an attempt to deliver a deployable, generalisable prediction tool. The value of a prediction model depends on whether it transports to a new cohort,

and we treated transportability itself as the object of study rather than as a property to be assumed.

Methods

Study design and setting

We conducted a retrospective two-cohort study with bidirectional temporal external validation at Qingyuan People's Hospital, Guangzhou Medical University, a residency training base in Guangdong Province, China. Cohort A (derivation) consisted of the 2023 residency intake, who took monthly mock examinations during training and sat the 2024 NMLE. Cohort B (validation) consisted of the 2024 intake, who took mock examinations before the 2025 NMLE. The two cohorts were consecutive and non-overlapping.

Participants

Trainees enrolled in the same residency year at the base were included if they appeared on the official mock-examination roster (the "same-year principle"). Trainees from previous years who retook the examination, those who withdrew from training, and those without a recorded NMLE outcome were excluded from the primary analysis. Retakers were excluded because they were not on the same-year roster and would have inflated the event rate from 6.1% to 14.4%, violating the study target of predicting first-attempt failure within the current cohort. The continuous-outcome analysis further required a valid first-attempt written examination score.

Outcomes

We used a co-primary outcome structure. The primary binary outcome was first-attempt NMLE failure, defined as any of: (1) first-attempt written score < 360; (2) an official record of "not passed" on the first attempt; (3) failure of the clinical-skills component on two occasions and therefore not sitting the written examination; or (4) the presence of a re-examination record (passing the re-examination still counted as first-attempt failure). The co-primary continuous outcome was the first-attempt written examination score. We additionally report the secondary binary outcome "below the safety line" (score < 400) for context.

Variables and feature engineering

Each mock examination contained four unit scores and a total score; we used the total score for all analyses. Because the two cohorts used different mock-examination papers and showed different raw-score distributions (Cohort A raw batch means 328.6–371.8; Cohort B 364.6–383.6), scores were standardised *within cohort and within batch* to z-scores. This removes batch-to-batch difficulty differences and cohort-specific ceiling effects, allowing the two cohorts to be combined and compared.

From the first 3 (T3) and first 5 (T5) mock examinations we derived three sets of features:

- **Level features:** mean standardised score (`level_mean`) and last standardised score (`level_last`). - **Shape features:** linear slope (`slope`), within-person standard deviation (`variability`), drop in the last two scores (`drop_last2`) and quadratic coefficient (`quad`). -

Carry indicators: counts of absence, approved leave, illness and zero-score records during the window (`n_absent`, `n_leave`, `n_ill`, `n_zero`).

The time axis was anchored to *examination order* (sequential batch index, 1 = most recent) rather than calendar date, because exact examination dates could not be retrieved. We prespecified, and subsequently verified, that the main conclusions were invariant to monotonic rescaling and robust to a wide range of non-uniform interval assumptions (see Sensitivity analyses).

Statistical and machine-learning analysis

The primary analysis was performed at T3 (first three mock examinations), with T5 and sensitivity analyses reported for comparison. The **prespecified primary model** was logistic regression using feature set A (level features plus carry indicators). Random forest, XGBoost and LightGBM were fitted as comparators only; they were neither prespecified nor candidates for deployment, and their full internal comparison is relegated to the Supplementary Material (Tables S6 and S7) so that the main text focuses on the prespecified model and on the prespecified test of trajectory shape. The prespecification of logistic regression was made before analysis because the event-per-variable ratio was only approximately 1.2 and because threshold-based early warning requires reliable absolute risks rather than the best possible ranking: the ensemble algorithms showed far greater in-sample optimism, and in the T5 sensitivity combination LightGBM had a clearly worse Brier score (0.116–0.138) than logistic regression (0.078–0.083). No resampling or synthetic oversampling was used; the prespecified logistic model was fitted without class weighting, whereas the two gradient-boosting comparators received a positive-class weight equal to the ratio of non-events to events within each training set. No hyperparameter search was performed: all algorithms were fitted with hyperparameters fixed in advance, because with 14 events an inner tuning loop would have fitted noise rather than signal (fixed settings: logistic regression, none; random forest, 1000 trees with default mtry; XGBoost, learning rate 0.05, maximum depth 2, minimum child weight 1, subsampling and column-sampling rates 0.8, 200 boosting rounds; LightGBM, learning rate 0.05, three leaves, minimum one observation per leaf, 200 rounds).

We used 5-fold cross-validation for internal validation and report the area under the receiver-operating-characteristic curve (AUC), precision-recall AUC (PR-AUC), Brier score and sensitivity at 90% specificity. We tested whether adding shape features improved prediction by comparing nested feature sets using 5-fold cross-validated R^2 for the continuous outcome and by computing category-free net reclassification improvement (NRI) and integrated discrimination improvement (IDI) for the binary outcome. We also report LASSO and Boruta variable-selection results as methodologic context, but because of the low event count these were not used to override the prespecified feature set.

Temporal external validation

We performed bidirectional leave-one-cohort-out validation: Cohort A trained and Cohort B tested, then Cohort B trained and Cohort A tested. We report AUC, calibration slope, calibration intercept and Brier score. Apparent (training-cohort) performance is reported only to illustrate optimism.

Calibration and decision-curve analysis

Calibration of the primary model was assessed with the van Houwelingen–Hosmer test and visualised with calibration plots. Decision-curve analysis (DCA) was used to identify threshold probabilities at which using the model to trigger intervention yields a higher net benefit than intervening in all or in no trainees. Net benefit was calculated directly from counts as TP/n − FP/n × pt/(1 − pt), where pt is the threshold probability [7].

Sensitivity analyses

1. **Outcome definition:** treating non-participants in the NMLE as failures instead of excluding them.

2. **Time window:** T5 (first five mock examinations) in addition to T3.

3. **Time-axis robustness:** we generated 202 interval scenarios (uniform scaling, random non-uniform intervals between 0.5 and 2 months, and extreme front-loaded or back-loaded schedules) and repeated the trajectory-shape comparison under each scenario.

Reporting framework and interpretation of results

Reporting follows TRIPOD+AI. Because the number of outcome events is small, we deliberately divide the findings, throughout the results and discussion, into two explicitly labelled tiers (Table 9): Tier 1, qualitative conclusions that were stable across every analysis performed; and Tier 2, numerical estimates whose magnitude is uncertain and which must not be transported to other settings. Only Tier 1 statements are offered as conclusions of the study. We describe the work as a methodological and educational validation rather than as the delivery of a generalisable prediction tool, and we ask readers to judge the two tiers accordingly.

Software and reproducibility

All analyses were performed in R 4.6.1 using publicly available packages (lcmm, tidymodels, ranger, xgboost, lightgbm, pROC, shapviz, dcurves, tidyverse). Random seeds were fixed for every stochastic step, including the stochastic gradient-boosting implementations and the Boruta shadow-feature procedure. The full R pipeline, the data-export scripts and a numerical checklist for this manuscript were linked so that every reported number is traceable to a comma-separated output file. Scripts are available from the corresponding author on reasonable request, subject to data-protection restrictions.

Results

Cohort and outcomes

Figure S6 shows the participant flow. Of 233 rostered trainees, 229 were included in the primary analysis: 116 from Cohort A (6 first-attempt failures, 5.2%) and 113 from Cohort B (8 failures, 7.1%), with a combined failure rate of 6.1% (14/229). The continuous-outcome analysis included 226 trainees with a valid first-attempt written score (116 in Cohort A, 110 in Cohort B); the mean first-attempt written score was 431.9 ± 38.7 overall, 434.5 ± 39.9 in Cohort A and 429.0 ± 37.3 in

Cohort B. Among the 226 trainees with valid written scores, 11 (4.9%) scored below the passing line of 360 and 43 (19.0%) scored below the safety line of 400 (Table 1).

Baseline standardised trajectory features were similar between cohorts (Table 1). Cohort A showed greater variation in raw mock-examination difficulty across batches than Cohort B, confirming the need for per-cohort z-standardisation (Table 2).

Model discrimination and calibration

At T3, the prespecified logistic model had a 5-fold cross-validated AUC of 0.701 in Cohort A and 0.684 in Cohort B (Table 3). We report the comparator algorithms only briefly, because they contributed nothing usable and their full comparison is available in the Supplementary Material (Tables S6 and S7). Internally they performed similarly (0.661–0.705) and even had lower Brier scores (0.048–0.104 versus 0.060–0.065), but a low Brier score is not evidence of good calibration when predicted probabilities are shrunk towards the event rate, and the calibration plots showed that absolute risks were imperfect for every algorithm (Figure 2). The decisive contrast was optimism: apparent AUCs were 0.70–0.71 for the prespecified logistic model but 0.993–1.000 for the tree-based comparators, which is why the analysis of record throughout this paper is the prespecified logistic model. For transparency we give the fitted primary model in full: at T3 in the primary analysis set the predicted log-odds of first-attempt failure were -3.558 - 0.6657 *level_mean - 0.4484* level_last + 2.095 *n_leave + 0.603* I(cohort = B), where level_mean and level_last are cohort- and batch-standardised z-scores, n_leave is the number of approved-leave records in the window, and I(cohort = B) is 1 for Cohort B and 0 otherwise (Table S8).

Trajectory shape added no incremental value

The key hypothesis (H3) was that trajectory-shape features would improve prediction. They did not. For the continuous first-attempt written score, 5-fold cross-validated R^2 was 0.3184 for the mean-only model, 0.3123 for mean + last, 0.2655 for last-only, and 0.0594 for the full model including shape features (Table 4). Thus, the mean score outperformed the most recent score by $\Delta R^2 = 0.0529$, and adding shape features reduced R^2 by 0.259 relative to the mean-only model.

For the binary outcome, adding shape features to level features produced a negative ΔAUC (−0.010), an NRI of 0.274 and an IDI of 0.018 (Table 5). The NRI deserves explicit interpretation: it arose entirely from reclassification of non-events (NRI in non-events 0.274; NRI in events 0). Not one trainee who actually failed the NMLE was reclassified upwards by the shape features. The apparent "improvement" is therefore a property of the majority who did not fail and provides no case for using trajectory shape to identify the trainees who need help.

Temporal external validation

Bidirectional temporal external validation gave two findings (Table 6). First, the prespecified logistic model transported consistently in both directions (A→B AUC 0.696; B→A AUC 0.608; directional difference 0.089) and is therefore the analysis of record. Second, the tree-based comparators, documented here only to explain why they were not adopted, were strongly direction-dependent: A→B AUCs were 0.523–0.529 whereas B→A AUCs were 0.724–0.758 (directional differences 0.196–0.230), and calibration was poor throughout — only the random-

forest model trained on Cohort B and tested on Cohort A had a calibration slope close to 1.0 (1.02).

The confidence intervals around these AUCs must be read alongside the point estimates. Because only 8 failure events occurred in the Cohort B hold-out and only 6 in the Cohort A hold-out, the intervals were wide: 0.519–0.874 for A→B and 0.249–0.966 for B→A. The B→A interval in particular spans almost the entire possible range. The external-validation AUCs should therefore be read as indicating low-to-moderate discrimination with appreciable uncertainty, not as precise performance estimates for deployment.

Decision curve and actionable thresholds

Using the out-of-fold predictions from the prespecified logistic model, DCA showed that the model yielded a higher net benefit than both "intervene in all" and "intervene in none" for decision thresholds between 3% and 29% (Figure 3). At the 7% risk threshold, 44 trainees (19.2%) would be flagged for targeted support, with sensitivity 35.7%, specificity 81.9%, positive predictive value 11.4% and negative predictive value 95.1% (Table 7).

The low positive predictive value requires interpretation rather than apology. With a baseline failure rate of 6.1%, the proportion of true failures among flagged trainees is bounded close to the prevalence, so a PPV of 11.4% is largely a consequence of the low event rate rather than a failure of the model; the decision-curve net benefit, which explicitly incorporates prevalence, remained positive across the useful threshold window. Equally, the absolute net benefit was modest, which is the expected counterpart of moderate discrimination and few events. A flagged trainee should therefore be understood as someone for whom additional support is warranted, not as someone predicted to fail.

Sensitivity analyses

When non-participants were treated as failures, the combined event count increased to 17/232 (7.3%). This sensitivity definition systematically weakened discrimination: the logistic model's cross-validated AUC fell to 0.59–0.60, supporting our primary outcome definition that excludes non-participants. Results at T5 were directionally consistent with T3 but showed greater shape-feature overfitting.

The most consequential sensitivity analysis concerned the time axis, and its result is one of the study's main findings. Under 202 different interval assumptions (uniform scaling factors 0.5–5.0, random non-uniform intervals, and extreme front-loaded or back-loaded schedules), the shape-feature set never outperformed the mean-only benchmark: it won in 0 of 202 scenarios. The mean-only model was best in 198/202 T3 scenarios (98.0%) and 201/202 T5 scenarios (99.5%) (Table 8, Figure 5). Because exact examination dates could not be retrieved, this analysis establishes that the negative result for trajectory shape does not depend on any particular assumption about the spacing of mock examinations, and that calendar dates are not required for the conclusion.

The latent class analysis of score trajectories is reported for description only. The model with the lowest Bayesian information criterion among 30 candidate models had an entropy of 0.075 and a

smallest class containing two trainees; it was therefore not used as a predictor and carries no clinical predictive value (Figure 1).

Discussion

This study is a methodological validation of an educational hypothesis rather than the development of a deployable prediction tool. In this retrospective two-cohort study of 229 Chinese residency trainees we tested whether the shape of a mock-examination trajectory adds predictive value beyond the level of performance, and found that it does not. The warning signal available before the examination came from the **mean level** of early mock-examination performance, not from trajectory shape. The prespecified logistic model using only the mean and last standardised scores plus carry indicators was the only algorithm that transported stably across consecutive cohorts. Tree-based comparators had comparable or better internal performance but large direction-dependent optimism and poor calibration, and are not suitable for deployment without extensive recalibration; they are reported in detail only in the Supplementary Material.

Our negative finding for trajectory shape is methodologically important. Previous work in medical education has often assumed that "growth" or "decline" carries independent prognostic information [8]. In this sample, the simplest summary—the mean of the first three mock scores—was more robust than any combination of slope, curvature or variability. This is consistent with the low signal-to-noise ratio created by only 14 failure events: with an event-per-variable ratio near 1.2, any additional feature is more likely to fit noise than to capture true signal. We therefore explicitly refrained from using LASSO or Boruta to select features, because those methods are unstable under such low event-per-variable conditions. The reclassification results reinforce the same conclusion from a different direction: the positive NRI was generated entirely by non-events, with no failure reclassified by shape features.

Why the negative conclusion remains credible at this event count. A small number of events undermines the precision of a performance estimate, but it does not undermine a negative comparison in the same way. Our central claim is not that the model predicts well; it is that adding trajectory-shape features makes prediction no better, and on our data worse. That direction was reproduced in every analysis we performed: across four algorithms, at two time windows (T3 and T5), under two outcome definitions, and in all 202 time-axis scenarios, where the shape-feature set never once outperformed the mean-only benchmark. A false *positive* conclusion that "shape helps" would require the shape features to win in at least some of these settings; they won in none. The low event count therefore limits how precisely we can state an *effect size* — the AUCs, thresholds and odds ratios that we label Tier 2 — but it does not materially weaken the *qualitative* conclusion that trajectory shape adds no incremental value. This asymmetry, stable direction with unstable magnitude, is exactly why the two tiers are separated in Table 9. We state the limitation as an asymmetry rather than as a blanket caveat so that readers can judge which of our claims the small sample actually threatens.

The temporal external validation exposed a second important phenomenon: the two cohorts differed in the correlation between mock-examination batch scores and the final NMLE written score. Although both cohorts came from the same base and followed the same training

programme, Cohort B had a steady upward raw-score trend and a weaker batch-to-final correlation than Cohort A. This means that a model trained on one cohort is not automatically valid for the next, even when the same examination is being predicted. Any deployed system should therefore include a recalibration step (for example, Platt scaling or isotonic regression) at the start of each academic year, using the first one or two mock examinations of the new cohort.

Decision-curve analysis showed that the model could be useful in practice despite modest AUCs. At a 7% risk threshold, roughly one in five trainees would be flagged for extra support, with a sensitivity of 36% and a specificity of 82%. The threshold range over which the model outperforms both default strategies (3%–29%) is wide enough to allow local adaptation: a base with ample remedial resources might use a lower threshold, whereas a resource-constrained base might use a higher one. The absolute gain in net benefit was nevertheless modest, which is consistent with moderate discrimination and a low event rate and is the reason we present the operating characteristics as local reference values rather than as transferable performance claims.

What is robust and what is not

Because the number of failure events was small, we separate two tiers of conclusions (Table 9). Tier 1 conclusions held in every analysis we performed: trajectory shape added no incremental predictive value; the warning window is best expressed in examination counts rather than calendar time; tree-based algorithms are markedly less transportable across cohorts than logistic regression; and the primary model was prespecified rather than selected by performance. These conclusions were consistent across algorithms, time points (T3 and T5), outcome definitions and 202 time-axis scenarios.

Tier 2 conclusions are numerical estimates whose magnitude is uncertain and which should not be carried to other settings: the absolute external-validation AUC values, the coefficients and odds ratios of individual predictors, and the operating characteristics at any particular risk threshold. The B→A AUC confidence interval (0.249–0.966) illustrates the precision problem, and no model coefficient approached significance (level_mean odds ratio 0.514, 95% CI 0.100–2.725, $p = 0.432$; every other $p \geq 0.34$). These estimates characterise the direction of an association but not its size, and they must not be quoted as fixed model performance.

Contribution relative to previous work

Previous Chinese work on predicting residency assessment outcomes has largely used cross-sectional logistic regression within a single examination cycle [2–4]. The present study differs in three respects: predictors are longitudinal rather than a single measurement; the hypothesis about trajectory shape and the primary model were prespecified rather than chosen after seeing model performance; and performance was evaluated by bidirectional temporal external validation across consecutive cohorts, which is uncommon in this literature and revealed that apparent performance and transportability can diverge sharply. The main contribution of this work is therefore methodological and educational rather than the delivery of a deployable tool: it offers a prespecified, temporally validated workflow for prediction studies in medical education, and the negative but practice-relevant finding that trend features should not be over-interpreted.

Strengths and limitations

Strengths include the prespecified primary model and comparator feature sets, the use of co-primary continuous and binary outcomes, the bidirectional temporal external validation, the explicit testing of whether trajectory shape adds value, and a formal demonstration that the results do not depend on assumptions about examination spacing.

Limitations are substantial, and we state them without softening. The paper carries an inherent, not incidental, statistical-power defect that governs how its results should be read: with only 14 binary events, the failure outcome cannot support precise estimates of discrimination or of any risk threshold, and this is the single most important constraint on the study. First, the effective sample size for the binary outcome is very small. With 14 failure events and roughly twelve candidate parameters across the compared models, the event-per-variable ratio was approximately 1.2, an order of magnitude below the conventional minimum of 10. Multivariable estimates are therefore unstable even though the primary model was prespecified: no coefficient approached significance, and several confidence intervals span more than an order of magnitude (for example, the count of approved leave, odds ratio 8.13, 95% CI 0.10–607.57). Under the T5 sensitivity definition one indicator became completely separated from the outcome and its odds ratio was not estimable, and we therefore report no numerical odds ratio for it. Second, this is a single-centre study, and external validity to other training bases or regions is unknown; the model cannot be assumed to transfer, and its parameters must be re-estimated locally. Third, mock-examination structure differed slightly between the two years (8 versus 7 batches), although within-cohort and within-batch z-standardisation mitigates this. Fourth, we could not retrieve exact calendar dates, so we express the warning window as "after the first three (or five) mock examinations" rather than "X months before the NMLE"; the sensitivity analysis supports this framing but does not restore the calendar information. Fifth, we did not have information on trainee study hours, attendance, rotation performance or the remediation received after each mock examination, so residual confounding cannot be excluded. Sixth, we had no item-level or clinical-skills scores, so the two components of the NMLE could not be modelled separately. Seventh, the latent class trajectory solution was exploratory only (entropy 0.075, smallest class two trainees) and should not be read as an identified typology. Eighth, and most consequential for how the results should be read, the binary (first-attempt failure) outcome is statistically underpowered and we therefore treat it as a secondary corroboration rather than as the primary evidence. The primary quantitative evidence in this paper is the continuous outcome — the 5-fold cross-validated R^2 for the first-attempt written score — which does not depend on dichotomising the outcome, uses all 226 trainees with a valid score rather than 14 events, and carries far more statistical information. The negative result for trajectory shape is established on that continuous outcome; the binary analysis is reported to show that dichotomising the outcome does not reverse the conclusion, not to supply a performance estimate. Readers should weight the two accordingly, and any future study that requires a precise binary risk model will need several times this sample.

Implications

For residency training bases, the practical message is simple: use the average of the first three mock-examination scores, not the latest score alone, to identify trainees who need extra support. Do not over-interpret trend, variability or curvature unless the sample is much larger. The 7%

threshold reported here is a local reference value, not a universal standard: it was derived from this base's own failure rate and resource assumptions, and other bases should re-derive a threshold from their own prevalence and remedial capacity. Because the model's purpose is to allocate scarce teaching time, its output should only be used to trigger supportive help; using it for appraisal, ranking or any punitive purpose would be inappropriate and would risk harming the trainees who most need support. For researchers, the study illustrates the importance of temporal external validation in medical-education prediction models: a good internal AUC can hide severe overfitting and poor transportability, and a prespecified negative result, supported by a formal robustness analysis, can be more informative than a marginally better AUC.

Conclusions

This study is a methodological validation of an educational hypothesis rather than the development of a deployable prediction tool. Among Chinese residency trainees the mean of early mock-examination scores, not trajectory shape, carried the available warning signal for first-attempt NMLE failure, and only two qualitative conclusions were robust across every analysis: trajectory shape added no incremental predictive value, and tree-based algorithms did not transport across consecutive cohorts. Every numerical estimate — AUCs, risk thresholds and odds ratios — is an exploratory single-centre value that must not be extrapolated. With only 14 failure events at one centre and a binary outcome that is inherently underpowered, the quantitative evidence rests primarily on the continuous (first-attempt written score) outcome, with the binary analysis as secondary corroboration. These conclusions are stated as a comparison of predictive value rather than as a performance claim, which is why they survive the small sample even though the magnitudes do not. Local recalibration is mandatory before any real-world use, and any implementation must follow a supportive, non-punitive protocol.

List of abbreviations

AUC: area under the receiver-operating-characteristic curve

CI: confidence interval

CV: cross-validation

DCA: decision-curve analysis

EPV: events per variable

IDI: integrated discrimination improvement

LightGBM: light gradient-boosting machine

NMLE: National Medical Licensing Examination

NPV: negative predictive value

NRI: net reclassification improvement

PPV: positive predictive value

PR-AUC: area under the precision-recall curve

RF: random forest

SD: standard deviation

T3: the first three mock examinations

T5: the first five mock examinations

TRIPOD+AI: transparent reporting of a multivariable prediction model for individual prognosis or diagnosis, artificial intelligence extension

XGBoost: extreme gradient boosting

Declarations

Ethics approval and consent to participate

This study used de-identified retrospective administrative and examination records and involved no new intervention, no identifiable personal information and no additional contact with trainees. Because only de-identified routine educational and examination records were analysed, the study was submitted to the Medical Ethics Committee of Qingyuan People's Hospital, Guangzhou Medical University for a formal written determination that ethical approval is not required and that the requirement for written informed consent is waived, under the national guidance on ethics review for life-science and medical research involving humans (Guo Wei Ke Jiao Fa [2023] No. 4) [9]. The Medical Ethics Committee of Qingyuan People's Hospital, Guangzhou Medical University confirmed in writing on 15 October 2026 that, because this study analysed only de-identified retrospective educational and examination records and involved no intervention, ethical approval was not required, and that the requirement for written informed consent was waived. The committee issues determinations of this kind as a signed written confirmation rather than a numbered approval certificate, so no reference number is available. The study was conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Availability of data and materials

The data that support the findings of this study were derived from administrative records of Qingyuan People's Hospital and are not publicly available because of privacy restrictions. De-identified aggregated data and the analysis scripts are available from the corresponding author on reasonable request, subject to approval by the hospital.

Competing interests

The authors declare that they have no competing interests.

Funding
This work was supported by the Institutional Research Project Fund of Qingyuan People's Hospital, Guangzhou Medical University (institutional approval document No. 8 [2025]; grant GYQY-2025-8). The funder had no role in study design, data collection, analysis, interpretation or manuscript preparation.

Authors' contributions
WX conceived and designed the study, developed the statistical analysis plan, performed the analyses, interpreted the results and drafted the manuscript. JZ contributed to data acquisition, verification and adjudication of the examination outcomes and critically revised the manuscript. JL contributed to data curation, extraction and verification of the training records and critically revised the manuscript. All authors read and approved the final manuscript and agree to be accountable for all aspects of the work.

Use of AI tools
Generative artificial-intelligence tools were used solely for language editing and to assist with drafting and debugging the analysis code. All study design decisions, statistical analyses, interpretation of results and the final content of the manuscript were made and verified by the authors, who take full responsibility for the work. No generative AI tool is listed as an author.

Acknowledgements
We thank the Medical Education Office of Qingyuan People's Hospital for providing de-identified mock-examination and NMLE records and for adjudicating outcome queries.

References
1. Standing Committee of the National People's Congress. Law of the People's Republic of China on Physicians. Adopted 20 August 2021; effective 1 March 2022. Available from: http://www.npc.gov.cn/. Accessed 14 October 2026.

2. Wang XB, Zhou A, Yan XL, et al. Analysis of factors influencing the results of the standardized residency training completion theoretical examination [in Chinese]. Zhonghua Yixue Jiaoyu Tansuo Zazhi. 2024;23(6):835–840.

3. He GL, Chen SY, He SZ, et al. Ordinal logistic regression analysis of factors affecting the theoretical scores of standardized residency training completion assessment [in Chinese]. Zhongguo Biye Hou Yixue Jiaoyu. 2024;8(8):587–591.

4. Chen L, Yuan LX, Wang Q, et al. Research on core factors in process assessment of standardized residency training [in Chinese]. Zhongguo Biye Hou Yixue Jiaoyu. 2023;7(4):330–334.

5. Li R, Wu T. Application of artificial intelligence generated content in medical examinations. Adv Med Educ Pract. 2025;16:492895. doi:10.2147/AMEP.S492895.

6. Yao LF, Chen YJ, Shen J, et al. Development and external validation of a machine learning model for automated feedback quality assessment in Chinese anesthesiology residency training. Adv Med Educ Pract. 2025;16:599543. doi:10.2147/AMEP.S599543.

7. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18.

8. Payne VL, Lingard L, van Melle E, et al. Using growth curve analyses to examine the trajectory of resident performance on in-training examinations. Acad Med. 2015;90(11):1534–1540.

9. National Health Commission, Ministry of Education, Ministry of Science and Technology, National Administration of Traditional Chinese Medicine of the People's Republic of China. Measures for the ethical review of life science and medical research involving humans (Guo Wei Ke Jiao Fa [2023] No. 4). Issued 18 February 2023.

Tables

Table 1

Table 1 Baseline standardised trajectory features and first-attempt outcome, by cohort.

Variable	Overall (n=229)	Cohort A (n=116)	Cohort B (n=113)
First-attempt NMLE failure - n (%)	14 (6.1)	6 (5.2)	8 (7.1)
First-attempt written score - mean ± SD	431.9 ± 38.7	434.5 ± 39.9	429.0 ± 37.3
Below safety line (<400) - n (%)	43 (19.0)	21 (18.1)	22 (20.0)
Standardised mean score (level_mean)	0.0 ± 0.8	-0.0 ± 0.9	0.0 ± 0.8
Standardised last score (level_last)	0.0 ± 1.0	0.0 ± 1.0	0.0 ± 0.9
Slope (standardised)	0.0 ± 0.4	0.0 ± 0.4	0.0 ± 0.5
Variability (standardised)	0.4 ± 0.3	0.4 ± 0.3	0.4 ± 0.3
Drop in last two scores (standardised)	0.2 ± 0.4	0.1 ± 0.3	0.3 ± 0.4
Count of absence	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Count of approved leave	0.0 ± 0.1	0.0 ± 0.1	0.0 ± 0.0
Count of illness	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Count of zero-score records	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

Table 2

Table 2 Cohort composition, outcomes and mock-examination difficulty.

Item	Cohort A	Cohort B	Overall
On roster - n	118	115	233
Primary analysis - n (%	116 (98.3)	113 (98.3)	229 (98.3)

of roster)			
First-attempt failures - n (%)	6 (5.2)	8 (7.1)	14 (6.1)
Sensitivity analysis (non-participants as failures) - n/N	8/118	9/114	17/232
Continuous-outcome analysis (valid written score) - n	116	110	226
Mean first-attempt written score	434.5 ± 39.9	429.0 ± 37.3	431.9 ± 38.7
Mock examinations available per trainee	8	7	-
Raw mock-examination batch mean range	328.6–371.8	364.6–383.6	-

Table 3

Table 3 Internal (5-fold cross-validated) performance at T3. The prespecified logistic model is the analysis of record; the comparator tree-based algorithms are shown for completeness only, with fuller detail in Supplementary Tables S6 and S7.

Feature set	Algorithm	AUC	PR-AUC	Brier	Sensitivity at 90% specificity
Cohort A (derivation)	Logistic (prespecified primary)	0.701	0.158	0.060	0.286
Cohort A	Random forest	0.700	0.297	0.048	0.429
Cohort A	XGBoost	0.705	0.303	0.082	0.429
Cohort A	LightGBM	0.703	0.223	0.104	0.429
Cohort B (validation)	Logistic (prespecified primary)	0.684	0.157	0.065	0.286
Cohort B	Random forest	0.661	0.306	0.048	0.429
Cohort B	XGBoost	0.702	0.265	0.072	0.429
Cohort B	LightGBM	0.684	0.163	0.090	0.357

Table 4

Table 4 Incremental value of trajectory shape for the continuous outcome (5-fold cross-validated R^2).

Feature set	5-fold CV R^2	ΔR^2 vs mean-only	ΔR^2 vs last-only
Mean only	0.3184	0	0.0529
Mean + last	0.3123	-0.0061	0.0468
Last only	0.2655	-0.0529	0
Mean + last + shape	0.0594	-0.2590	-0.2061

Table 5

Table 5 Incremental value of trajectory shape for the binary outcome (ΔAUC, NRI, IDI).

Comparison	AUC	ΔAUC	NRI	NRI events	NRI non-events	IDI
Level features only	0.683	-	-	-	-	-
Level + shape features	0.672	-0.010	0.274	0	0.274	0.018

Table 6

Table 6 Bidirectional temporal external validation (A→B and B→A). Confidence intervals are wide because only 6–8 failure events occurred in each held-out cohort.

Direction	Algorithm	N	Events	AUC (95% CI)	Brier	Calibration slope
A → B	Logistic	113	8	0.696 (0.519–0.874)	0.078	0.239
A → B	Random forest	113	8	0.523 (0.259–0.788)	0.066	0.165
A → B	XGBoost	113	8	0.529 (0.255–0.802)	0.085	0.136
A → B	LightGBM	113	8	0.528 (0.246–0.810)	0.074	0.109
B → A	Logistic	116	6	0.608 (0.249–0.966)	0.051	-0.056
B → A	Random forest	116	6	0.745 (0.479–1.000)	0.041	1.021
B → A	XGBoost	116	6	0.724 (0.438–1.000)	0.068	0.366
B → A	LightGBM	116	6	0.758 (0.459–1.000)	0.061	0.384

Table 7

Table 7 Decision-curve operating characteristics of the prespecified logistic model. Operating characteristics at any threshold are local reference values derived from this base's 6.1% failure rate and should be re-derived locally.

Threshold (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Flagged - n (%)
6.1	50.0	78.1	13.0	96.0	54 (23.6)
7.0	35.7	81.9	11.4	95.1	44 (19.2)
10.0	28.6	89.8	15.4	95.1	26 (11.4)
15.0	21.4	95.3	23.1	94.9	13 (5.7)
20.0	21.4	95.8	25.0	94.9	12 (5.2)
30.0	21.4	96.7	30.0	95.0	10 (4.4)

Table 8

Table 8 Robustness of the trajectory-shape comparison to the time-axis definition. Under all 202 interval scenarios the shape-feature set never outperformed the mean-only benchmark.

Scenario set	Time point	Scenarios (n)	Mean-only R^2	Shape R^2 (median)	Shape beats mean (n)	Mean-only best (%)	Conclusion
Uniform scaling	T3	4	0.3184	0.0594	0	100	Identical across all scale factors
Random non-uniform intervals	T3	200	0.3184	0.0693	0	98.0	Shape never superior
Extreme front/back loaded	T3	2	0.3184	-3.240 / -0.062	0	-	Shape substantially worse under both extremes
Uniform scaling	T5	4	0.2966	0.2243	0	100	Identical across all scale factors
Random non-uniform intervals	T5	200	0.2966	0.2242	0	99.5	Shape never superior
Extreme front/back loaded	T5	2	0.2966	0.241 / 0.222	0	-	Still below mean-only under both extremes

Table 9

Table 9 Two-tier interpretation of the findings: conclusions that were robust across all analyses, and numerical estimates that are exploratory only.

Tier	Conclusion	Key supporting evidence	Status and scope
TIER 1 (robust)	Trajectory shape adds no incremental predictive value beyond the mean of early mock-examination scores	Shape features outperformed the mean-only benchmark in 0 of 202 interval scenarios (T3: delta R2 -3.56 to -0.03; T5: -0.08 to -0.06); the same rank order held in all four time-point x analysis-set combinations	Direction-consistent across algorithms, time points, outcome definitions and time-axis assumptions; based on fixed-fold 5-fold cross-validation
TIER 1 (robust)	Calendar examination dates are not required for this analysis	Rescaling the time axis by factors of 0.5-5.0 left every cross-validated R2 identical (maximum absolute deviation 0)	Robust by construction: level and shape features span the same space under any linear rescaling
TIER 1 (robust)	Tree-based algorithms do not transport across consecutive cohorts	Direction-dependent AUC differences of 0.196-0.230 versus 0.089	Qualitative finding; consistent across time points and analysis sets

		for the prespecified logistic model, while apparent AUCs were 0.993-1.000	
TIER 1 (robust)	The primary model is prespecified logistic regression rather than a selected ensemble	Chosen before analysis because events per variable was approximately 1.2 and threshold-based early warning needs absolute risks; in the T5 sensitivity combination LightGBM had a clearly worse Brier score (0.116-0.138) than logistic regression (0.078-0.083)	Prespecified; not revised after seeing the results
TIER 2 (exploratory)	Absolute AUC values in temporal external validation (A to B 0.696; B to A 0.608)	Point estimates rest on only 6-8 failure events per held-out cohort; the B to A 95% CI spans 0.249-0.966	Direction of transport is stable but magnitude is uncertain; must not be quoted as a fixed performance figure
TIER 2 (exploratory)	The 7% risk threshold and its operating characteristics (sensitivity 35.7%, specificity 81.9%, PPV 11.4%, 44 trainees flagged)	Derived from 14 events and this base's 6.1% prevalence; net benefit exceeds both default strategies only between thresholds of 3% and 29%	Base-specific reference value; must be re-derived locally from each base's prevalence and remedial capacity
TIER 2 (exploratory)	Individual model coefficients and odds ratios	Events per variable approximately 1.2; no coefficient approached significance (level_mean OR 0.514, 95% CI 0.100-2.725, p = 0.432; all other p >= 0.34); under the T5 sensitivity definition one indicator was completely separated and its OR was not estimable	Indicate direction only; not interpretable as effect sizes
TIER 2 (exploratory)	Latent class trajectory solution (five classes)	Lowest BIC among 30 candidate models, but entropy 0.075 and the smallest class contained only 2 trainees	Descriptive only; never used as a predictor and of no clinical predictive value

Figures

Figure 1

Figure 1 Latent class trajectories of standardised mock-examination scores (five-class solution). This analysis is exploratory and descriptive only. Although this model had the lowest Bayesian information criterion among 30 candidate models, entropy was low (0.075) and the smallest class contained only two trainees; class membership was therefore not used as a predictor and carries

no clinical predictive value. The figure should not be read as identifying meaningful trainee subgroups.

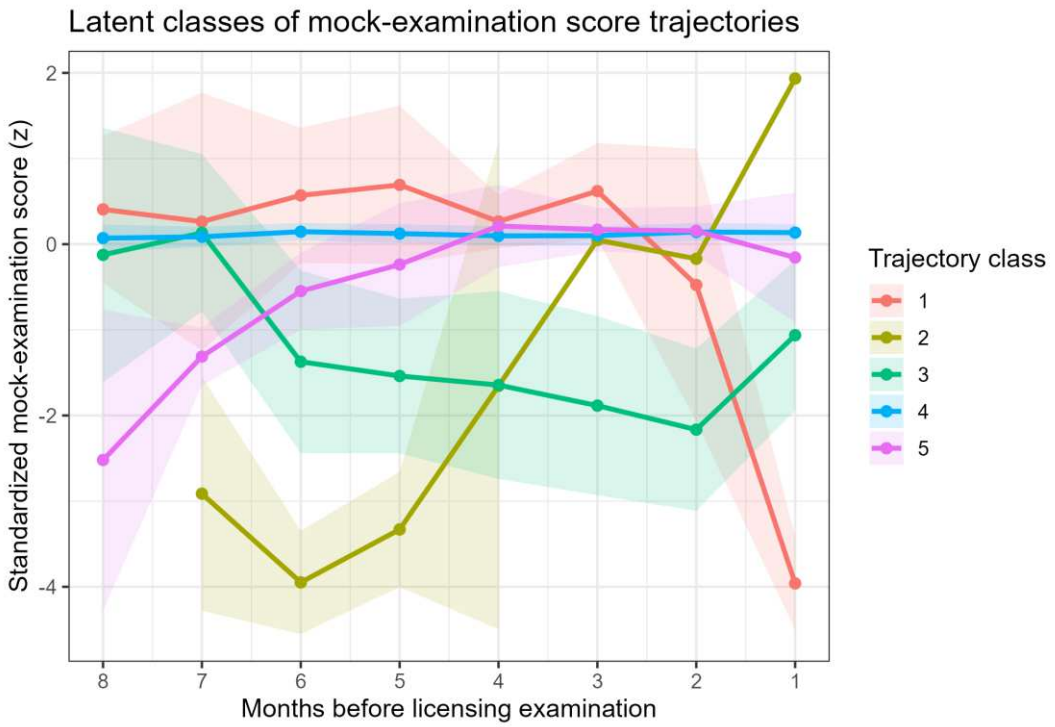


Figure 2

Figure 2 Calibration of the prespecified logistic model (T3; main analysis). The calibration slope shown in the panel is below the ideal value of 1.00, and observed proportions deviate from predicted probabilities across the range: absolute risks from the model are imperfect and require local recalibration before any threshold-based use.

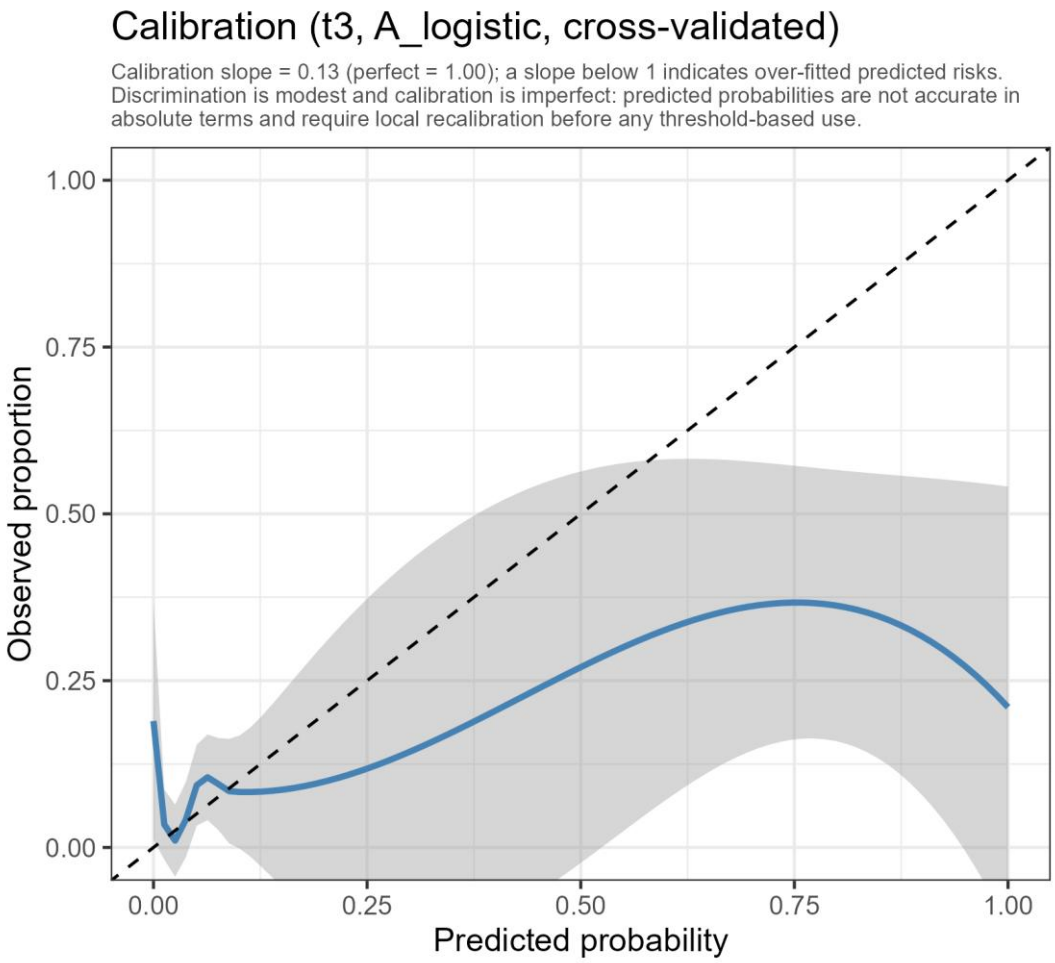


Figure 3

Figure 3 Decision-curve analysis of the prespecified logistic model (T3; main analysis). The panel indicates the threshold window within which the model outperforms both default strategies. The window is wide (3%–29%), but the absolute gain in net benefit is modest, as expected with moderate discrimination and a low event rate (prevalence 6.1%); the low positive predictive value is a consequence of that low prevalence rather than a failure of the model.

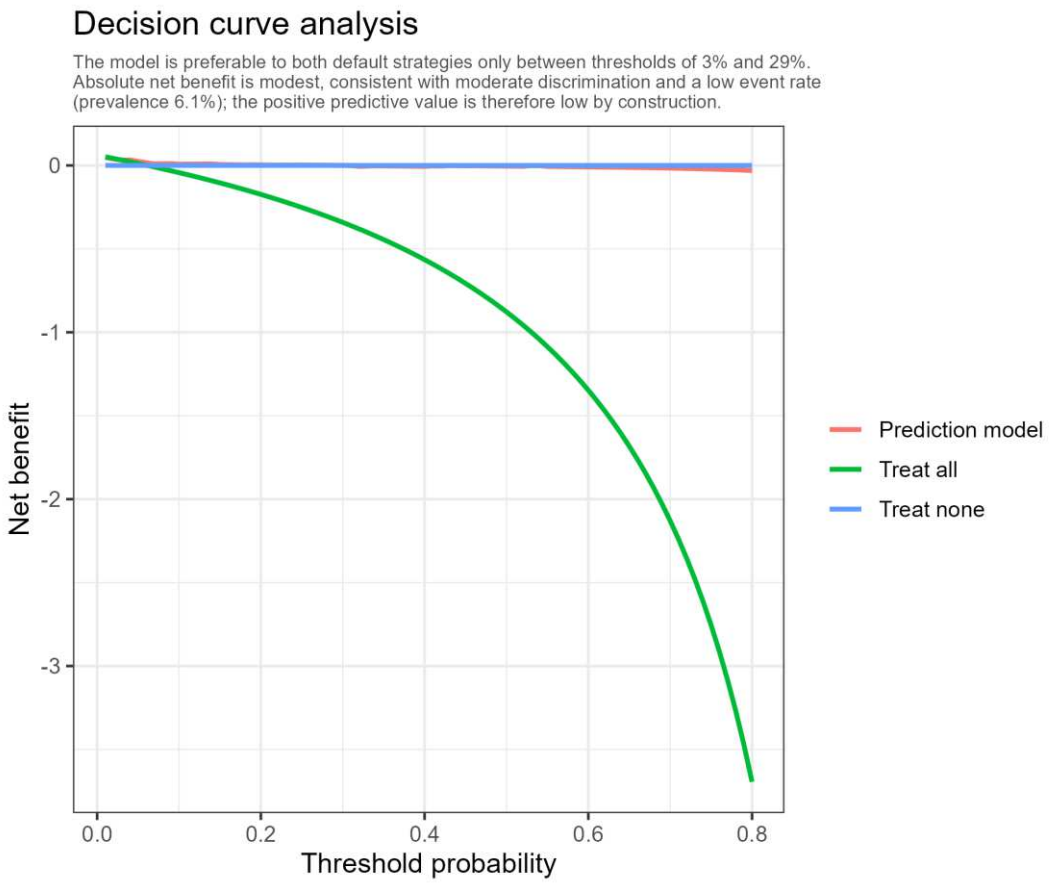


Figure 4

Figure 4 Temporal external validation (T3; main analysis) of the prespecified logistic model: receiver-operating-characteristic curves and calibration in the leave-one-cohort-out analysis. The calibration slopes for each held-out cohort are annotated; they deviate markedly from the ideal value of 1.00, so predicted probabilities do not transport between cohorts and local recalibration is mandatory. The number of failure events in each held-out cohort is also annotated; with so few events the confidence intervals around the AUCs are wide, and the point estimates should be treated as indicative only.

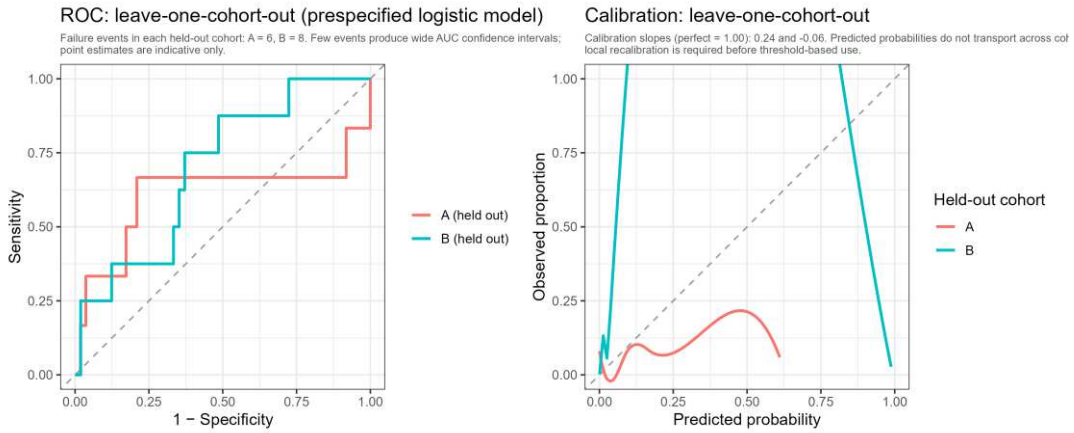
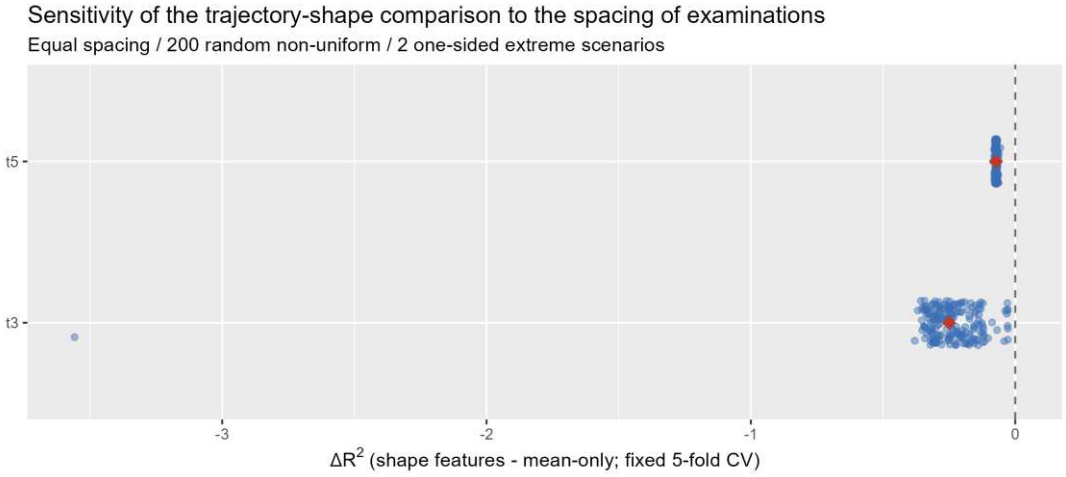


Figure 5

Figure 5 Sensitivity of the trajectory-shape comparison to 202 time-axis scenarios. The mean-only feature set matched or exceeded the shape-feature set in every scenario, confirming that the negative result for trajectory shape does not depend on assumptions about the spacing of mock examinations.



Supplementary Files

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

- TableS1.csv
- TableS2.csv
- TableS4.csv
- TableS8.csv
- TableS3.csv
- TableS5.csv
- TableS6.csv
- TableS7.csv
- TableS9.csv
- TableS10.csv
- TableS11.csv
- TableS12.csv
- TableS12coperatingpointst3main.csv
- TableS13.csv
- FigureS1.tif
- FigureS2.tif
- TableS12coperatingpointst5main.csv
- TableS12coperatingpointst3sensitivity.csv
- FigureS3.tif
- FigureS4.tif
- TableS12bnetbenefitgrid.csv
- TableS12coperatingpointst5sensitivity.csv
- FigureS5.tif
- FigureS6.tif
- TableS13brescalingdetail.csv
- TRIPODAIChecklistBMCMedEduc.csv
- TableS14fullintervalscenarios.csv