International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 25, No. 1, 2026

Risk-Set Misalignment Inflates Apparent Performance in OBE-Oriented Landmark Temporal Prediction

Ya Zhao[1, *], Yang Xu[2], Juan Huang[3], and Yang Zhang[3]

[1] School of Computer Science and Engineering, Hunan University of Information Technology, Changsha 414000, China
[2] School of International Business, Hunan University of Information Technology, Changsha 414000, China
[3] School of Management, Hunan University of Information Technology, Changsha 414000, China

Abstract: Temporal event prediction models are widely deployed in outcome-based education (OBE) and other domains where timely intervention depends on reliable prospective estimates. However, the trustworthiness of reported performance hinges on whether the evaluation cohort is aligned with the prospective decision context-specifically, whether subjects whose outcomes have already been realized are excluded at the prediction landmark. This paper presents a systematic evaluation framework that formalizes landmark temporal prediction as a risk-set alignment task, in which eligibility, feature availability and future outcomes are rigorously anchored to a common time point. Using 32,593 enrolment records and over ten million activity logs from the Open University Learning Analytics Dataset (OULAD), we evaluate logistic regression and histogram gradient boosting at five weekly landmarks under both a prospective landmark-aligned risk set and a retrospective diagnostic cohort. The results reveal a striking discrepancy: aligned ROC-AUC ranges from 0.634 to 0.679, whereas the retrospective cohort yields 0.872 at week 12-a difference driven by simultaneous changes in population, prevalence and outcome composition. Event-time stratification confirms that already-resolved events are substantially easier to classify than future events (ROC-AUC 0.905–0.966 vs. 0.623–0.683). These findings demonstrate that risk-set misalignment inflates apparent discrimination and compromises the trustworthiness of temporal prediction models in OBE and analogous prospective decision settings. We provide a reproducible reporting protocol to support comparable evaluation across studies.

Keywords: Outcome-based Education, Landmark Temporal Prediction, Risk-set Alignment, Trustworthy Evaluation, Machine Learning.

1. Introduction

Temporal event prediction-forecasting whether and when a subject will experience a target event-underpins decision-making in domains ranging from clinical medicine to predictive maintenance and outcome-based education (OBE). In all these settings, the practical value of a model depends on whether its evaluation faithfully reflects the prospective decision context: predictions must be assessed against subjects who remain at risk at the moment the forecast is issued. OBE provides a particularly instructive case because it requires continuous monitoring of progress toward defined outcomes, and withdrawal prediction serves as a natural landmark temporal task where risk-set alignment is both critical and frequently violated. The credibility of any such prediction system depends not only on the model but on whether the evaluation cohort excludes subjects whose outcomes have already been realized before the prediction landmark.

The Open University Learning Analytics Dataset (OULAD) [1] is a widely used benchmark for temporal event prediction in the OBE domain. It contains 32,593 enrolment records spanning 28,785 unique subjects and more than ten million virtual-learning-environment (VLE) activity records, making it a concrete testbed for examining how evaluation protocol choices influence reported prediction performance. However, the prediction tasks constructed from OULAD are heterogeneous. Some studies exclude withdrawn enrolments and predict pass versus fail [2,3]; others merge withdrawal with failure into a composite negative outcome [4,5]; and still others treat withdrawal as the primary binary endpoint [6,7]. These tasks address different estimands and their reported performance is not directly comparable.

Beyond task heterogeneity, a subtler source of incomparability involves temporal eligibility at the prediction landmark. When a prediction is issued at week k, some enrolments may have already been unregistered before that point. Retaining these resolved records changes the evaluated population: their outcome has already been realized and their behavioural traces may reflect patterns that are observable only in retrospect. Prior work has employed temporal feature windows [2,5,6,8], yet whether already-resolved events are included in, excluded from, or left ambiguous in the evaluation cohort has not been consistently reported. Jawad et al. [4] explicitly merged pre-window withdrawals into the negative class. Ujkani et al. [7] filtered early withdrawn records but did not provide a reproducible cutoff rule. Zhang and Ahn [5] and Shou et al. [6] did not explicitly report eligibility at the landmark. The distinction between specifying a feature window and enforcing landmark-aligned eligibility has not been systematically examined.

Within the full-text studies audited in this work, we found no controlled evaluation that held the landmark, feature window, student-level split and model family fixed while contrasting aligned and retrospective risk sets. This gap matters because changing the risk-set definition simultaneously alters the target population, event prevalence, outcome horizon and feature availability, making it impossible to attribute performance differences to any single factor.

To address this gap, we present, for the first time, a controlled evaluation that systematically contrasts landmark-aligned and retrospective risk sets under identical experimental conditions on a single dataset. This study makes

three contributions. First, we formalize withdrawal prediction on OULAD as a landmark risk-set task in which eligibility, feature availability and future outcomes are rigorously aligned at each prediction time point-a framework applicable to any temporal event prediction problem with time-varying risk sets. Second, we provide novel quantitative evidence showing how cohort composition and apparent performance diverge between the prospective estimand and a retrospective diagnostic cohort under fixed subject-grouped folds and model families. Third, we propose a comprehensive and reproducible reporting protocol-covering the landmark, eligibility rule, feature timing, split unit, prevalence and alert capacity-to support trustworthy and comparable evaluation of temporal prediction models in OBE and related domains. The evaluation is conducted on OULAD at five prediction landmarks (weeks 1, 2, 4, 8, 12) using class-weighted logistic regression (LR) and histogram gradient boosting (HGB) with fixed five-fold subject-grouped cross-validation.

2. Related Work

2.1. Task Definition Heterogeneity

OULAD records four final outcomes: Pass, Distinction, Fail and Withdrawn. How these categories are mapped to binary labels determines the estimand. He et al. [2] excluded withdrawn enrolments and predicted pass versus fail using weekly RNN features. Liu et al. [3] similarly removed withdrawals and focused on clickstream-based pass/fail classification. In contrast, Jawad et al. [4] merged Fail and Withdrawn into a single negative class and explicitly included enrolments that had withdrawn within the observation window. Qiu et al. [8] used a four-class outcome without collapsing categories. Shou et al. [6] proposed a withdrawal-specific binary task with course-percentage windows. These differences mean that reported accuracy, AUC or F1 values across studies reflect different target populations and cannot be directly compared. While task heterogeneity itself is well recognized, its interaction with temporal eligibility-which compounds the comparability problem-has received little attention.

2.2. Temporal Windows and Landmark Eligibility

Several studies constructed temporal feature windows to simulate real-time prediction. He et al. [2] used weekly cutoffs across presentations. Zhang and Ahn [5] employed semester-aware weekly windows with cross-presentation validation. Shou et al. [6] defined windows as percentages of course duration. However, specifying a feature window does not automatically determine who remains eligible for evaluation at that time point. An enrolment that was unregistered before the window cutoff has an already-realized outcome; retaining it changes the prediction task from a prospective forecast to a partly retrospective classification. Ujkani et al. [7] acknowledged this issue and filtered early withdrawn students, but the filtering rule was not fully reproducible. In summary, existing work has addressed when features are observed but not who should be evaluated at each landmark-a gap that conflates prospective and retrospective estimands and undermines cross-study comparability.

2.3. Data Splitting and Evaluation Validity

The 3,538 students with multiple enrolments in OULAD create a risk of information leakage when records from the same student appear in both training and test sets. Student-level grouping has been recommended for temporal prediction on longitudinal datasets [9], yet many OULAD studies use record-level random splits [10,11,12]. Additionally, class imbalance varies with the prediction task and landmark: aligned withdrawal prevalence ranges from 22.9% at week 1 to 12.5% at week 12, making threshold-dependent metrics such as accuracy or fixed-threshold F1 potentially misleading. Reporting ROC-AUC alongside prevalence-sensitive metrics such as Average Precision (AP) and capacity-aware metrics such as Recall@10% provides a more complete picture of operational utility. Taken together, the issues reviewed in Sections 2.1–2.3 indicate that a controlled evaluation isolating the effect of risk-set definition-while holding task definition, splitting strategy and metrics constant-is currently lacking.

3. Data and Task Formulation

3.1. Dataset and Analysis Unit

We used the Open University Learning Analytics Dataset (OULAD) [1], which records enrolment, activity and assessment data from courses delivered through a virtual learning environment. The dataset contains seven linked tables: studentInfo (32,593 rows), studentRegistration (32,593 rows), studentVle (10,655,280 rows), studentAssessment (173,912 rows), vle (6,364 rows), assessments (206 rows) and courses (22 rows). The unit of analysis is a single student–module–presentation enrolment record, identified by the triplet (id_student, code_module, code_presentation). Because 3,538 students held enrolments in more than one module (up to five), the 32,593 enrolment records correspond to 28,785 unique students. Throughout this paper, counts refer to enrolment records unless explicitly stated otherwise.

The official OULAD download endpoint has been discontinued. We obtained the data from the GitHub mirror maintained alongside the CRAN package ouladFormat (version 1.2.2) at ehoward1/oulad_data, with commit d451a05 for studentVle and 5bbf34a for the remaining six tables.

3.2. Prediction Landmarks and Temporal Boundaries

We defined five prediction landmarks at weeks $k \in \{1, 2, 4, 8, 12\}$, corresponding to exclusive cutoff days $t_k = 7k$. These landmarks were chosen to span the early engagement period (weeks 1–2), mid-course consolidation (week 4) and later stages (weeks 8–12), providing a comprehensive view of how predictive performance evolves as the observation window lengthens. At each landmark, the following temporal constraints were strictly enforced:

- *Registration requirement:* only enrolments with date_registration < t_k were eligible.
- *Feature window*: only events timestamped in the half-open interval [0, t_k) were used. Pre-course activity (negative dates) was excluded from the primary analysis.
- *Outcome timing*: positive labels (withdrawal) referred to events occurring at or after t_k, ensuring that the outcome had not yet been realized at the time of prediction.

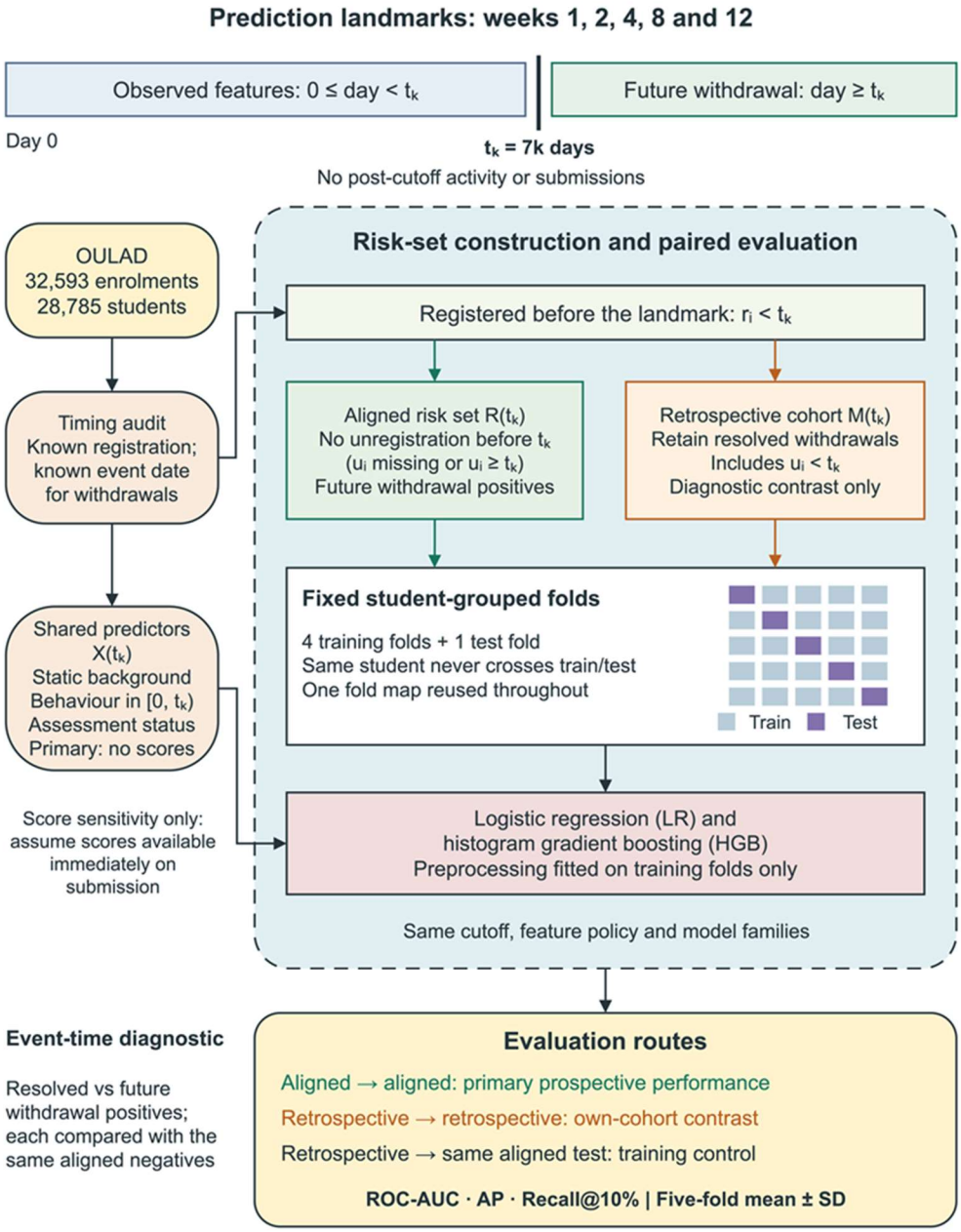


Fig. 1 illustrates the landmark evaluation framework.

Fig. 1. Risk-set-aware landmark evaluation for temporal event prediction in OBE. At each landmark, predictor construction is restricted to information timestamped before the cutoff. The prospective risk set contains enrolments that started before the landmark and were not unregistered before it. A retrospective cohort that retains already-resolved events is evaluated only as a diagnostic contrast under the same cutoff, feature policy, fixed subject-level fold map and model families.

3.3. Cohort Definitions

Let $Y_i = 1$ denote that enrolment i has a final result of Withdrawn, r_i its registration date and u_i its unregistration date. The *landmark-aligned risk set* at landmark t_k is:

$$R(t_k) = \{i : r_i < t_k \text{ and } (u_i \text{ is missing or } u_i \geq t_k)\} \quad (1)$$

Within $R(t_k)$, the positive class comprises enrolments that withdrew at or after t_k. This corresponds to the prospective estimand: predicting future withdrawal among enrolments currently amenable to intervention.

The *retrospective diagnostic cohort* is:

$$M(t_k) = \{i : r_i < t_k\} \quad (2)$$

This cohort retains enrolments whose withdrawal was already resolved before t_k. It does not estimate the same quantity as $R(t_k)$; differences in apparent performance between the two cohorts reflect simultaneous changes in population, prevalence and outcome composition.

3.4. Timing Inconsistencies

Three categories of records required special handling. First, 93 enrolment records had final_result = Withdrawn but no date_unregistration; because we could not determine whether the withdrawal occurred before or after each landmark, these

records were excluded from all landmark analyses. Second, 45 records lacked a registration date and were similarly excluded. Third, 9 records had a date_unregistration but a final result other than Withdrawn; these were handled by the standard eligibility rule and reported separately in the cohort audit.

3.5. Feature Construction

We constructed three feature groups from information available before each landmark. All preprocessing was fitted on training folds only.

Static features (8 variables): gender, region, highest education, Index of Multiple Deprivation (IMD) band, age band, number of previous attempts, studied credits and disability status.

Behavioural features (31–32 variables): click counts by each of the 20 VLE activity types; total clicks, number of active weeks, first and last active week; weekly click mean, standard deviation and maximum; silence duration (weeks since last activity), inactive weeks; clicks in the most recent one and two weeks and their share of total clicks; and a linear trend slope fitted to the weekly click series. Weekly statistics were computed on a complete week grid from week 0 to week $k-1$, with zero-filled entries for weeks with no activity, so that silence was explicitly represented rather than treated as missing data.

Assessment features (12–13 variables): number of submissions before t_k, number of assessments due before t_k, number of due assessments not yet submitted, mean and maximum submission delay, total assessment weight, and TMA/CMA-specific submission counts. The variable asm_missed was computed by matching individual assessment IDs between the due and submitted sets.

Score-timing sensitivity: The primary analysis excluded assessment scores because OULAD does not record when scores were released to students. A sensitivity analysis added five score-related features (mean, minimum, last, TMA mean, CMA mean score) under the assumption that scores were available immediately upon submission.

4. Experimental Design

4.1. Cross-validation

A single five-fold stratified student-grouped split was generated (seed 20260916) and reused across all landmarks, cohort definitions, feature policies and models. All enrolments from the same id_student were confined to one fold. This design ensured that fold-to-fold comparisons were paired and that no student-level information crossed from training to test.

4.2. Models

We evaluated two model families with complementary inductive biases, deliberately chosen to represent the two dominant paradigms in tabular prediction-linear and tree-based-so that any protocol effect observed would not be contingent on a single model architecture:

Logistic regression (LR): categorical variables were imputed with the training-fold mode and one-hot encoded; numerical variables were imputed with the training-fold median (with missing indicators) and standardized. Class weight was set to "balanced" with regularization $C = 1.0$ and a maximum of 3,000 iterations.

Histogram gradient boosting (HGB): categorical variables were imputed with the training-fold mode and ordinally encoded; numerical variables retained native missing values handled by the HGB splitting procedure. Class weight was "balanced", with 300 boosting iterations and learning rate 0.1. Internal early stopping was disabled because its random validation split would break student grouping.

4.3. Evaluation Comparisons

Four comparisons were conducted at each landmark:

(1) *Primary estimand:* aligned training → aligned test. This is the main result for prospective withdrawal prediction.

(2) *Retrospective own-cohort*: retrospective training → retrospective test. This shows apparent performance under the commonly used cohort definition, but differs from the primary estimand in population, prevalence and outcome composition.

(3) *Training-population control*: retrospective training → the same aligned test folds. By holding the test records identical to comparison (1), any difference isolates the effect of changing the training population.

(4) *Event-time stratification*: within retrospective test folds, positives were separated into resolved-before-landmark positives ($u_i < t_k$) and future-withdrawal positives ($u_i \geq t_k$), each evaluated against the same aligned at-risk negatives. This diagnostic assessed whether already-resolved withdrawals constituted a more easily separable subgroup.

4.4. Metrics

Three complementary metrics were selected to capture different facets of prediction utility:

- ROC-AUC: threshold-free ranking discrimination, chosen because it is invariant to the decision threshold and widely reported in prior OULAD studies, facilitating comparison.
- *Average Precision* (AP): the area under the precision–recall curve computed by average_precision_score; reported alongside positive prevalence because the no-skill AP baseline equals prevalence. AP is particularly informative under class imbalance, where ROC-AUC alone may overstate practical utility.
- *Recall@10%*: the fraction of true positives recovered when alerts are issued to the top 10% of ranked test records, reflecting a fixed intervention capacity. This metric directly addresses the operational question facing practitioners: how many at-risk subjects can be identified given limited resources?
- *Uncertainty*: arithmetic mean ± sample standard deviation across five folds. These are descriptive fold-level statistics, not confidence intervals; no hypothesis test was performed.

5. Results

5.1. Cohort Composition

Fig. 2 shows how landmark-aligned and retrospective cohort definitions produced markedly different evaluation populations. The aligned risk set decreased from 29,002 enrolment records at week 1 to 25,629 at week 12 as resolved withdrawals were progressively excluded. The retrospective cohort remained nearly constant at approximately 32,400 records. Aligned positive prevalence declined from 22.9% to 12.5%, whereas retrospective prevalence was stable at 30.9%. Critically, the proportion of retrospective positives whose withdrawal had already occurred before the landmark grew

from 33.5% at week 1 to 68.0% at week 12 (Table 1).

Table 1. Cohort composition and timing audit at five prediction landmarks.

Week	Aligned n	Aligned prevalence	Retrospective n	Retrospective prevalence	Resolved-positive share
1	29,002	22.9%	32,363	30.9%	33.5%
2	27,970	20.0%	32,397	30.9%	44.2%
4	27,435	18.3%	32,437	30.9%	49.8%
8	26,423	15.2%	32,446	30.9%	60.0%
12	25,629	12.5%	32,451	30.9%	68.0%

Note. Forty-five records without registration dates and 93 withdrawn records without event dates were excluded from all landmark analyses. Nine non-withdrawn records with a dated unregistration were handled by the eligibility rule and retained in the audit.

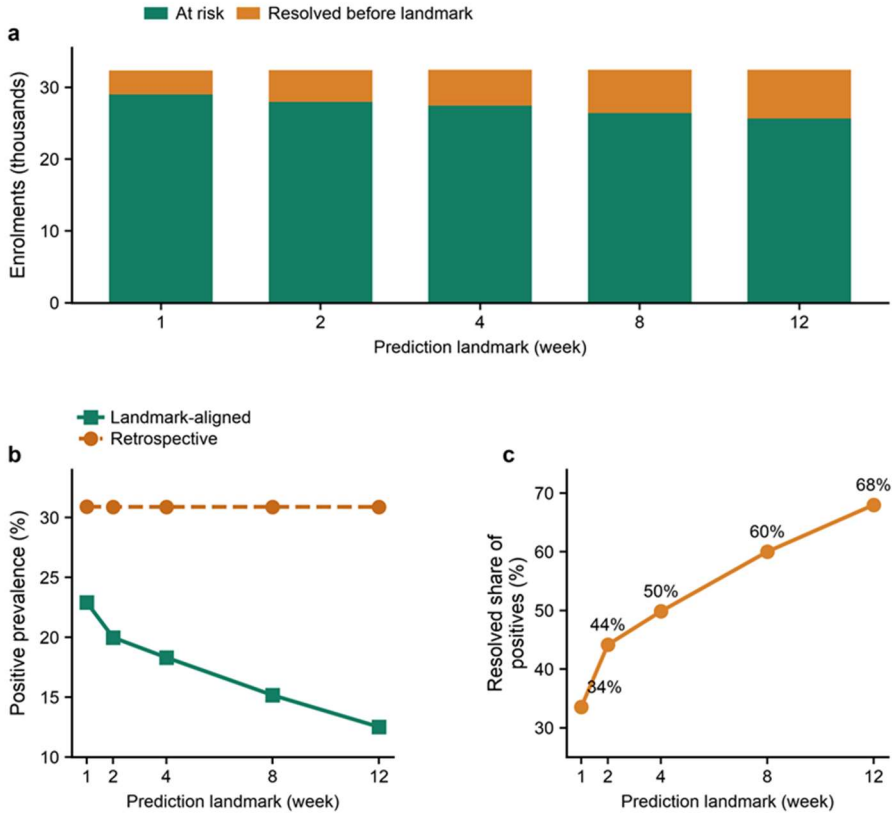


Fig. 2 Landmark eligibility changes cohort composition.

Fig. 2 Landmark eligibility changes cohort composition. (a) Enrolment records retained at each prediction landmark, partitioned into the landmark-aligned risk set (green) and additional already-resolved records retained by the retrospective definition (orange). (b) Positive prevalence under both definitions. (c) Proportion of retrospective positives whose withdrawal had already occurred before the landmark. Counts are deterministic cohort summaries; no uncertainty intervals are shown.

5.2. Landmark-aligned Primary Performance

Table 2 and Fig. 4a–c present the primary results on the landmark-aligned risk set under the no-score feature policy. HGB ROC-AUC was 0.646 ± 0.008 at week 1, dipped to 0.634 ± 0.011 at week 2, then recovered to 0.675 ± 0.012 at week 8 and stabilized through week 12 (0.675 ± 0.007). LR followed a similar trajectory, reaching 0.679 ± 0.008 at week 8. Notably, the pattern was non-monotonic: extending the observation window did not uniformly improve discrimination, suggesting that the additional behavioural signal accumulated between weeks 1 and 2 was insufficient to offset the loss of the earliest withdrawals from the risk set.

AP declined from 0.354 ± 0.011 at week 1 to 0.231 ± 0.013 at week 12 for HGB, and from 0.348 ± 0.010 to 0.220 ± 0.009 for LR. This decline must be interpreted alongside the concurrent drop in aligned prevalence from 22.9% to 12.5%, which lowered the no-skill AP baseline. The week-4 AP for both models (HGB: 0.311; LR: 0.318) exceeded the week-2 values, indicating a non-monotonic pattern that cannot be attributed to prevalence alone and likely reflects genuine improvement in the informational content of the accumulated behavioural features.

Recall@10% increased modestly from 0.188 ± 0.007 at week 1 to 0.219 ± 0.012 at week 12 (HGB), meaning that a fixed 10% alert capacity covered only about one-fifth of future withdrawals even at the latest landmark. This finding

has important practical implications: practitioners operating under constrained intervention resources should expect that the majority of positive cases will fall outside the alert threshold.

Table 2. Landmark-aligned early-withdrawal prediction under the no-score primary feature policy.

Week	Model	ROC-AUC	AP	Recall@10%	n	Prevalence
1	HGB	0.6456 ± 0.0081	0.3535 ± 0.0113	0.1884 ± 0.0067	29,002	22.9%
1	LR	0.6425 ± 0.0096	0.3484 ± 0.0095	0.1863 ± 0.0070	29,002	22.9%
2	HGB	0.6343 ± 0.0106	0.3012 ± 0.0139	0.1809 ± 0.0111	27,970	20.0%
2	LR	0.6323 ± 0.0124	0.2995 ± 0.0101	0.1806 ± 0.0137	27,970	20.0%
4	HGB	0.6601 ± 0.0136	0.3115 ± 0.0136	0.2064 ± 0.0098	27,435	18.3%
4	LR	0.6748 ± 0.0112	0.3180 ± 0.0114	0.2130 ± 0.0091	27,435	18.3%
8	HGB	0.6751 ± 0.0115	0.2722 ± 0.0109	0.2218 ± 0.0128	26,423	15.2%
8	LR	0.6786 ± 0.0076	0.2652 ± 0.0092	0.2129 ± 0.0107	26,423	15.2%
12	HGB	0.6754 ± 0.0073	0.2310 ± 0.0133	0.2192 ± 0.0117	25,629	12.5%
12	LR	0.6730 ± 0.0079	0.2197 ± 0.0094	0.2160 ± 0.0139	25,629	12.5%

Note. Values are mean ± sample standard deviation across five fixed student-grouped test folds. AP denotes Average Precision. Recall@10% is the fraction of positives among alerts issued to the highest-scored 10% of each test fold. Fold variability is descriptive; no confidence interval or significance test is implied.

5.3. Descriptive Contrast with the Retrospective Cohort

Fig. 3a–b present the own-cohort comparison. Retrospective HGB ROC-AUC rose from 0.735 ± 0.003 at week 1 to 0.872 ± 0.003 at week 12, whereas aligned HGB ROC-AUC reached only 0.675 ± 0.007 at week 12. The descriptive ROC-AUC difference grew substantially from +0.089 at week 1 to +0.197 at week 12 for HGB, and from +0.085 to +0.186 for LR. Retrospective AP similarly exceeded aligned AP by +0.228 at week 1 and by as much as +0.570 at week 12 (HGB). These widening gaps underscore a critical observation: the longer the course progresses, the larger the proportion of already-resolved withdrawals in the retrospective cohort, and the more inflated the apparent performance becomes. These contrasts are descriptive comparisons of different estimands-the populations, prevalences, prediction horizons and outcome compositions all differed between the two cohorts.

5.4. Training-population Control and Event-time Stratification

When the test records were held identical (the same aligned folds) and only the training cohort was changed from aligned to retrospective, the effect was small. HGB ROC-AUC increased by +0.003 to +0.009 across the five landmarks. LR ROC-AUC moved in the opposite direction, declining by −0.007 to −0.023, with the magnitude growing at later landmarks. Both sets of differences were an order of magnitude smaller than the own-cohort contrasts reported in Section 5.3 (Fig. 3d). The large own-cohort apparent difference could not be explained by changing the training population alone; it co-occurred with changes in the test population, prevalence, prediction horizon and event composition.

Event-time stratification (Fig. 3c) provided further insight. When evaluated against the same aligned at-risk negatives, resolved-before-landmark positives achieved ROC-AUC values of 0.905–0.966 across landmarks and models, whereas future-withdrawal positives achieved only 0.623–0.683. At week 12, HGB yielded 0.963 ± 0.002 for resolved positives versus 0.680 ± 0.009 for future positives; LR yielded 0.958 ± 0.002 versus 0.650 ± 0.011. Already-resolved withdrawals constituted a markedly more separable subgroup, which helps explain why the retrospective cohort-where such records formed an increasing share of positives-appeared to present an easier classification task.

5.5. Score-timing Sensitivity

Under the assumption that assessment scores were available immediately upon submission, aligned ROC-AUC at weeks 1 and 2 changed negligibly (paired difference < 0.005 for both models). At weeks 4 through 12, score availability provided modest gains: +0.019 to +0.028 in ROC-AUC and +0.023 to +0.043 in AP (Fig. 4d). At week 12, the LR gains were +0.028 ± 0.010 (ROC-AUC), +0.043 ± 0.011 (AP) and +0.041 ± 0.010 (Recall@10%). Because OULAD does not record score-release timestamps, these gains remain conditional on an unverified immediate-availability assumption and should not be interpreted as observed real-time performance improvements.

5.6. Cross-model and Cross-landmark Synthesis

Across both models and all five landmarks, the retrospective own-cohort ROC-AUC and AP consistently exceeded the aligned values, and resolved-positive ROC-AUC consistently exceeded future-positive ROC-AUC. The robustness of these two patterns across the evaluated LR and HGB models-which represent fundamentally different inductive biases-provides strong evidence that the observed discrepancy is a property of the evaluation protocol rather than an artifact of a particular model family. In contrast, the training-population control yielded differences that were model-dependent in direction and small in magnitude, confirming that the dominant source of apparent performance inflation lies in the test-side cohort composition, not in the training data.

6. Discussion

6.1. Risk-set Alignment Changes the Estimand

The central finding of this study is that the risk-set definition materially changed the evaluated cohort and the apparent performance of temporal event prediction. At week 12, the retrospective cohort retained 68.0% of its positives as already-resolved withdrawals, yet its ROC-AUC of 0.872 might be taken at face value as the achievable discrimination

for prospective prediction. Our landmark-aligned evaluation yielded a ROC-AUC of 0.675 at the same time point-a striking descriptive difference of nearly 0.20 points. This gap arose not from a single mechanism but from the simultaneous change in population, prevalence (12.5% vs. 30.9%), prediction horizon and outcome composition. The training-population control showed that changing the training cohort alone, while holding the test records fixed, accounted for only a small and model-dependent fraction of this difference. To the best of our knowledge, this is the first controlled evaluation to systematically quantify this discrepancy on a widely used benchmark dataset.

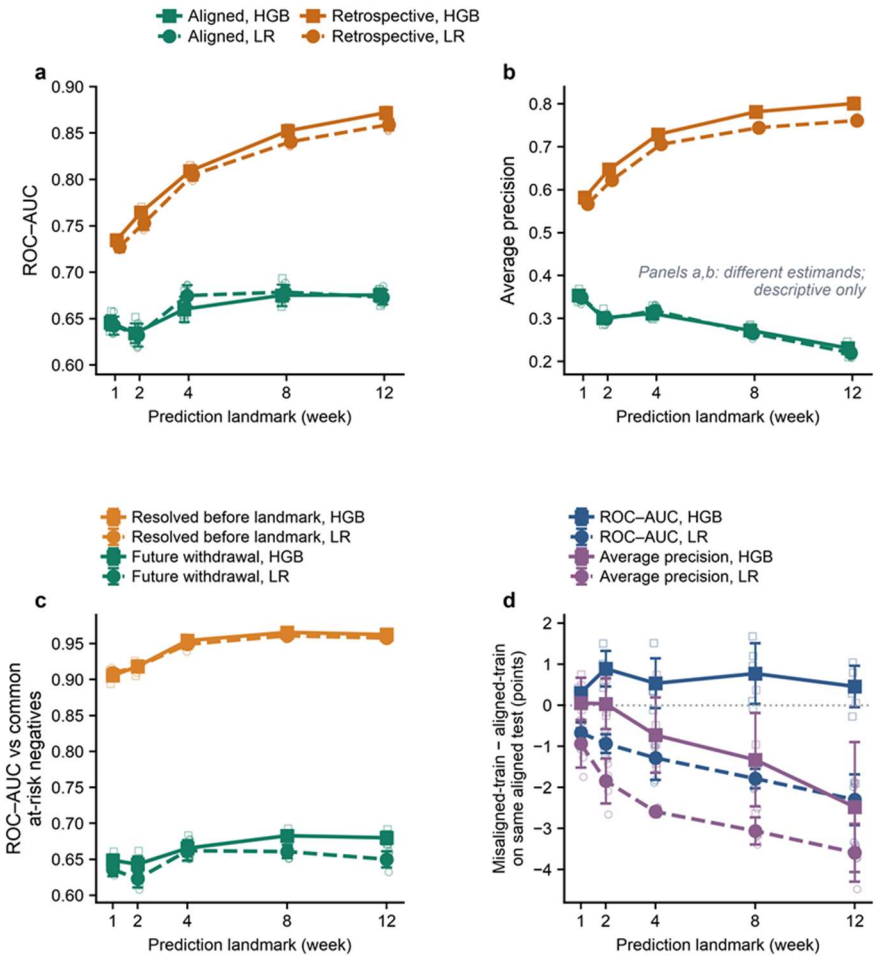


Fig. 3 Risk-set definition changes apparent discrimination and reveals event-time separability.

Fig. 3. Risk-set definition changes apparent discrimination and reveals event-time separability. (a) ROC-AUC and (b) AP for landmark-aligned (green) and retrospective (orange) cohorts. (c) ROC-AUC for resolved-before-landmark positives (orange) and future-withdrawal positives (green), each compared with the same aligned at-risk negatives. (d) Paired differences in ROC-AUC and Average Precision (AP), expressed in percentage points (retrospective minus aligned), between models trained on the retrospective and aligned cohorts when both were evaluated on identical aligned test records. HGB: squares and solid lines; LR: circles and dashed lines. Hollow symbols show fold values; filled symbols show fold means; error bars are sample standard deviations.

6.2. Already-resolved Withdrawals are Easier to Distinguish

Event-time stratification revealed that resolved-before-landmark positives were far more separable from at-risk negatives (ROC-AUC 0.905–0.966) than future-withdrawal positives (0.623–0.683). This is consistent with the expectation that subjects who have already experienced the target event exhibit distinctive behavioural patterns-such as prolonged silence and abrupt cessation of activity-that are observable in the logged data, whereas subjects who will experience the event in the future may still be actively participating at the time of prediction. The retrospective cohort mixes these two subgroups, producing an apparent discrimination that substantially understates the difficulty of the prospective prediction task. This mechanism provides a concrete explanation for why retrospective evaluations in the literature may report optimistically high performance figures.

6.3. Mapping to the Literature

Our literature audit identified three distinct evidence states among OULAD studies. Jawad et al. [4] explicitly merged pre-window withdrawals into the failure class-a confirmed instance of risk-set misalignment. Ujkani et al. [7] explicitly filtered early withdrawn students but without a reproducible cutoff. Zhang and Ahn [5] and Shou et al. [6] employed temporal windows but did not report whether already-resolved withdrawals were retained. The distinction between these states matters: “confirmed misalignment”, “explicit

filtering without a reproducible rule" and "not reported" have different implications for result interpretation. Our controlled comparison provides the first quantitative context for assessing how much these protocol choices may influence reported performance within a single dataset, offering a reference point for future meta-analytic comparisons.

6.4. Implications for Trustworthy Temporal Prediction in OBE and Beyond

The risk-set alignment problem identified in this study is not unique to OBE-oriented withdrawal prediction; it generalizes to any temporal event prediction task where subjects leave the at-risk pool before the prediction landmark-including clinical event forecasting, subscription churn prediction and predictive maintenance. In each setting, retaining already-resolved subjects conflates prospective and retrospective estimands and inflates apparent discrimination.

Within the OBE context, outcome-based frameworks require that progress toward defined outcomes be monitored through formative mechanisms. Temporal prediction models serve as a component of this monitoring loop: they translate behavioural signals into actionable forecasts that enable timely intervention. Our results demonstrate that the trustworthiness of these forecasts depends fundamentally on whether the evaluation protocol reflects the prospective decision context in which interventions are deployed.

The achievable prospective discrimination (ROC-AUC 0.634–0.679) and the limited coverage of a 10% alert capacity (Recall@10% of 0.186–0.222) provide substantially more realistic planning benchmarks than the retrospective figures commonly reported in the literature. For practitioners designing intervention workflows, these aligned estimates should inform resource allocation decisions: the gap between retrospective and prospective performance suggests that systems calibrated on retrospective metrics may systematically overestimate their reach.

The reporting protocol formalized in this study-covering landmark, eligibility rule, feature timing, split unit, prevalence and alert capacity-may serve as a template for specifying evaluation conditions in any temporal prediction task with time-varying risk sets, supporting the transparency and reproducibility that trustworthy machine learning demands.

6.5. Limitations

Several limitations should be acknowledged. The evaluation was conducted on a single public dataset from one institution offering distance-delivered courses; generalizability to other institutional settings and application domains remains untested. OULAD does not record score-release timestamps, so the score-timing sensitivity analysis relied on an unverified assumption of immediate availability.

On the data side, 93 withdrawn records without event dates and 45 records without registration dates were excluded. Including the former as a sensitivity analysis would change aligned prevalence but not the fundamental protocol comparison. Nine non-withdrawn records with unregistration dates were handled by the eligibility rule and reported in the audit.

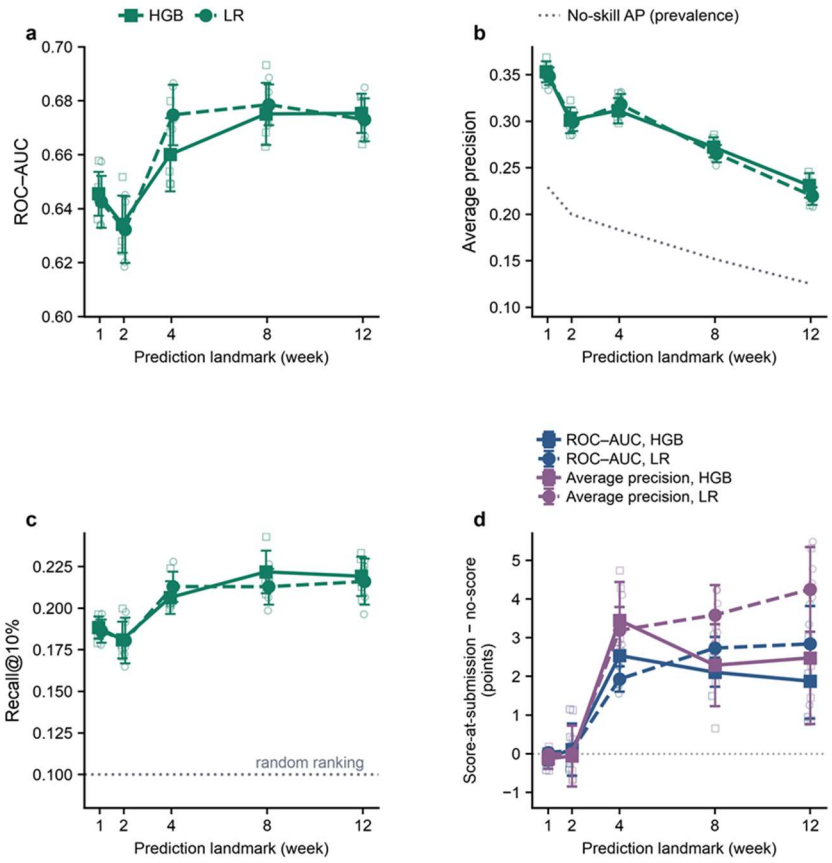


Fig .4 Prospective aligned performance and sensitivity to assumed score availability.

Fig. 4. Prospective aligned performance and sensitivity to assumed score availability. (a) ROC-AUC. (b) AP, with the

landmark-specific positive prevalence as the no-skill baseline (grey dotted line). (c) Recall@10%, with the random-ranking baseline at 0.10 (grey dotted line). (d) Paired difference between the score-at-submission sensitivity policy and the no-score primary policy. HGB: squares and solid lines; LR: circles and dashed lines. Hollow symbols show individual fold values; filled symbols and connecting lines show fold means; error bars are sample standard deviations across the five folds.

On the modelling side, hyperparameters were not tuned; the evaluation aimed to compare estimands under fixed modelling choices rather than to maximize predictive performance. The reported uncertainties are sample standard deviations across five folds and do not constitute formal confidence intervals. No external validation on a second dataset was performed; cross-presentation holdout within OULAD provides only internal validation.

6.6. Future Directions

Several avenues merit further investigation. First, replicating the landmark-aligned evaluation on datasets from other institutions and domains (clinical event prediction, subscription churn, predictive maintenance) would test the generalizability of the risk-set alignment effect. Second, incorporating richer temporal representations—such as sequence models over weekly engagement trajectories—may improve prospective discrimination while maintaining alignment constraints. Third, integrating the proposed evaluation protocol into real-time deployment pipelines would allow the framework to be validated under operational conditions where intervention decisions are actually made. Finally, coupling the temporal prediction framework with closed-loop intervention studies would provide causal evidence on whether aligned predictions lead to improved outcomes in OBE programs and other applied settings.

7. Conclusion

This study has formalized landmark temporal prediction as a risk-set alignment task and provided the first controlled quantitative evidence of how the risk-set definition changes the evaluated cohort and apparent performance on the OULAD benchmark. Three principal conclusions emerge from this work.

First, we have demonstrated that eligibility, predictors and outcomes must be rigorously aligned at a common landmark. When already-resolved events were retained in the evaluation cohort, the apparent ROC-AUC at week 12 was 0.872, compared with 0.675 under landmark alignment-a descriptive difference of nearly 0.20 points that reflected simultaneous changes in the evaluated population, prevalence and outcome composition rather than a single causal mechanism.

Second, our results establish that performance metrics must be reported alongside the eligibility rule, positive prevalence and alert capacity. Average Precision and Recall@10% are both sensitive to the risk-set definition and the resulting prevalence. Without these contextual details, cross-study comparison of reported prediction performance remains unreliable.

Third, we have shown that feature-availability assumptions require explicit disclosure. Assessment-score availability provided meaningful gains only at later landmarks and only under the assumption of immediate release-a condition that cannot be verified from OULAD alone.

These findings carry direct implications for temporal prediction in OBE and analogous prospective decision settings: models calibrated on retrospective metrics risk overestimating their intervention reach, potentially misguiding resource allocation decisions. The risk-set alignment framework and reporting protocol proposed here generalize beyond the OBE domain to any temporal event prediction task with time-varying at-risk populations. We encourage future studies to adopt this protocol-covering landmark, eligibility rule, feature window boundary, split unit, prevalence and alert capacity-as standard evaluation metadata, and to extend this framework to multi-institutional settings and other application domains where the alignment challenge may manifest differently.

Acknowledgments

This work was supported by the University-Level Teaching Reform Project of Hunan University of Information Technology (Project No. 6), “Exploration and Practice of an OBE-Based Talent Training Model for Engineering Management under the New Engineering Education Initiative”.

References

[1] Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. Scientific Data, 4, 170171. https://doi.org/10.1038/sdata.2017.171

[2] He, Y., Chen, R., Li, X., Hao, C., Liu, S., Zhang, G., & Jiang, B. (2020). Online at-risk student identification using RNN-GRU joint neural networks. Information, 11(10), 474. https://doi.org/10.3390/info11100474

[3] Liu, Y., Fan, S., Xu, S., Sajjanhar, A., Yeom, S., & Wei, Y. (2023). Predicting student performance using clickstream data and machine learning. Education Sciences, 13(1), 17. https://doi.org/10.3390/educsci13010017

[4] Jawad, K., Shah, M. A., & Tahir, M. (2022). Students’ academic performance and engagement prediction in a virtual learning environment using random forest with data balancing. Sustainability, 14(22), 14795. https://doi.org/10.3390/su142214795

[5] Zhang, C., & Ahn, H. (2023). E-learning at-risk group prediction considering the semester and realistic factors. Education Sciences, 13(11), 1130. https://doi.org/10.3390/educsci13111130

[6] Shou, Z., Xie, M., Mo, J., & Zhang, H. (2024). Predicting student performance in online learning: A multidimensional time-series data analysis approach. Applied Sciences, 14(6), 2522. https://doi.org/10.3390/app14062522

[7] Ujkani, B., Minkovska, D., & Hinov, N. (2024). Course success prediction and early identification of at-risk students using explainable artificial intelligence. Electronics, 13(21), 4157. https://doi.org/10.3390/electronics13214157

[8] Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., Jiang, B., & Chen, P. K. (2022). Predicting students’ performance in e-learning using learning process and behaviour data. Scientific Reports, 12(1), 453. https://doi.org/10.1038/s41598-021-03867-8

[9] Bilal, M., Omar, M., Anwar, W., Bokhari, R. H., & Choi, G. S. (2022). The role of demographic and academic features in a student performance prediction. Scientific Reports, 12(1), 12508. https://doi.org/10.1038/s41598-022-15880-6

[10] Poudyal, S., Mohammadi-Aragh, M. J., & Ball, J. E. (2022). Prediction of student academic performance using a hybrid 2D CNN model. Electronics, 11(7), 1005. https://doi.org/10.3390/electronics11071005

[11] Qiu, F., Zhu, L., Zhang, G., Sheng, X., Ye, M., Xiang, Q., & Chen, P. K. (2022). E-learning performance prediction: Mining the feature space of effective learning behavior. Entropy, 24(5), 722. https://doi.org/10.3390/e24050722

[12] Bettahi, A., Belouadha, F. Z., & Harroud, H. (2025). A modular and explainable machine learning pipeline for student dropout prediction in higher education. Algorithms, 18(10), 662. https://doi.org/10.3390/a18100662