education sciences

MDPI

Article

Evaluating Data-Minimized Early-Warning Systems for Educator Decision Support: Evidence Across Review Capacities and Educational Stages

Aigul Shaikhanova [1,2], Oleksandr Kuznetsov [3,4,*], Gulmira Shangytbayeva [5,*], Kamila Bakenova [1,6], Kainizhamal Iklassova [7] and Dana Tulemisova [8]

1 Global Education and Training, iSchool at Illinois, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA; aiguls@illinois.edu (A.S.); bakenova.k@au.edu.kz (K.B.)
2 Department of Information Security, L.N. Gumilyov Eurasian National University, Satpayev 2, Astana 010008, Kazakhstan
3 Department of Theoretical and Applied Sciences, eCampus University, Via Isimbardi 10, 22060 Novedrate, CO, Italy
4 Department of Intelligent Software Systems and Technologies, School of Computer Science and Artificial Intelligence, V. N. Karazin Kharkiv National University, 4 Svobody Sq., 61022 Kharkiv, Ukraine
5 Department of Informatics and Information Technologies, Zhubanov University, 34 Aliya Moldagulova Avenue, Aktobe 030000, Kazakhstan
6 Center for Scientific Research and Commercialization, Arkalyk Pedagogical University, 17 Auelbekov Str., Arkalyk 110300, Kazakhstan
7 Department of Information and Communication Technologies, Manash Kozybayev North Kazakhstan University, Pushkin Str. 86, Petropavlovsk 150000, Kazakhstan; keiklasova@ku.edu.kz
8 School of Cybersecurity, Astana IT University, 55/11 Mangilik El Avenue, Astana 010000, Kazakhstan; d.tyulemissova@astanait.edu.kz
* Correspondence: oleksandr.kuznetsov@uniecampus.it (O.K.); gshangytbayeva@zhubanov.edu.kz (G.S.)

Abstract

Educational early-warning systems rank students for review when support capacity is limited. We examined how removing background information changes retrospective review-list performance in the Open University Learning Analytics Dataset (OULAD) and two separately analyzed UCI education datasets. Full-information models were compared with reduced-input models across several educational stages and four review capacities (5%, 10%, 20%, and 30%). The primary outcome was the difference in review-list precision, with paired bootstrap intervals adjusted across stages and capacities within each policy family. Sex/gender removal showed no statistically detectable difference after this adjustment; this does not establish equivalence. Removing socioeconomic/family information reduced whole-cohort predictive performance in OULAD and UCI 697 and reduced precision in selected larger OULAD lists and at UCI 697 enrollment. UCI 320 estimates were inconclusive. Some adjusted precision findings varied across calibration methods. Diagnostic models showed that excluded attributes remained partly predictable from retained records. These post-development analyses evaluate predictive performance, not causal effects on educator workload, interventions, or student outcomes. The findings support evaluating data minimization at the intended educational stage and review capacity, while separately assessing whole-cohort performance and list membership. Prospective evaluation with educators and students is needed before drawing conclusions about educational benefit.

Keywords: learning analytics; educational early-warning systems; educator decision support; data minimization; review capacity; capacity-aware educational technology integration; responsible educational technology

Check for updates

Academic Editors: Musa Adekunle Ayanwale and Damola Olugbade

Received: 5 September 2026
Revised: 15 October 2026
Accepted: 21 October 2026
Published: 28 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Educational early-warning systems help educators identify students who may need support. Their usefulness depends on how predictions fit into institutional review workflows: a teacher, adviser, or support team may be able to review only a small proportion of a cohort. Evaluation should therefore consider review-list precision, recall, and membership at that capacity, as well as predictive performance across the whole cohort.

This distinction has a direct practical meaning. If an institution can review 10% of a cohort, review-list precision is the proportion of students in that 10% list who later experience the adverse outcome; recall is the proportion of all subsequent adverse outcomes represented in the list. Precision–recall measures are particularly informative when outcomes are imbalanced and capacity is constrained (Saito & Rehmsmeier, 2015). Yet a useful retrospective list is only a prerequisite for effective support. Whether an alert improves counseling, teaching, retention, or achievement must ultimately be established in prospective educational practice (Plak et al., 2022).

The choice of information used to construct a review list is equally important. Student-background variables may improve prediction, but their collection raises questions about necessity, proportionality, interpretation, and governance. Data minimization is a lifecycle design principle; deleting a field does not by itself make a system fair, private, or lawful (Drachsler & Greller, 2016; European Parliament and Council of the European Union, 2016; Rets et al., 2023). Information related to an excluded attribute may remain in other records. Separately, two models with similar aggregate performance may direct educators' attention to different students. Evaluating data minimization therefore requires the intended educational stage, review quota, comparison model, and consequences for review-list membership to be stated explicitly.

Previous work has established the importance of capacity-based alerting, calibration, subgroup auditing, repeated prediction stages, and alert-list stability (Wu et al., 2026; Wu & Luo, 2026). Leakage-aware OULAD protocols likewise show why only information genuinely available at the prediction time should be used (Le et al., 2026). Empirical educational studies have examined different ways to restrict predictive inputs. Chiu et al. obtained early performance predictions across 236 undergraduate subjects using relative learning-management-system engagement measures without demographic or prior academic inputs (Chiu et al., 2026). Hasan and Fritz used OULAD to show that gender could be inferred from behavioral records, and examined feature selection and adversarial censoring to reduce this inference while retaining course-outcome prediction (Hasan & Fritz, 2022). These studies distinguish prediction with fewer inputs from changes to the representation of retained information; neither approach makes field deletion alone a privacy guarantee.

Implementation research adds a complementary perspective. Rets et al. reviewed ten publications from four years of developing and implementing an early-alert dashboard, linking ethical data use to stakeholder involvement, inclusive support, and checks for hidden bias (Rets et al., 2023). The present study addresses a narrower question: how field-removal policies change review lists across educational stages and review capacities. It complements these studies with paired estimates of predictive changes, list membership, and residual attribute predictability. It does not measure privacy improvement or the effects of educational interventions.

Our earlier article used OULAD to study early risk prediction, interpretability, and educator workload under temporal validation (Shaikhanova et al., 2025). The present study reuses the educational setting and elements of that modeling design; neither OULAD nor workload-aware prediction is claimed as new. The added contribution is a paired evaluation of targeted and broader information removal across educational stages and four review capacities, with family-wise uncertainty intervals, separate UCI analyses, and diagnostic

tests of residual attribute predictability. This evaluates whether a data-removal conclusion at one operating point remains supported under other conditions. OULAD, UCI 697, and UCI 320 are interpreted separately rather than pooled as replications.

The study addresses four questions:

1. How does removing student-background information change the usefulness of review lists when educators can examine 5%, 10%, 20%, or 30% of a cohort?
2. Do these consequences change as additional evidence becomes available at later educational stages?
3. Do sex/gender-only removal and socioeconomic/family removal show different predictive consequences across the three educational datasets?
4. Can an excluded attribute still be predicted from retained records, and what does this imply for responsible institutional use?

2. Materials and Methods

2.1. Study Design, Evidentiary Status, and Provenance

This is a retrospective methodological evaluation of early-warning models using existing, de-identified educational datasets. The framework was developed iteratively from earlier analyses and then fixed before the final multi-capacity, attribute-predictability, and harmonized-block evaluations. No review capacity, educational stage, feature definition, model grid, alert rule, resampling plan, or reporting threshold was changed after that point. The study is a post-development retrospective robustness evaluation, not pre-registered confirmatory research.

Table 1 links the four research questions to the analyses. Review capacity denotes a specified fraction of records to examine, not a measured amount of staff time.

Table 1. Research questions, educational stages, information-removal scenarios, and outcome measures.

Question	Datasets and Stages	Information-Removal Comparison	Main Measures
RQ1: capacity	All three datasets; all stages listed below	Dataset-specific reduced-input policies versus full information at 5%, 10%, 20%, and 30% capacity	Primary: precision difference; secondary: recall and list overlap
RQ2: stage	OULAD: days 14–90 (six stages); UCI 697: enrollment and semester one; UCI 320: baseline, G1, G1+G2, separately by subject	Each policy versus its stage-matched full control; OULAD common-cohort sensitivity	Precision differences by stage; secondary AP, AUROC, and Brier score
RQ3: information group	Same datasets and stages as RQ2	Sex/gender-only and socioeconomic/family removal versus full information	Precision differences across stages and capacities; secondary whole-cohort metrics
RQ4: residual predictability	OULAD and UCI 697: gender and disability/special needs; UCI 320: sex, at the corresponding stages	Diagnostic prediction of each excluded attribute from retained records	Diagnostic AUROC, AP relative to prevalence, and positive-case counts

AP: average precision; AUROC: area under the receiver operating characteristic curve.

A provenance audit identified that an earlier recovery copy of OULAD encoded 1111 missing area-deprivation values in a manner inconsistent with the official archive. Because this changed preprocessing, all current OULAD results were regenerated from a

fixed version of the official data and compared with controls from the same execution. The original release was also replayed on that version. Checksums identify the exact input files; the Supplementary Materials document the transition between runs and the computational evidence. Current inferences use the corrected run rather than assuming that all historical results are unaffected.

Figure 1 summarizes the study as an educator-facing decision process. The datasets supply evidence available at different educational stages, the paired models test alternative information policies, and the capacity rule converts model scores into lists containing specified fractions of eligible records. The corresponding staff time remains to be measured.

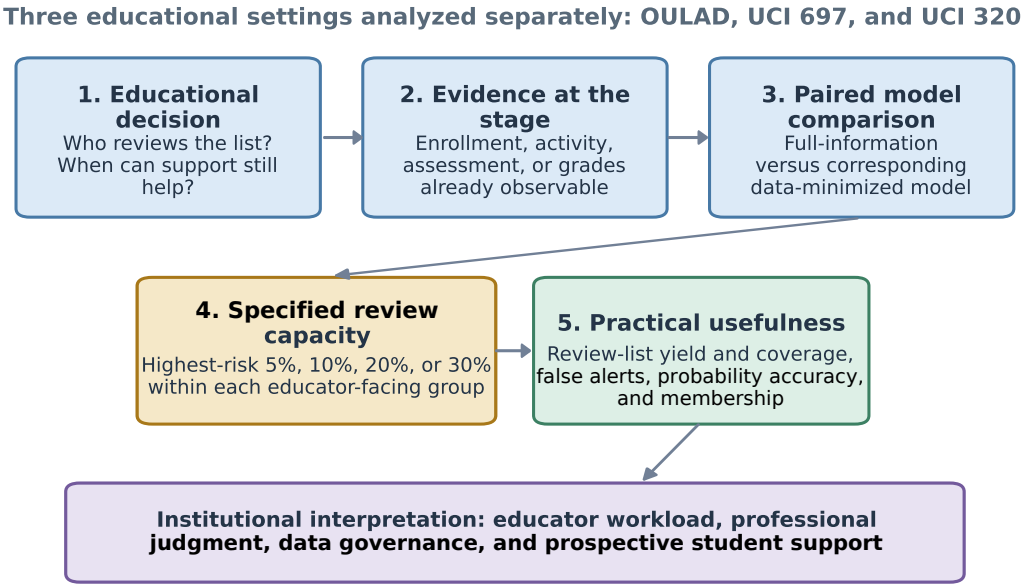


Figure 1. Capacity-aware framework for evaluating an educational early-warning system before prospective use. OULAD, UCI 697, and UCI 320 represent different educational settings and were analyzed separately. The framework links the intended educational decision and evidence available at that stage to paired model comparisons, specified review quotas, and measures of retrospective predictive performance.

2.2. OULAD Setting, Outcome, and Temporal Partitions

OULAD contains 32,593 student–module–presentation registrations across 22 presentations of seven anonymized modules, together with assessment and virtual-learning-environment records (Kuzilek et al., 2017). AtOn each observation day, the analysis included registrations that had begun and had not yet ended in withdrawal. The binary outcome was defined as future Fail or Withdrawn versus Pass or Distinction. Data from 2013B and 2013J were used for fitting, 2014B for model selection and calibration, and 2014J for temporal evaluation. These partitions contain distinct, temporally ordered presentations; they are not necessarily disjoint at the student level because a student may have multiple registrations. Review lists were constructed after days 14, 28, 42, 56, 70, and 90.

Because students who withdraw leave later risk sets, the size and composition of the evaluated cohort change over time. This dynamic truncation limits interpretation of stage-wise comparisons, which do not identify the optimal time for intervention. A sensitivity analysis followed registrations present across the first four or all six observation days. Conditioning on later continuation can introduce selection effects, as discussed in Section 4.4.

2.3. Time-Aware Evidence and Data-Removal Scenarios

Only information available aton the observation day was used. Virtual-learning-environment features summarized cumulative and recent activity, active days, timing, inactivity, pre-start activity, and change in click activity. An assessment submission was included only if both its due date and submission date were no later than the observation day. Scores were excluded because OULAD does not record when grades became visible to educators or students, avoiding the use of grade information with uncertain temporal availability.

All standard models used course context, previous attempts, studied credits, registration timing, and the activity and assessment-process evidence available at that stage. The *full-information* model additionally used gender, region, highest education, Index of Multiple Deprivation band, age band, and disability. The *six-field removal* policy omitted all six; *gender/disability removal* omitted only those two fields. Single-field analyses and the sex/gender-only and socioeconomic/family blocks were separate comparisons. Supplementary Table S14 lists the exact removed fields for each field-exclusion policy and dataset. The term *data-minimized* refers to model inputs, not deletion throughout the source dataset or institutional data lifecycle.

2.4. Predictive Models, Selection, and Calibration

For OULAD, histogram gradient boosting was implemented using scikit-learn (Pedregosa et al., 2011). Missing numeric values were imputed using training-set medians, with missingness indicators retained. Categorical variables were imputed using the most frequent training category and one-hot encoded. The numeric branch, including missingness indicators, was standardized. All preprocessing transformations were fitted exclusively on training data. Model complexity was selected on the validation presentation by average precision, with simpler models preferred in a tie. Platt scaling was fitted on validation predictions and applied without modification to the temporal evaluation presentation. The implementation fits a logistic regression to the logit of the raw model probability after clipping it to $[10^{-6}, 1 - 10^{-6}]$; the fitted slope was required to be positive in the OULAD implementation. This two-parameter form provides a parsimonious recalibration step in the fixed modeling pipeline, adjusting the location and spread of the log odds while retaining the direction of the risk score. It was not selected through an empirical comparison of calibration methods. Platt scaling has been evaluated for boosted trees, but this does not establish its optimality for the histogram gradient-boosting models used here (Niculescu-Mizil & Caruana, 2005). Probability accuracy was evaluated using the Brier score; calibration was not assumed to guarantee accurate probabilities (Collins et al., 2024; Moons et al., 2025; Van Calster et al., 2019). The released code specifies the model and preprocessing parameters.

To assess calibration-method sensitivity, we additionally compared the archived Platt probabilities with raw model probabilities and probabilities from isotonic regression fitted only to the original validation or inner-calibration samples. Outcome models, hyperparameters, evaluation records, group quotas, and deterministic tie keys were held fixed. Isotonic regression offers a flexible monotone mapping but can overfit small calibration samples and introduce tied probabilities (Niculescu-Mizil & Caruana, 2005). The comparison was a post-development sensitivity analysis; no calibrator was selected using evaluation outcomes. Brier score, log loss, reliability summaries, review-list membership, and paired precision differences were evaluated. Log loss used the same 10^{-6} clipping for all methods. Conditional paired precision intervals used 5000 common resamples per stage, with the original student-cluster or row-level resampling and stage–capacity multiplicity families. Probability-score differences used pointwise 95% intervals. Full methods and aggregate results are provided

in the Supplementary Materials. Beta calibration remains an untested parametric alternative (Kull et al., 2017). Neither Brier score nor log loss isolates calibration quality.

All comparisons pair a data-minimized model with a full-information control created under the same data split, feature-availability rules, model family, and software environment. Selected models and predictions were fixed before the uncertainty calculations.

2.5. Review Capacity and Educational Interpretation of the Metrics

Students were ordered by estimated risk separately within each operational group g: a module in OULAD, a course in UCI 697, or a school in UCI 320. Review capacity was allocated within these groups, rather than using a single global ranking. At capacity b, the number placed in each group's review list was

$$k_g = \lceil bn_g \rceil, \qquad b \in \{0.05, 0.10, 0.20, 0.30\}.$$

Here n_g is the number of eligible records in group g. Review capacity is a quota of records, not staff hours or the number of interventions delivered. In OULAD, a record is a student–module–presentation registration, so different selected records can refer to the same student. Rounding within groups can make the overall selected fraction slightly exceed b. The primary measure was review-list precision, defined as the proportion of reviewed records with the subsequent adverse outcome. Recall was the proportion of all records with that outcome captured by the list:

$$P_b = \frac{\sum_g TP_g}{\sum_g k_g}, \qquad R_b = \frac{\sum_g TP_g}{\sum_g N_{+,g}}.$$

At a 10% review capacity, these measures are denoted by Precision@10% and Recall@10%. Exact score ties at a list boundary were resolved by a pre-specified deterministic key; a sensitivity analysis averaged over possible orders within tied boundary scores.

Average precision (AP) summarized ranking quality across possible thresholds, area under the receiver operating characteristic curve (AUROC) measured overall discrimination, and the Brier score assessed probability accuracy. The original OULAD decision rule treated a loss greater than three percentage points in day-42 Precision@10% as practically concerning. Because this threshold was selected during development, it serves as an operational heuristic rather than a formal non-inferiority margin. Appendix A provides a plain-language guide to these measures and their institutional interpretation.

2.6. OULAD Paired Uncertainty and Common-Cohort Sensitivity

Uncertainty in the fixed-10% OULAD analysis was estimated by resampling students and retaining all their registrations. Both review lists were reconstructed from the same resampled records in every replicate. Day 42 used 5000 replicates, and the other stages used 2000. The family-wise intervals in this analysis adjust across six observation days at the fixed 10% capacity. The final multi-capacity analysis instead adjusts across all 24 stage–capacity combinations per policy (Section 2.12); its intervals support the overall capacity conclusions. Models were not refitted during resampling, so model-selection and calibration decisions remained fixed.

The four-horizon common cohort contained 8853 registrations from 8536 students; the six-horizon cohort contained 8507 registrations from 8216 students. Within each, day-42-minus-day-14 difference-in-differences compared the change in each exclusion gap with the full-control change. These analyses are conditional on later registration retention and may therefore be subject to selection effects.

2.7. Fallback Models Without Clickstream Data

To examine a setting in which learning-platform activity data are unavailable, day-42 fallback models excluded all clickstream variables and retained course context and assessment-process information that was due and observable by that day. An enriched version added further process counts and ratios, but no grades, click telemetry, or future records. Model selection and calibration followed the same validation protocol as the standard models.

At a 10% review capacity, the pre-specified operational criterion permitted the enriched fallback to fall no more than five percentage points below the standard six-field removal comparator. This comparison evaluates whether a clickstream-free fallback remains practically adequate, rather than only improving on a less informative fallback.

2.8. External Educational Datasets and Cross-Fitted Controls

Two official UCI datasets were evaluated subsequent to the OULAD analysis and are reported as distinct external analyses. UCI 697, *Predict Students' Dropout and Academic Success*, contains 4424 higher-education students, 36 predictors, and the outcomes Dropout, Enrolled, and Graduate (Realinho et al., 2021). The primary comparison contrasted Dropout with Graduate and excluded students who remained enrolled; a secondary outcome compared Dropout with all other statuses. The primary analysis thus targets a selected group with resolved outcomes. It may be subject to selection bias when interpreted as performance for all entrants; the direction and size of that bias were not estimated. The secondary outcome retains Enrolled records but treats current status as observed, without establishing their eventual outcomes. Enrollment models used only information available on entry. Semester-one models added first-semester curricular-unit evidence while prohibiting all second-semester fields. Data-removal scenarios omitted direct personal information, family and financial information, or gender and educational special-needs status. A sparse operational comparator retained only a small set of admissions and course variables.

UCI 320, *Student Performance*, provides separate Mathematics ($n = 395$) and Portuguese ($n = 649$) samples from two secondary schools (Cortez, 2008). The two subject samples were not pooled because the available linkage fields did not uniquely identify the same students across files. The outcome was final-grade failure (G3 < 10). Baseline models excluded all grades; the next stages added G1 and then G2, while always excluding the final grade G3. Data-removal scenarios omitted sex, omitted demographic and family fields, omitted sensitive and wellbeing fields, or retained only school-observable information.

Both UCI analyses used ten shared stratified outer cross-validation folds. Model selection and calibration were conducted only within each outer training fold, and logistic regression served as a sensitivity model. Review capacity was allocated within course for UCI 697 and within school for UCI 320. Paired resampling used 5000 replicates for the primary gradient-boosting comparisons and 2000 for secondary comparisons. Family-wise intervals covered the two UCI 697 or three UCI 320 educational stages within each policy family. These datasets broaden the settings examined. They are not treated as independent confirmations because their analyses followed the OULAD work and their educational contexts and outcome definitions differ.

2.9. Evaluation Across Four Review Capacities

Before the final capacity analysis was computed, review fractions of 5%, 10%, 20%, and 30% were pre-specified within the post-development protocol. For each fraction, we recalculated the number of selected records, true and false alerts, precision, recall, lift over prevalence, list overlap, and the proportion retained from the full-information list. Existing OULAD test probabilities and UCI out-of-fold probabilities were reused; only the

decision capacity changed, without model refitting. Paired uncertainty for performance differences is described in Section 2.12; Supplementary Table S2 summarizes the primary precision families.

2.10. Residual Predictability After Direct Attribute Removal

Separate diagnostic models evaluated whether an excluded attribute could still be predicted from the information retained in the corresponding data-minimized model. Logistic regression served as the primary diagnostic model, with gradient boosting used as a nonlinear sensitivity analysis. OULAD diagnostics targeted gender and disability and were evaluated temporally on 2014J, including a subset of students not observed during fitting. UCI 697 diagnostics targeted gender and educational special-needs status, and UCI 320 diagnostics targeted sex within the saved outer folds. Educational outcomes, identifiers, split fields, future-stage information, and the target attribute itself were excluded from the diagnostic models. These diagnostics measure residual predictability. They do not establish actual proxy use, proxy discrimination, or privacy leakage by the early-warning models.

2.11. Comparable Information Groups Across Datasets

Two information groups were defined to support transparent cross-dataset comparisons. The *sex/gender* group comprised the corresponding recorded sex or gender field in each dataset. The *socioeconomic/family* group comprised highest education and area deprivation in OULAD; parental qualifications and occupations, debtor and tuition status, and scholarship status in UCI 697; and parental education and occupations in UCI 320. These operational groupings do not imply that the datasets measure an identical social construct or that the measures are invariant across datasets.

Existing paired comparisons were reused where available. Only missing gradient-boosting comparisons needed for the two information groups were fitted, each with a full-information control created in the same execution. Controls were required to reproduce the corresponding previously saved probabilities to numerical precision before a comparison was accepted. No UCI 697 outcome model was refitted. A previously preserved XuetangX analysis was excluded because the official dataset schema did not support verifiable demographic comparisons.

2.12. Cross-Dataset Paired Uncertainty and Multiplicity

Review lists were reconstructed in every bootstrap replicate. OULAD resampled at the student level and retained repeated registrations, whereas the UCI analyses resampled at the row level. The same resampling multiplicities were applied to full-information and data-minimized predictions at all four capacities. Primary gradient-boosting comparisons used 5000 replicates and sensitivity comparisons used 2000. For review-list metrics, each family covered all pre-specified stages and capacities for a given dataset, outcome, model, and data-removal policy; global probability metrics covered stages. We report pointwise and Bonferroni-adjusted family-wise 95% percentile intervals. The primary comparison was the difference in review-list precision between the data-minimized and full-information models. Recall, list overlap, AP, AUROC, and Brier-score differences were secondary. An interval that includes zero is interpreted as inconclusive, not as proof of equivalence. This distinction matters particularly when the datasets differ in sample size and precision.

2.13. Computational Reproducibility

Inputs, data partitions, feature-availability rules, fitted pipelines, predictions, resampling draws, tables, and figures were verified using executable scripts. The final audit independently reconstructed all primary capacity results and statistical intervals from

archived computational artifacts. The Supplementary Materials document counts, hashes, and replay tolerances; the public repository provides the released analysis and verification code, aggregate results, and bootstrap draws. Student-level records, individual predictions, and fitted model bundles are not redistributed. These checks establish internal reproducibility, but do not replace independent institutional validation or prospective evaluation with educators and students.

3. Results

The primary outcome is the data-minimized-minus-full difference in review-list precision. Unless identified as calibration sensitivity, results use the original Platt-calibrated primary analysis. Results are organized by dataset, with the final comparisons across stages and all four capacities presented before supporting results at the fixed 10% capacity. Whole-cohort metrics and list membership are secondary outcomes; single-field, common-cohort, and clickstream-free analyses provide supporting evidence. Residual attribute predictability addresses RQ4 separately. Section 3.4 summarizes the findings across datasets without pooling them.

3.1. OULAD

3.1.1. Primary Comparisons Across Review Capacities

The OULAD results use the official archive, a deterministic tie-breaking rule, and corrected same-run controls. Some intermediate counts differed from the superseded recovery run; no result was adjusted to match historical values. Supplementary Table S11 documents the transition.

Six-field removal produced lower precision point estimates in 19 of the 24 stage–capacity combinations. After family-wise adjustment across all 24 combinations, losses remained at the 30% capacity on days 14, 28, and 56. Gender/disability removal produced no precision difference whose adjusted interval excluded zero at any evaluated stage or capacity. Thus, the fixed-10% result alone did not describe performance across the full range of evaluated review fractions. Table 2 summarizes these comparisons; Supplementary Table S15 retains the detailed point estimates.

Table 2. Summary of primary Platt-calibrated review-list precision findings across the four capacities. Findings refer to intervals adjusted across stages and capacities within each policy family; detailed selected-policy point estimates appear in Supplementary Table S15.

Dataset	Policies Summarized	Adjusted Precision Findings
OULAD	Six-field; gender/disability removal	Six-field losses at 30% on days 14, 28, and 56. No detected gender/disability difference.
UCI 697	Sparse admissions versus full information	Enrollment losses at all four capacities; semester-one losses at 20% and 30%.
UCI 320	Evaluated removal policies; subjects separate	No detected precision difference in Mathematics or Portuguese.

No detected difference means that the adjusted interval includes zero; it does not establish equivalence. UCI 697 uses the primary Dropout-versus-Graduate outcome.

For the comparable information groups, sex/gender-only removal also produced no detectable precision difference after adjustment. Removing socioeconomic/family information—highest education and area deprivation—reduced whole-cohort ranking and probability accuracy aton every observation day. Adjusted precision losses remained at the 30% capacity on days 28 and 56 (Table 3; Figure 2). AThe absence of an adjusted precision difference does not establish equivalence.

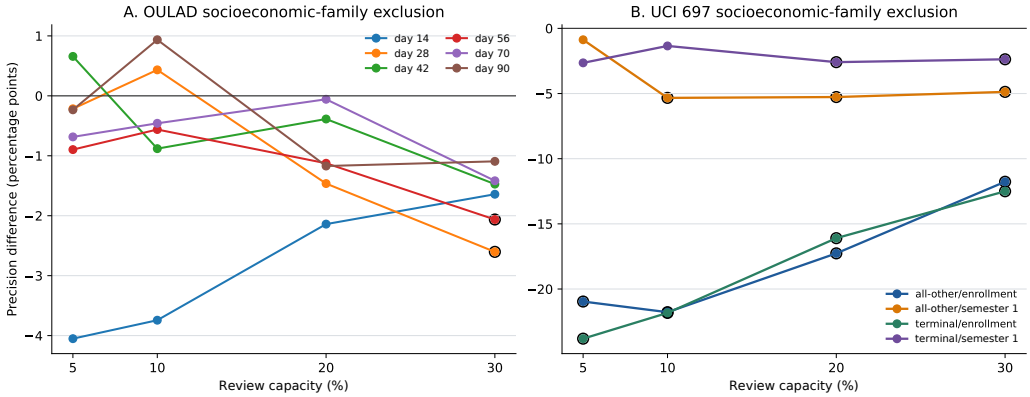


Figure 2. Change in review-list precision after removing socioeconomic and family-related information under the original Platt-calibrated primary analysis. Open black circles identify family-wise intervals that exclude zero. Lines connect capacities within one educational stage. OULAD and UCI 697 use different populations, outcomes, and field definitions and are not pooled.

Table 3. Comparable information groups and adjusted precision findings under the primary Platt calibration. The fields represent dataset-specific constructs. Full point-estimate ranges and interval counts, including the secondary UCI 697 outcome, appear in Supplementary Table S16.

Dataset	Sex/Gender-Only Removal	Socioeconomic/Family Removal
OULAD	Gender: no detected difference	Highest education and deprivation: losses at 30% on days 28 and 56.
UCI 697	Gender: no detected difference	Parental and financial fields: enrollment losses at all capacities; semester-one losses at 20% and 30%.
UCI 320, Mathematics	Sex: no detected difference	Parental education and occupations: no detected difference.
UCI 320, Portuguese	Sex: no detected difference	Parental education and occupations: no detected difference.

All statements concern family-wise precision intervals across stages and capacities. No detected difference does not establish equivalence. UCI 697 uses Dropout versus Graduate. Exact field names are listed in Supplementary Table S14.

3.1.2. Supporting Comparisons at the Fixed 10% Capacity

At day 42, 9040 active registrations were eligible for review and 3584 later ended in Fail or Withdrawn. At a 10% capacity, each model produced a list of 908 registrations. The full-information model identified 769 subsequent adverse outcomes, the six-field removal model identified 762, and the gender/disability removal model identified 765.

Six-field removal achieved day-42 precision of 0.8392, compared with 0.8469 for full information. Both lists identified approximately 84 subsequent adverse outcomes per 100 reviewed registrations. The paired difference was -0.0077 (95% interval $[-0.0265, 0.0155]$); its lower bound remained above the negative three-percentage-point operational reference.

At day 14, six-field removal changed precision by -0.0406: the list contained 670 subsequent adverse outcomes, compared with 708 among the same number of reviewed registrations in the full-information list. The interval adjusted across six days at 10% was below zero ($[-0.0692, -0.0036]$). Under the final adjustment across all 24 stage–capacity combinations, however, the interval reached zero ($[-0.0735, 0.0000]$). These are different multiplicity families; the latter does not establish a difference at this operating point. AP was lower aton every observation day, a separate finding about whole-cohort ranking.

Removing gender and disability produced day-42 precision of 0.8425, compared with 0.8469 for full information (Table 4). The difference was less than half a percentage point,

and its interval included zero. Small AUROC and Brier-score differences nevertheless favored the full-information model. Similar list yield therefore did not imply identical ranking or probability estimation across the full cohort.

Table 4. Day-42 comparison of targeted gender/disability removal with the full-information model.

Metric	G/D Excluded	Full	Difference	95% Interval
Alerts	908	908	—	—
True-risk alerts	765	769	−4	—
False alerts	143	139	+4	—
Precision@10%	0.8425	0.8469	−0.0044	[−0.0110, 0.0099]
Recall@10%	0.2134	0.2146	−0.0011	[−0.0028, 0.0025]
Average precision	0.6934	0.6943	−0.0009	[−0.0022, 0.0005]
AUROC	0.7585	0.7603	−0.0018	[−0.0027, −0.0009]
Brier score	0.1915	0.1909	+0.0006	[0.0003, 0.0008]

Intervals use 5000 paired student-level bootstrap replicates and are conditional on the fitted models. Positive Brier differences favor the full-information model.

Across the six observation days, gender/disability precision differences ranged from −0.0044 at day 42 to +0.0035 at day 90 (Table 5; Figure 3). Every pointwise interval included zero, and all six-stage family-wise lower bounds remained above the negative three-percentage-point reference. This post-development comparison is descriptive, not a confirmatory equivalence or non-inferiority test.

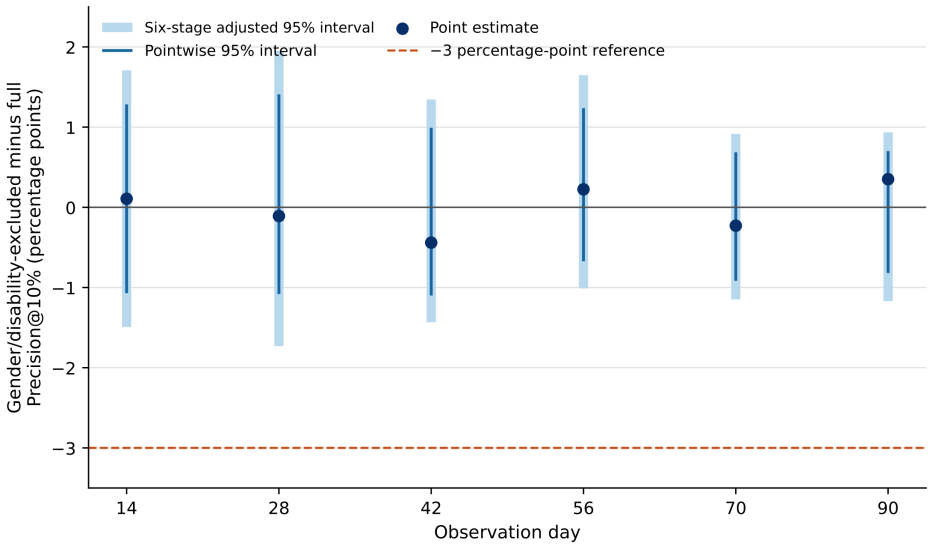


Figure 3. Difference in review-list precision after targeted removal of gender and disability. Dark intervals are pointwise paired bootstrap intervals; pale intervals adjust across six days at the fixed 10% capacity, not across the four capacities. The dashed line marks the three-percentage-point operational reference.

The secondary AP difference at day 14 was −0.00342 (six-stage family-wise interval [−0.00613, −0.00071]). Together with the day-42 AUROC and Brier results in Table 4, this shows that small precision differences at 10% can coexist with detectable whole-cohort differences.

Review-list membership also changed. Compared with the full-information list, gender/disability removal replaced between 17 and 61 registrations in each direction, depending on the observation day. Six-field removal replaced 290 registrations at day 14 and 169 at day 42. Similar aggregate yield therefore did not mean that the same registrations were selected for review.

Table 5. Gender/disability removal versus full information at 10% capacity; family-wise intervals adjust across six days only.

Day	Reviewed	Targeted TP	Targeted P	Full TP	Full P	Difference	Pointwise 95%	Family-Wise 95%
14	935	709	0.7583	708	0.7572	+0.0011	[−0.0107, 0.0128]	[−0.0149, 0.0171]
28	923	763	0.8267	764	0.8277	−0.0011	[−0.0108, 0.0141]	[−0.0173, 0.0195]
42	908	765	0.8425	769	0.8469	−0.0044	[−0.0110, 0.0099]	[−0.0143, 0.0134]
56	890	803	0.9022	801	0.9000	+0.0022	[−0.0067, 0.0124]	[−0.0101, 0.0165]
70	873	813	0.9313	815	0.9336	−0.0023	[−0.0092, 0.0069]	[−0.0115, 0.0092]
90	855	804	0.9404	801	0.9368	+0.0035	[−0.0082, 0.0070]	[−0.0117, 0.0093]

3.1.3. Single-Field, Common-Cohort, and Clickstream-Free Analyses

The largest single-field loss occurred when highest education was removed at day 14. The precision difference was −0.0342 (95% interval [−0.0544, −0.0085]). By day 42, all six single-field differences lay between −0.0066 and +0.0077. This pattern suggests that prior educational information may be most informative before richer evidence of current participation becomes available; it does not establish a causal effect of that field on the outcome.

Following the same registrations across observation days did not establish a consistent improvement in gender/disability removal relative to full information. Six-field removal had more favorable point estimates at later stages, but the family-wise intervals included zero. These conditional-cohort results do not establish that later observation reduces the removal gap or improves the opportunity for intervention.

Adding richer assessment-process information improved the clickstream-free fallback, with AP increasing by 0.0256, AUROC by 0.0315, and the Brier score decreasing by 0.0070 relative to the sparse fallback. Its 10% precision increased by only 0.0055, and the interval included zero. Table 6 reports the point estimates.

Table 6. Day-42 results for models that do not use clickstream information.

Policy	Model Size	Relevant Cases	Precision@10%	AP
Six-field-excluded standard	7	762	0.8392	0.6798
Strict original	31	669	0.7368	0.5883
Strict enriched due-process	7	674	0.7423	0.6139
Strict typed alternative	7	670	0.7379	0.5945

Model size is the selected maximum number of terminal nodes per tree.

Against the standard six-field removal model, the enriched fallback achieved precision of 0.7423 rather than 0.8392. The precision difference was −9.69 percentage points (95% interval [−12.89, −7.38] percentage points), exceeding the five-percentage-point operational loss criterion. The enriched fallback therefore improved on the sparse fallback but did not meet the stated adequacy criterion.

3.1.4. Residual Attribute Predictability

OULAD diagnostic gender AUROC ranged from 0.690 to 0.812 across observation days, while disability AUROC ranged from 0.663 to 0.680 (Table 7). Residual predictability remained when evaluation was limited to students not observed during diagnostic-model fitting. Retained records therefore contained information associated with the deleted attributes, but this does not show that the early-warning model used a particular proxy.

Table 7. Held-out predictability of directly excluded attributes. Ranges cover the evaluated stages; AP lift is average precision divided by target prevalence.

Dataset	Target	Stages	Positive Cases	AUROC Range	AP-Lift Range
OULAD	disability	6	761–868	0.663–0.680	1.52–1.68
OULAD	gender	6	4905–5425	0.690–0.812	1.31–1.43
UCI 320 Math.	sex	3	187	0.746–0.750	1.50–1.50
UCI 320 Port.	sex	3	266	0.734–0.734	1.63–1.64
UCI 697	Educational special needs	2	40	0.637–0.655	1.77–1.88
UCI 697	Gender	2	1249	0.771–0.774	1.89–1.90

3.1.5. Calibration Sensitivity

Raw and archived Platt probabilities selected identical OULAD lists in all 128 model–capacity combinations. Isotonic regression introduced ties and changed list membership; some stages and capacities with adjusted precision losses also changed (Supplementary Table S18). Sex/gender-only and gender/disability removal remained inconclusive under all three methods. A separate Monte Carlo sensitivity affected the borderline six-field-removal interval at day 28 and 30%: it included zero when unchanged Platt predictions were evaluated using the common resampling stream for the calibration comparison. This difference is not a calibration effect; the original primary estimates are retained. The enriched fallback still failed the five-percentage-point adequacy criterion under each method.

3.2. UCI 697: Higher-Education Dropout

For the primary Dropout-versus-Graduate outcome, the sparse operational model had lower precision than the full-information model at enrollment at all four review capacities, with deficits ranging from 34.92 percentage points at 5% to 17.97 points at 30%. After semester one, adjusted differences remained detectable only at the larger 20% and 30% capacities. Socioeconomic/family removal showed the same pattern: losses remained across all four enrollment capacities and at 20% and 30% after semester one. Sex/gender-only removal produced no precision difference whose family-wise interval excluded zero (Tables 2 and 3; Figure 2).

At the supporting fixed-10% operating point, the sparse admissions model had a 32.61-percentage-point precision deficit at enrollment (Table 8; Figure 4). This corresponds to approximately 33 fewer observed dropout cases per 100 reviewed records in the retrospective terminal-status sample. After semester one, the deficit narrowed to 1.35 percentage points, and the two-stage family-wise interval reached zero. This comparison does not estimate how an intervention would change dropout.

Table 8. UCI 697 review-list precision for the primary terminal-status outcome at 10% capacity; family-wise intervals adjust across two stages only.

Stage	Full Information	Sparse Operational	Difference (Points)	Family-Wise 95% CI
Enrollment	0.9111	0.5849	−32.61	[−39.25, −27.06]
After semester 1	0.9865	0.9730	−1.35	[−4.04, 0.00]

The type of information removed mattered. At enrollment, removing family and financial fields reduced 10% precision by 21.83 percentage points. Removing seven direct personal fields changed it by only −0.54 points, and removing gender plus special-needs status changed it by +0.81 points; both intervals included zero. Small whole-cohort ranking differences remained detectable, and after semester one some global metrics still favored full information. Stronger current academic evidence narrowed some predictive gaps but did not make every reduced model interchangeable.

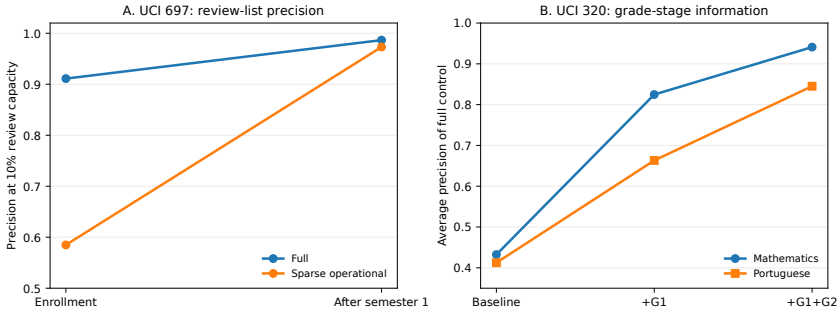


Figure 4. How later academic evidence changed prediction in two distinct datasets. (**A**) shows UCI 697 review-list precision for full-information and sparse operational models. (**B**) shows UCI 320 average precision as G1 and G2 become available; Mathematics and Portuguese are analyzed separately. The panels are not pooled estimates or direct replications.

These primary results are conditional on a resolved outcome because students recorded as Enrolled were excluded. Such selection can change the sample composition and the estimated predictive gaps; their direction and magnitude in the complete entering cohort are not established here. In the secondary Dropout-versus-all-other analysis, socioeconomic/family removal yielded seven of eight stage–capacity intervals below zero (Supplementary Table S16). This provides a comparison under an alternative outcome definition, but it does not recover the eventual outcomes of students still enrolled or replace prospective cohort evaluation.

Diagnostic gender AUROC reached 0.771–0.774. Educational special-needs prediction was less stable: only 40 positive cases were available, and AP was approximately 0.020 at a prevalence of 0.011 (Table 7). These attribute-prediction results are separate from the dropout-model comparisons.

Both raw and isotonic probabilities changed pooled out-of-fold review-list membership relative to Platt. Raw probabilities retained the primary family-wise precision classifications. With isotonic regression, socioeconomic/family-removal losses remained across all four enrollment capacities for both outcome definitions, but one semester-one loss per outcome became inconclusive: at 20% for Dropout versus Graduate and 10% for Dropout versus all other statuses. Sex/gender-only removal remained inconclusive. These results qualify the specific affected operating points rather than establish calibration-method equivalence (Supplementary Table S18).

3.3. UCI 320: Secondary-School Performance

Mathematics and Portuguese were analyzed separately. None of the review-list precision comparisons had a family-wise interval excluding zero after adjustment across stages and capacities, including the sex/gender-only and socioeconomic/family removal blocks (Table 3). The estimates therefore did not establish a consistent precision effect of background-field removal in either subject; they do not demonstrate equivalence.

Full-information prediction improved substantially after grades became available (Figure 4). In Mathematics, AP increased from 0.432 before grades to 0.825 with G1 and 0.941 with G1 and G2, while 10% precision increased from 0.575 to 0.925 and 0.975, respectively. In Portuguese, AP increased from 0.412 to 0.663 and 0.845, and precision increased from 0.409 to 0.591 and 0.727.

For background-field removal, nearly all pointwise and family-wise intervals included zero for ranking, probability accuracy, and review-list precision. The only exception was a small Mathematics period-one AUROC improvement after removing sex. Given the small samples and the many settings examined, this isolated secondary result does not support a general policy conclusion. Separately, diagnostic sex AUROC was approximately 0.75 in Mathematics and 0.73 in Portuguese (Table 7).

Alternative calibration changed the pooled out-of-fold review lists, but all family-wise precision intervals still included zero for both subjects under raw, Platt, and isotonic probabilities. This remains an imprecise result, not evidence of equivalent policies or calibrators (Supplementary Table S18).

3.4. Cross-Dataset Summary

Across the 90 fixed gradient-boosting configurations, no calibration method had the lowest Brier score in every setting. Supplementary Table S17 reports full-control Brier scores and log loss, with illustrative reliability curves in Figure S1 and complete configuration-level results in the accompanying aggregate files. No raw-versus-Platt or isotonic-versus-Platt change in the policy-minus-full precision gap had a family-wise interval excluding zero. Thus, changes in the classification of individual policy contrasts do not by themselves demonstrate differences between calibration methods. These comparisons do not select a preferred method using evaluation outcomes.

Tables 2 and 3 summarize the adjusted precision findings from the original Platt-calibrated primary analysis; detailed numerical versions are retained in Supplementary Tables S15 and S16. Detectable losses depended on the dataset, information group, educational stage, and review fraction. Across the sex/gender-only block, none of the 56 stage–capacity evaluations from 14 stage-specific model pairs had a precision interval excluding zero. This does not establish equivalence: some secondary global metrics differed, and review-list membership was not always identical.

In the original Platt-calibrated primary analysis, across 296 planned stage–capacity results, 30 family-wise intervals were below zero, and none was above zero. Some rows are duplicate reporting views of the same paired comparison, so these counts are descriptive rather than independent confirmations. Figures 2 and 4 juxtapose stage and capacity profiles while retaining the separate datasets, outcomes, and field definitions.

Every primary diagnostic model had a held-out AUROC point estimate above 0.5 (Table 7; Figure 5). This residual predictability identifies associations in retained records. It does not identify their source or establish actual proxy use, discrimination, or privacy leakage by the early-warning models.

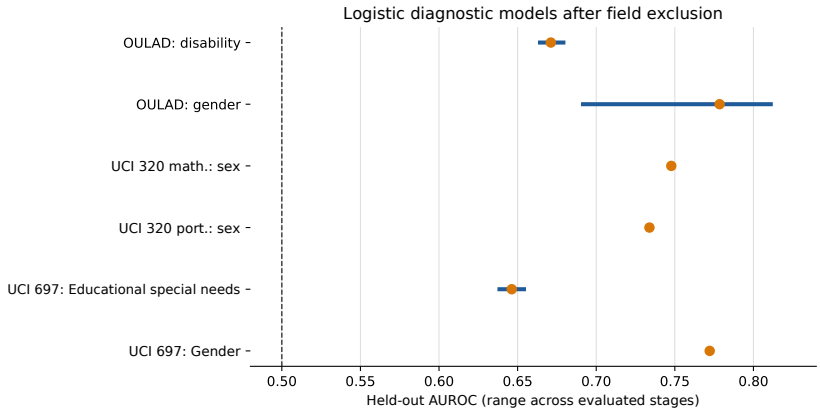


Figure 5. Held-out AUROC ranges after direct attribute removal. Each segment spans the evaluated stages, and the point is their mean; the dashed line represents chance discrimination. Residual predictability does not show that the early-warning model used a specific proxy.

4. Discussion

4.1. Principal Interpretation

The contribution is a capacity-aware retrospective evaluation framework for data-minimized early-warning systems. A conclusion about the cost of removal must specify the

fields, educational stage, review capacity, and performance measure. The OULAD results illustrate why one operating point is insufficient: the small six-field removal difference at day 42 and 10% did not describe the detectable losses in several 30% lists. Gender/disability removal showed no family-wise precision loss, whereas socioeconomic/family removal affected whole-cohort performance and selected larger lists. These findings concern predictive performance under the evaluated conditions, not causal effects on educational practice.

Data minimization can have practical value even where some predictive loss occurs. If omitted fields are also withheld from the operational data flow, fewer sensitive attributes need to be accessed, transferred, or stored for prediction. This may simplify data governance and reduce direct exposure of those attributes. These potential benefits were not quantified here, and removing model inputs alone does not change institutional collection or retention of the original records. Residual attribute predictability further limits any privacy inference from deletion (Hasan & Fritz, 2022). The predictive changes therefore inform a broader decision involving educators and students (Rets et al., 2023); they do not establish the overall balance of benefits and costs of a removal policy.

The UCI analyses show how this interpretation depends on the educational setting. Family/financial removal and sparse admissions data weakened enrollment-time prioritization in UCI 697, with smaller gaps after semester one. UCI 320 showed better prediction after grades became available but no stable review-list removal effect. Later evidence may improve prediction while leaving less time for support; the present data do not determine that trade-off or an optimal intervention stage. The balance should be evaluated in the intended institution with educators.

Review-list precision, whole-cohort performance, and list membership answer different questions. Precision measures the yield within a fixed review capacity; AP and AUROC describe ranking across the cohort, while the Brier score measures probability accuracy. Similar list precision can therefore coexist with poorer overall prediction or different students being selected. For example, OULAD gender/disability removal at day 42 changed the number of observed adverse outcomes in the list by only four, but replaced 37 registrations in each direction. Institutional evaluation should examine both aggregate performance and who receives attention.

Residual attribute predictability adds a distinct governance lesson. Gender, disability, and special-needs information remained partly recoverable from other educational records even after direct removal. This is consistent with overlapping information and correlated social or behavioral processes, but it does not establish that the outcome model used a specific proxy. Direct deletion alone neither removes statistical association nor demonstrates fairness (Gardner et al., 2019; Pham et al., 2025). A stronger fairness claim would require subgroup allocation and performance analyses, a model-specific investigation, and an explicit normative standard.

The clickstream-free results illustrate the same need for context. Assessment-process evidence improved a sparse fallback but did not recover the practical yield of the standard comparator. Institutions may still need a fallback when platform telemetry is unavailable or should not be used, but the system should then disclose reduced confidence, abstain where appropriate, and avoid presenting a degraded list as operationally equivalent.

4.2. Why Provenance Matters for Educational Decisions

The provenance correction is not merely a software detail. A different representation of missing area-deprivation values changed how the full-information and partially minimized models interpreted student records. An implicit rule for tied scores also changed two selections at a review-list boundary. In a teacher-facing system, such changes can determine which individual students appear for review. Reproducibility must therefore cover source

data, missing-value semantics, information timing, list construction, and software versions, not only the final model.

4.3. Implications for Educator-Facing Digital Technology

Module- or course-specific lists define a local allocation of review slots. Their precision, false-alert counts, membership changes, and fallback performance describe retrospective selection performance, not actual staff workload. Reviewing equally many records can take different amounts of time because cases differ in complexity, information completeness, contact requirements, and follow-up. Repeated OULAD registrations may also generate several reviews for one student. Translating a quota into staffing requirements would require observations of review time, repeat contacts, and intervention activity in the intended institution. Large-scale teacher-facing implementations and human-centered dashboard research nevertheless demonstrate that workflow, explanation, and professional interpretation must be studied with educators (Alfredo et al., 2024; Herodotou et al., 2019). Data-driven decision making should treat platform evidence as one input to contextual professional judgment rather than as an automated instruction (Evenstein Sigalov et al., 2026).

These findings motivate a possible institutional evaluation procedure rather than a validated deployment recommendation. An institution could define the educational decision, the responsible professional, the stage at which support is possible, and a review quota informed by measured workload. Input policies could then be compared using list precision, false-alert counts, probability accuracy, membership, and subgroup allocation at the intended stages. Procurement and deployment decisions would also require separate consideration of data governance, explanations, professional override, missing-data handling, and prospective student outcomes. The present analyses do not establish the effectiveness of this proposed procedure.

This proposed procedure connects predictive evaluation to educators' professional activity. Whether it reduces workload, improves decisions, or benefits students requires prospective institutional validation with educators and students.

Figure 6 separates the retrospective findings from proposed institutional checks. These checks are suggestions for prospective evaluation, not demonstrated effects on practice.

4.4. Limitations

The analysis was developed iteratively and is not pre-registered confirmatory research. Its bootstrap intervals are conditional on the fitted models, selected hyperparameters, and fitted calibration maps; they do not account for model-selection, calibration-method, institutional-transport, or development-process uncertainty. For a fixed OULAD model, a single strictly increasing calibration map preserves ranking and capacity-based selection, apart from ties introduced by clipping or numerical precision; Brier scores can nevertheless change. This invariance cannot be assumed for pooled UCI out-of-fold predictions because separate calibrators are fitted within different folds and may change ordering between folds. Isotonic calibration can additionally change selection through tied probabilities. The additional raw/Platt/isotonic comparison assesses only these fixed models and calibration samples; it does not identify a universally preferable calibrator or include uncertainty from refitting calibration maps. Beta calibration and other alternatives were not evaluated. The primary model family was gradient boosting, with logistic regression used only as a sensitivity in the UCI analyses. Family-wise intervals cover the pre-specified stages and capacities within each named policy and metric, not every analysis conducted during the broader project.

The OULAD outcome combines Fail and Withdrawn, which may require different responses and occur on different timelines. Later risk sets exclude earlier withdrawals. The common-cohort sensitivity holds membership fixed by retaining registrations present at later

stages, but consequently selects students who continue: it excludes early withdrawals and cannot recover the timing effect for the original entry cohort. Thus, neither analysis separates information accumulation from all consequences of cohort selection. In UCI 697, excluding Enrolled students conditions the primary analysis on resolved status. Selection can change the outcome prevalence, predictor distribution, and difficulty of the classification task; consequently, the reported precision deficits cannot be transferred directly to all entrants. We did not estimate the direction or size of this selection bias. The secondary Dropout-versus-all-other analysis includes Enrolled students but uses a different outcome definition and is not prospective follow-up or a correction for censoring. Incomplete timestamps further limit reconstruction of when results became visible. These analyses do not identify optimal intervention timing or distinct intervention pathways for failure and withdrawal.

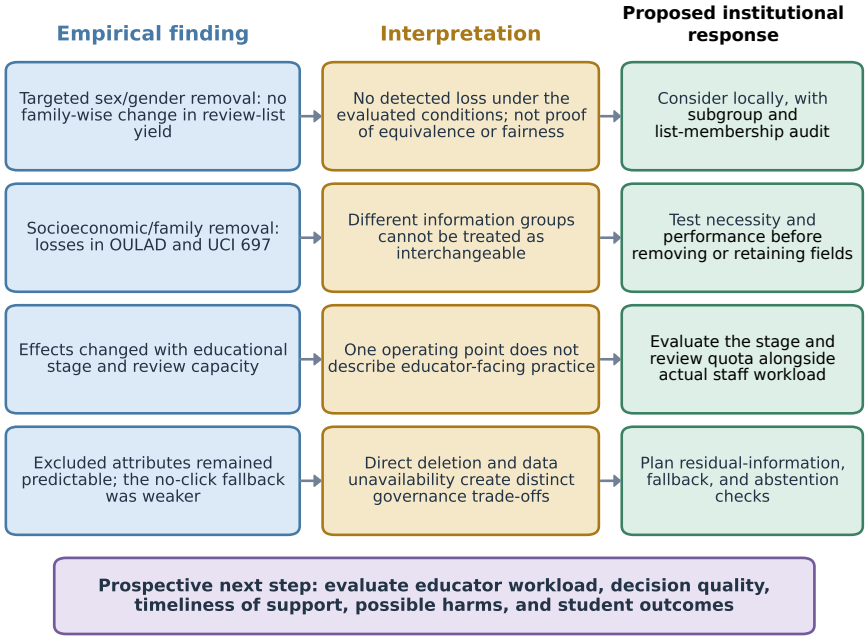


Figure 6. Retrospective findings, their interpretation, and proposed institutional checks. The proposed responses require prospective evaluation; the arrows organize the discussion and do not represent causal effects or validated deployment instructions.

The datasets differ in educational level, population, outcome definition, observation timing, sample size, and validation design. OULAD uses a later presentation for temporal evaluation, whereas UCI uses within-dataset cross-fitting; the latter does not establish performance in a later cohort or another institution. UCI 320 has small subject-specific samples that could not be uniquely linked or pooled. The socioeconomic/family blocks also contain different fields: prior education and area deprivation in OULAD, parental and current financial indicators in UCI 697, and parental education and occupation in UCI 320. Their labels do not imply measurement equivalence. Similar directions are therefore descriptive observations across distinct settings, not a common effect estimate or independent institutional replication. Context-specific evaluation is needed before deployment.

Residual predictability is descriptive and does not establish actual proxy use, privacy leakage, or unfair allocation by the early-warning models. The special-needs diagnostic in UCI 697 is particularly uncertain because only 40 positive cases were available. We did not conduct a complete subgroup fairness audit, privacy attack, or prospective evaluation of educator workload, intervention uptake, possible harms, or student benefit. Those analyses require an institutionally defined use case and participation by the educators and students affected.

Finally, the official XuetangX schema contained no demographic fields, and a preserved derivative could not be independently connected to that schema. It was therefore

excluded from all demographic evidence rather than used as unverifiable support. Detailed provenance exclusions, superseded results, and reproducibility records remain available in the Supplementary Materials.

5. Conclusions

This retrospective study found that the predictive consequences of removing student-background fields depended on the information group, educational stage, and review capacity. Sex/gender-only removal showed no family-wise difference in review-list precision, without establishing equivalence. Socioeconomic/family removal reduced whole-cohort predictive performance in OULAD and UCI 697 and affected selected review lists; UCI 320 estimates remained inconclusive. Excluded attributes remained partly predictable from retained records. The calibration sensitivity changed some stage–capacity interval classifications, while sex/gender-only comparisons remained inconclusive. The findings are conditional on the evaluated populations, outcome definitions, fitted models, and calibration procedures. They do not quantify privacy benefits, actual staff workload, or the effects of interventions on students.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/educsci16111816/s1: Table S1: Computational evidence and provenance ledger; Table S2: Review-list precision families across stages and capacities; Table S3: Comparable information-group contrasts at a 10% review capacity; Table S4: Complete residual-attribute predictability results; Table S5: OULAD main-policy point estimates; Table S6: UCI review-capacity contrasts; Table S7: UCI review-capacity contrasts; Table S8: Single-field removal results; Table S9: Common-cohort timing sensitivity; Table S10: Review-list overlap; Table S11: Provenance-transition records; Table S12: Provenance-transition records; Table S13: Production and verification inventory; Table S14: Exact information-removal policies and dataset-specific fields; Table S15: Detailed precision contrasts across four review capacities; Table S16: Detailed contrasts for comparable information groups; Table S17: Full-control probability scores under alternative calibration methods; Table S18: Calibration sensitivity of adjusted precision findings; Figure S1: Illustrative reliability curves; accompanying aggregate results and code for the calibration sensitivity analysis. Supplementary Materials S1 is provided as a separate MDPI supplementary-file document and is numbered independently from the main text.

Author Contributions: Conceptualization, O.K.; methodology, O.K.; software, O.K., K.B. and K.I.; validation, O.K., K.B. and K.I.; formal analysis, A.S.; investigation, G.S.; resources, K.B. and K.I.; data curation, K.B. and K.I.; writing—original draft preparation, O.K.; writing—review and editing, D.T.; visualization, D.T.; supervision, A.S.; project administration, G.S.; funding acquisition, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number AP23489228.

Institutional Review Board Statement: Not applicable. This study used only publicly accessible, de-identified secondary datasets (OULAD; UCI "Predict Students' Dropout and Academic Success"; UCI "Student Performance"), with no recruitment, contact, intervention, re-identification attempt, or access to non-public records.

Informed Consent Statement: Not applicable. This study used only pre-existing, publicly available, de-identified secondary data; no human subjects were recruited or contacted by the authors.

Data Availability Statement: OULAD is available from the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/349/open+university+learning+analytics+dataset (accessed on 15 October 2026) and is described by Kuzilek et al. (2017). The official archive used here had SHA-256 `f2ed1902616c1fe8d2824d872c0b7d2d72be435bf0124d077044fe4be2c6d3e4`. UCI 697 is available at https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success (accessed on 15 October 2026); the outer ZIP used here had SHA-256 `e90e55fd65ec462ae283ebeb2cca409319e3460ed898d8754f62fb35cc83a65d`. UCI 320 is available at https://archive

.ics.uci.edu/dataset/320/student+performance (accessed on 15 October 2026); the outer ZIP had SHA-256 `82ae9d66437b9808df42e8c89d2bb179c46e9cfbcf06f38abc1d20b3b747e177`. The current nested `student.zip`, not `.student.zip_old`, was analyzed. All three UCI pages identify the datasets as CC BY 4.0. The aggregate-only reproducibility release accompanying this submission is available at https://github.com/KuznetsovKarazin/SecureEWS (accessed on 15 October 2026); it contains analysis code, the locked protocol, machine-readable aggregate results and bootstrap draws, verification reports, and manuscript sources. Code and aggregate outputs for the additional calibration sensitivity analysis are included with the Supplementary Materials. Raw or processed student-level data, individual predictions, fitted model bundles, private clean-room archives, and XuetangX derivative files are not redistributed.

Acknowledgments: During the preparation of this manuscript, the authors used OpenAI ChatGPT (GPT-6 Astra Ultra) and Codex (exact version not recorded) to assist with code development, verification workflows, visualization, and language and drafting support. The authors reviewed and validated the code, analyses, references, figures, and final text, revised the outputs as necessary, and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Practical Interpretation Guide

Table A1 maps the principal evaluation quantities to questions faced by educators and educational institutions. The quantities should be interpreted together; no single metric establishes readiness for deployment. Institutional readiness also requires evidence about workload, equity, and professional judgment.

Table A1. Plain-language interpretation of the evaluation quantities.

Quantity	Plain-Language Meaning	Institutional Question
Review capacity	The specified fraction of eligible records selected for review	How does the selected quota relate to measured staff time?
Precision at capacity b	Among selected records, the proportion with the later adverse outcome	What fraction of selected records has the adverse outcome?
Recall at capacity b	The proportion of all records with a later adverse outcome included in the list	How many subsequent cases are missed because the list is limited?
AP and AUROC	How well the model ranks records across the full range of possible thresholds	Does the model retain useful ordering beyond the chosen review boundary?
Brier score	How closely predicted probabilities correspond to observed outcomes	How accurate are the predicted probabilities overall?
List overlap	The extent to which two models select the same records for review	Does data minimization change who is reviewed even when aggregate yield is similar?

References

Alfredo, R. D., Echeverria, V., Zhao, L., Lawrence, L., Fan, J. X., Yan, L., Li, X., Swiecki, Z., Gasevic, D., & Martinez-Maldonado, R. (2024). Designing a human-centred learning analytics dashboard in-use. *Journal of Learning Analytics*, *11*(3), 62–81. [CrossRef]

Chiu, H. S., Wong, A., & Wong, T. L. (2026). A privacy-conscious AI framework for early identification of at-risk students across disciplines using LMS engagement data. *Education Sciences*, *16*(7), 1046. [CrossRef]

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. [CrossRef]

Cortez, P. (2008). *Student performance*. UCI Machine Learning Repository. [CrossRef]

Drachsler, H., & Greller, W. (2016). Privacy and learning analytics: It's a DELICATE issue. A checklist for trusted learning analytics. In *Proceedings of the sixth international conference on learning analytics and knowledge* (pp. 89–98). Association for Computing Machinery. [CrossRef]

European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016. *Official Journal of the European Union*, *L 119*, 1–88.

Evenstein Sigalov, S., Hershkovitz, A., Cohen, A., & Nachmias, R. (2026). Trusting the data: An updated framework for teachers' data-driven decision-making (DDDM) in higher education. *Education and Information Technologies*, *31*, 953–981. [CrossRef]

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the ninth international conference on learning analytics and knowledge* (pp. 225–234). Association for Computing Machinery. [CrossRef]

Hasan, R., & Fritz, M. (2022). Understanding utility and privacy of demographic data in education technology by causal analysis and adversarial-censoring. *Proceedings on Privacy Enhancing Technologies*, *2022*(2), 245–262. [CrossRef]

Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z., & Hlosta, M. (2019). A large-scale implementation of predictive learning analytics in higher education: The teachers' role and perspective. *Educational Technology Research and Development*, *67*, 1273–1306. [CrossRef]

Kull, M., Silva Filho, T., & Flach, P. (2017). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th international conference on artificial intelligence and statistics* (Vol. 54, pp. 623–631). PMLR. Available online: https://proceedings.mlr.press/v54/kull17a.html (accessed on 8 October 2026).

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific Data*, *4*, 170171. [CrossRef]

Le, N. L., Abel, M.-H., & Laforge, B. (2026). When can we trust early warnings? Leakage-Excluded early outcome prediction from LMS interaction logs. *arXiv*. [CrossRef]

Moons, K. G. M., Damen, J. A. A., Kaul, T., Hooft, L., Andaur Navarro, C. L., Dhiman, P., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., & Collins, G. S. (2025). PROBAST+AI: An updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, *388*, e082505. [CrossRef]

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on machine learning* (pp. 625–632). Association for Computing Machinery. [CrossRef]

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pham, N., Pham Ngoc, H., & Nguyen-Duc, A. (2025). Fairness for machine learning software in education: A systematic mapping study. *Journal of Systems and Software*, *219*, 112244. [CrossRef]

Plak, S., Cornelisz, I., Meeter, M., & van Klaveren, C. (2022). Early warning systems for more effective student counselling in higher education: Evidence from a dutch field experiment. *Higher Education Quarterly*, *76*(1), 131–152. [CrossRef]

Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). *Predict students' dropout and academic success.*. UCI Machine Learning Repository. [CrossRef]

Rets, I., Herodotou, C., & Gillespie, A. (2023). Six practical recommendations enabling ethical use of predictive learning analytics in distance education. *Journal of Learning Analytics*, *10*(1), 149–167. [CrossRef]

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*(3), e0118432. [CrossRef]

Shaikhanova, A., Kuznetsov, O., Iklassova, K., Tokkuliyeva, A., & Sugurova, L. (2025). Interpretable predictive modeling for educational equity: A workload-aware decision support system for early identification of at-risk students. *Big Data and Cognitive Computing*, *9*(11), 297. [CrossRef]

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The achilles heel of predictive analytics. *BMC Medicine*, *17*, 230. [CrossRef]

Wu, S., Duan, J., & Luo, M. (2026). Calibrated early-warning models with fairness auditing and selective prediction for course withdrawal risk: Evidence from OULAD. *PLoS ONE*, *21*(7), e0352867. [CrossRef]

Wu, S., & Luo, M. (2026). *Reproducibility code for evaluating repeated course-withdrawal early-warning systems in distance higher education*. Zenodo. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.