Serican Journal of Medicine
2026, VOL. 3, NO. 2, 25575
https://doi.org/10.17161/sjm.v3i2.25575

Limited fitness of routinely captured student participation data for predicting graduation examination performance in undergraduate medical education

Chujie Chen[1], Zhen Zhang[2] and Peng Yun[2]*

Departments of [1]Urology and [2]Endocrinology, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen 518000, China

***Corresponding author**: Peng Yun, MD/PhD, Department of Endocrinology, The Seventh Affiliated Hospital, Sun Yat-sen University, 628 Zhenyuan Road, Guangming District, Shenzhen, China. Tel: +86-0755-81206795; Email: yunpeng@sysush.com

ABSTRACT

Introduction: Clinical-teaching platforms now routinely capture student-level participation records, and learning analytics promises to turn these into predictions of educational outcomes. Yet the completeness, linkability, and predictive validity of such data are seldom audited, especially in new programmes where capture is still maturing.

Subjects and Methods: We conducted a single-centre secondary analysis at a new tertiary teaching hospital. Teaching-participation records were linked by name, within enrolment cohort, to graduation examination scores (theory, skill, total) for three cohorts of five-year clinical-medicine interns (2022-2024). We assessed completeness, linkability, and the predictive validity of overall, domain-matched (theory- vs. skill-oriented), and rotation-based exposure, using Spearman correlations with Fisher confidence intervals, Benjamini-Hochberg correction, and cohort-fixed-effects regressions on within-cohort standardized scores.

Results: The platform held 3,216 activities and 16,391 participation records, yet only 61 students linked individually (5.5% of the 1,101-student roster), rising steeply. No association survived multiplicity correction (0 of 15). Domain-matched associations were weak, with confidence intervals spanning zero (theory exposure vs. theory $\rho = 0.00$; skill vs. skill $\rho = 0.14$). Rotation duration was unrelated to performance; rotation breadth correlated with theory score ($\rho = 0.39$) and persisted after adjustment for participation volume, a single hypothesis-generating signal. Given the small, 2024-dominated sample, the nulls exclude only moderate-to-large associations, not small effects.

Conclusions: Routinely captured teaching-participation data showed limited completeness, linkability, and predictive validity for graduation outcomes and are not yet fit for student-level prediction. Institutions considering such use should audit linkability and predictive validity before relying on participation dashboards for high-stakes inference.

ARTICLE HISTORY
Received: June 2, 2026
Revised: June 10, 2026
Accepted: June 11, 2026

KEYWORDS
learning analytics, data quality, medical education, clinical clerkship, educational measurement

Introduction

Clinical-teaching management platforms increasingly record granular, student-level traces of educational activity (attendance at ward rounds and lectures, completion of procedural-skill sessions, submission of reflective notes), and the field of learning analytics has framed these accumulating traces as a resource for understanding and improving educational outcomes [1], while cautioning that analytics must stay anchored in evidence about learning rather than in data availability alone [2]. One commonly used case is that what is routinely captured can be turned into actionable signals: dashboards that flag disengaged learners, predictive models that anticipate examination failure, or evidence that particular forms of teaching exposure drive competence. Yet engagement traces from learning-management systems have predicted achievement only inconsistently across courses [3].

Any such inference, however, inherits the quality of the data beneath it. In the neighbouring domain of electronic health records, the recognition that routinely collected data are generated for operational rather than analytic purposes prompted a formal vo-

cabulary for data-quality assessment: completeness, correctness, currency, plausibility, and, above all, fitness for a specific use [4]. Educational platforms collect data under similar operational pressures (compliance, attendance accounting, accreditation), and there is no guarantee that metrics optimized for those purposes are fit for predicting learning. Attendance and participation count, in particular, record presence rather than learning, and may bear little relation to what a student has actually mastered [5-6].

Despite the enthusiasm for educational big data, the upstream questions are seldom asked: can routinely captured teaching-participation records actually be linked, at the level of the individual student, to summative competency outcomes; and once linked, do they predict those outcomes? These are validity questions, and a contemporary validity argument requires evidence that a proposed score interpretation and use are defensible before the score is acted upon [7-8]. They are especially pressing in newly established teaching programmes, where data capture is still ramping up, and an apparently rich dataset may in fact be sparse and selective at the level that matters.

We therefore audited, in a single newly established tertiary teaching hospital, the fitness-for-use of its routinely captured teaching-participation data. We asked, for three graduating cohorts of five-year clinical-medicine interns, (i) how complete and linkable the captured participation data were against graduation comprehensive examination records, and (ii) whether overall, domain-matched, and rotation-based exposure metrics predicted theory, skill, and total graduation performance.

Subjects and Methods

Study design and setting

This was a retrospective secondary analysis of routinely collected administrative and educational data at a newly established tertiary teaching hospital in China. The analysis was framed as a data-quality (fitness-for-use) audit of the institution's clinical-teaching management platform, drawing on the data-quality dimensions developed for routinely collected health records [4]. The study population comprised five-year clinical-medicine interns who sat the graduation comprehensive examination in 2022, 2023, or 2024. Other affiliated hospitals and other programmes were out of scope.

Data sources

Two routinely maintained data sources were used. The first was the teaching module of the institutional clinical-teaching management platform, from which official student-level participation exports were obtained; each record described a teaching activity (with an activity identifier, activity type, responsible department, and timestamp) together with the participating student, an attendance sign-in flag, and an indicator of whether a reflective note had been submitted. The second was the graduation comprehensive examination record for each cohort, providing a theory score, a skill score, and a composite total score per student.

Record linkage and de-identification

Because the examination records and the platform records shared no common numeric key, students were linked deterministically by name within the cohort of enrolment. For the five-year programme, the cohort of enrolment plus five years defines the graduation (and examination) year; the enrolment field was internally consistent (of 429 five-year clinical students, none spanned more than one enrolment year). Linkage was first restricted to the relevant graduation year and then matched on name within that year's examination roster; students whose names were ambiguous within a year on either side were dropped. Of the 1,101 students on the examination rosters, 62 five-year clinical interns were identifiable in the platform for their graduation year, and 61 of these were matched to an examination record (2, 14, and 45 in the 2022, 2023, and 2024 cohorts); the single unmatched 2023 student gives that cohort a platform linkage rate of 0.93 (Table 1). Linkage was performed in memory; only de-identified, pseudonymized, student-level aggregates were written to disk, with no names, platform identifiers, identity-document numbers, or contact details retained. No demographic attributes (such as age or sex) were retained in the de-identified extract; the representativeness of the linked sample was therefore assessed against examination performance rather than demographic composition (see Results and Supplementary Table S4).

Exposure measures

We derived three families of exposure metrics from the captured participation records (one record per

Table 1. Student-level data capture and record linkage by graduation cohort.

Graduation cohort	Platform-linkable clinical students	Examination-roster students	Linked students	Linkage rate vs. platform	Linked / roster (%)
2022	2	372	2	1.00	0.5
2023	15	370	14	0.93	3.8
2024	45	359	45	1.00	12.5
Pooled	—	1,101	61	—	5.5

Platform-linkable clinical students = five-year clinical interns identifiable in the platform for that graduation year; linked students = those further matched to an examination record. The 2023 difference (15 vs. 14) reflects one platform student who could not be matched to an examination record (platform linkage rate 0.93). "Linked / roster (%)" expresses linked students as a percentage of that cohort's examination roster.

student-activity attendance; the attendance sign-in flag and the reflective-note indicator are recorded on the same record, but were treated as separate metrics rather than as part of the activity count). Overall participation comprised the number of distinct activities attended, the number of distinct activity types, the number of distinct departments covered, the attendance sign-in rate, and the number of reflections submitted (a count of the reflective notes a student submitted on the platform, which records submission rather than the depth or quality of reflection, neither of which the platform captures). Domain-matched exposure classified each activity by its platform-assigned type into theory-oriented (didactic lectures and case discussions), skill-oriented (procedural-skill sessions), and bedside/mixed (teaching ward rounds, and night ward rounds where present) categories; collective lesson-preparation activities (faculty-facing) were excluded. Rotation measures comprised total internship length (the interval between roster-recorded start and end dates), two activity-timestamp proxies for rotation duration (each student's overall first-to-last activity span and the mean within-department activity span), and rotation breadth (the number of distinct departments in which a student attended activities). The rotation-duration proxies are acknowledged surrogates, as no authoritative department-by-department rotation schedule was available, and attendance/participation counts index presence rather than instructional quality or dose. The platform-assigned activity-type labels were fixed before any examination outcome was examined, with the analyst blind to scores at classification; the full taxonomy (platform activity types, their domain assignment, and per-type activity counts within the analytic cohort) is given in Table 4.

Outcomes and standardization

The outcomes were the theory, skill, and total graduation examination scores. Only these component-level scores were available; the examination blueprint, station-level skill scores, examiner identifiers, and reliability indices were not retained in the records and could not be analyzed. Because the 2022 skill component was recorded on a different scale from later cohorts, all scores and count-based exposures were standardized within graduating cohort (z-scores) before pooling, so that associations reflect within-cohort rank rather than cross-cohort scale differences.

Statistical analysis

Exposure-outcome associations were quantified with Spearman correlations on within-cohort standardized scores, with 95% confidence intervals from the Fisher z-transformation. For the overall family (5 exposures × 3 outcomes = 15 tests) and the rotation family (12 tests), p-values were controlled with the Benjamini–Hochberg false-discovery-rate procedure. For the domain-matched analysis, we compared matched pairs (theory-oriented exposure with theory score; skill-oriented exposure with skill score) against cross-domain placebo pairs, and fitted ordinary-least-squares regressions of each standardized examination component on the raw domain counts with cohort fixed effects and heteroscedasticity-consistent (HC3) standard errors. A multivariable model of standardized total score on the overall exposure metrics with cohort fixed effects was also fitted. To test whether the association between rotation breadth and theory score was independent of overall participation, the standardized theory score was additionally regressed on rotation breadth, the total activity count, and cohort fixed effects (with HC3 standard errors), complemented by a partial Spearman correlation of breadth with theory score controlling for the total activity count. No prediction model was developed or validated; "predictive validity" here refers

Table 2. Distribution of captured exposure metrics and examination outcomes for the 61 linked students.

Variable	Mean	SD	Min	Median	Max
Attended activities (count)	34.6	10.7	1	36	54
Activity types (count)	4.2	0.7	1	4	5
Departments covered (count)	5.7	1.1	0	6	7
Sign-in rate	0.97	0.06	0.67	1.00	1.00
Reflections submitted (count)	24.5	13.7	0	26	47
Theory score	19.4	2.5	14.0	19.6	24.4
Skill score	62.8	38.0	46.4	56.6	267.5
Total score	75.5	6.0	62.4	76.3	86.8

to whether routinely captured metrics carried a reproducible, outcome-linked signal. The Fisher z confidence intervals are approximate, given ties in the count-based exposures and the near-ceiling sign-in rate. Because the 2022 cohort contributed only two linked students, its fixed-effect coefficient is not interpreted, and the regression is treated as auxiliary. Sensitivity analyses reclassified case discussions to the bedside/mixed category and excluded the small 2022 cohort, the latter applied to the overall, domain-matched, and rotation correlations and to the regression. We additionally compared the linked students with the unlinked remainder of each examination roster, using within-cohort standardized mean differences (SMD) and Mann–Whitney tests, to gauge whether the captured subset was representative of examination performance. Given the linked sample size, the study had approximately 80% power to detect only moderate correlations ($|\rho| \approx 0.35$); a null result therefore indicates the absence of a detectable association rather than proof of no effect. Analyses used Python 3.13 (pandas, SciPy, statsmodels) and R 4.5.1 (ggplot2). The study is reported in line with the RECORD extension of the STROBE statement for studies using routinely collected data [9]; the completed checklist is provided in Supplementary Appendix S1.

Ethics

The study was approved by the institutional ethics committee (KY-2026-111-01) and used only de-identified, routinely collected records; the requirement for individual informed consent was waived.

Results

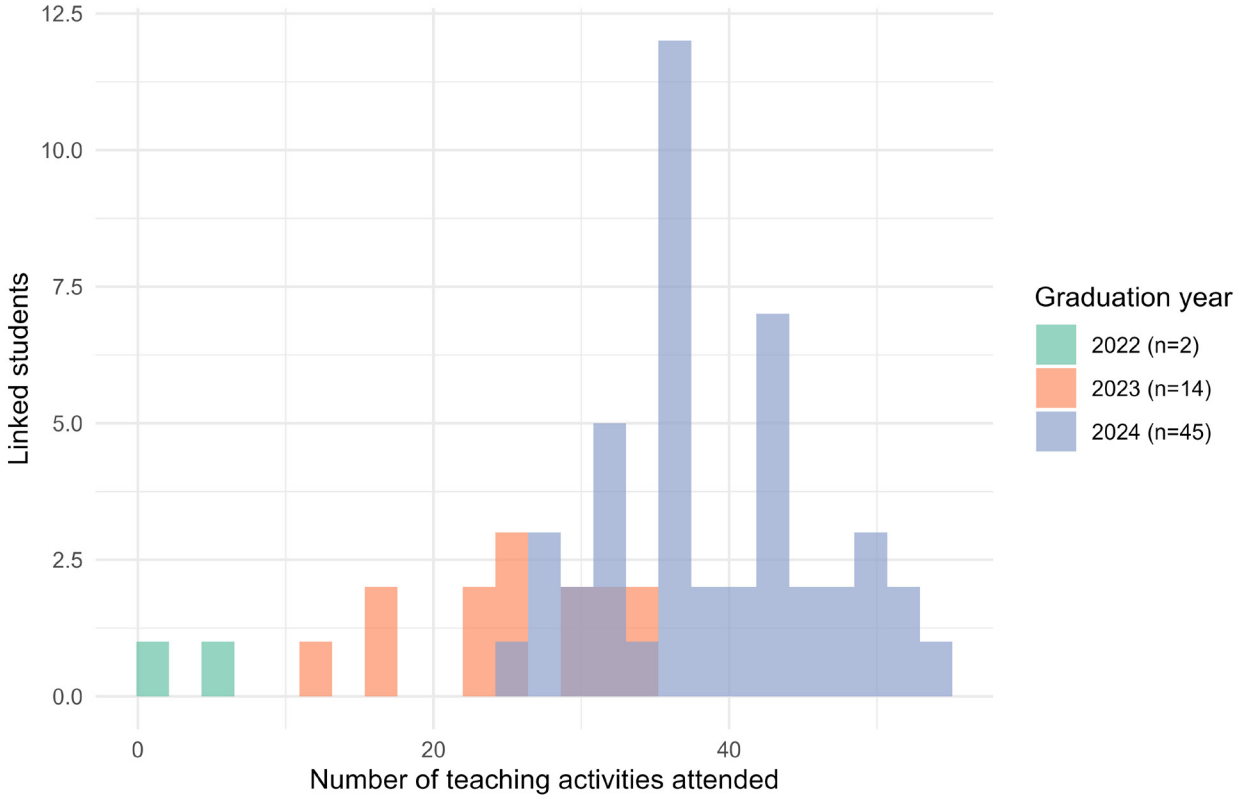
Data capture and record linkage

Over the study window, the teaching module recorded 3,216 distinct teaching activities comprising 16,391 student-level participation records. Despite this volume, the proportion of graduating clinical interns whose participation could be linked to an individual graduation examination record was small and strongly time-dependent. Among five-year clinical-medicine students, the number of platform-registered students who could be linked to the corresponding cohort's examination roster rose across successive graduating cohorts, from 2 in 2022 to 14 in 2023 and 45 in 2024, yielding a pooled analytic sample of 61 students (Table 1). The graduation rosters themselves were essentially complete (372, 370, and 359 unique students for 2022, 2023, and 2024; 1,101 in total), so the bottleneck was student-level capture and identifiability within the teaching platform rather than missing examination data. Expressed against the examination roster, only 61 of 1,101 graduating students (5.5%) were both captured and linkable, ranging from 0.5% (2 of 372) in 2022 to 3.8% (14 of 370) in 2023 and 12.5% (45 of 359) in 2024. Once a student appeared in the platform with an attributable identity, linkage to the examination record succeeded for almost all of them (linkage rate versus the platform 1.00, 0.93, and 1.00 across the three cohorts). The pooled sample was therefore dominated by the 2024 cohort (45 of 61), reflecting progressive uptake of the platform after its introduction rather than any change in the underlying student population. The captured students were broadly representative of examination performance: within-cohort standardised differences from the unlinked remainder of each roster were small for theory and total scores in the two well-populated cohorts ($|\text{SMD}| \leq 0.26$; pooled Mann-Whitney $p = 0.16$ and 0.91), the only notable difference being a modestly higher skill score among linked 2024 students (SMD = 0.38, $p = 0.02$; Supplementary Table S4).

Distribution of captured exposure metrics and examination outcomes

For the 61 linked students, the routinely captured participation metrics varied widely in their usable spread (Table 2). The count of attended activities was well dispersed (Figure 1), as was the number of submitted reflections. Other captured metrics were concentrated near a ceiling or otherwise compressed:

Figure 1. Distribution of the number of attended teaching activities among the 61 linked students, by graduation cohort.



the attendance sign-in rate was near-uniform, the number of distinct activity types was tightly clustered (most students 4-5), and the number of covered departments was similarly narrow (most students 5-7). Examination scores were standardized within the graduating cohort before pooling because the 2022 skill component used a different scale (raw skill-score range up to 267.5 in 2022 versus a tighter spread thereafter); theory and total scores were on a stable scale across cohorts.

Overall exposure–outcome associations

Across all five captured exposure metrics and the three examination outcomes (15 pairs), no association survived correction for multiple comparisons: 0 of 15 Spearman correlations were significant after Benjamini-Hochberg control (within-cohort standardized scores; full matrix in Supplementary Table S1). The strongest single association was between the number of covered departments and theory score ($\rho = 0.23$, uncorrected $p = 0.080$, adjusted $p = 0.73$); the most-attended single metric, the count of attended activities, was essentially unrelated to skill score ($\rho = 0.06$, $p = 0.62$; Figure 2). A multivariable model of standardized total score on the captured exposure metrics with cohort fixed effects explained almost no variance ($R^2 = 0.026$), and every exposure coefficient was non-significant. Excluding the small 2022 cohort left this pattern unchanged (0 of 15 associations significant after correction, $n = 59$; auxiliary regression $R^2 = 0.048$; Supplementary Table S3).

Domain-matched exposure–outcome associations

Because an omnibus null could conflate conceptually distinct activities, we examined domain-matched dose-response relationships, aligning each exposure with the examination component in the same competency domain (Table 3; Figure 3). The matched associations were weak, and their 95% confidence intervals included zero: theory-oriented exposure versus theory score $\rho = 0.00$ (95%CI: 0.25 to 0.25) and skill-oriented exposure versus skill score $\rho = 0.14$ (95% CI: 0.12 to 0.38). The analysis did not provide evidence that the matched pairs were stronger than the cross-domain placebo pairs (theory exposure versus skill score $\rho = 0.07$; skill exposure versus theory

Figure 2. Number of attended teaching activities versus skill examination score (within-cohort standardized) among the 61 linked students, with a single pooled linear fit (Spearman ρ = 0.06, p = 0.62); points are coloured by graduation cohort, but no per-cohort line is fitted given the small 2022 subgroup (n = 2).

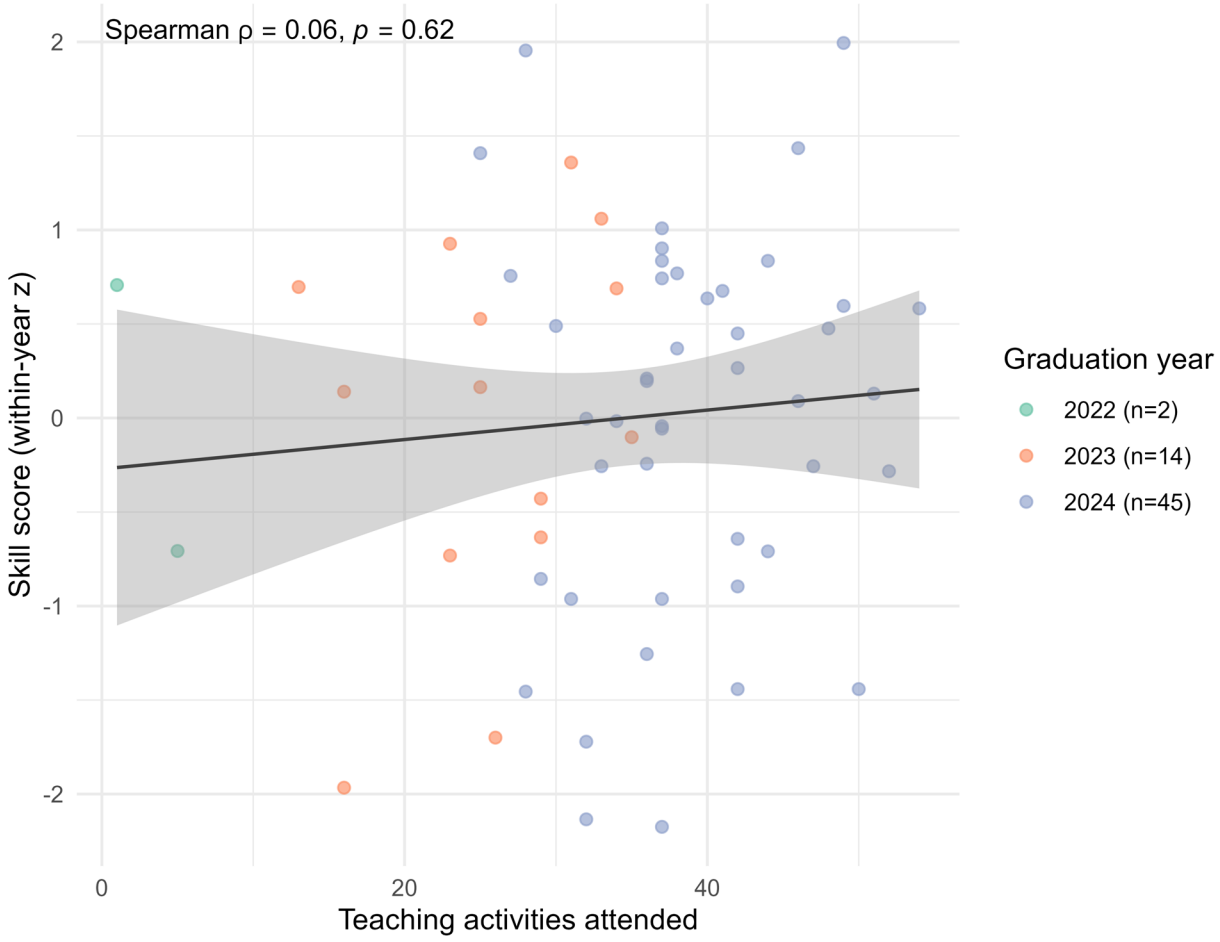
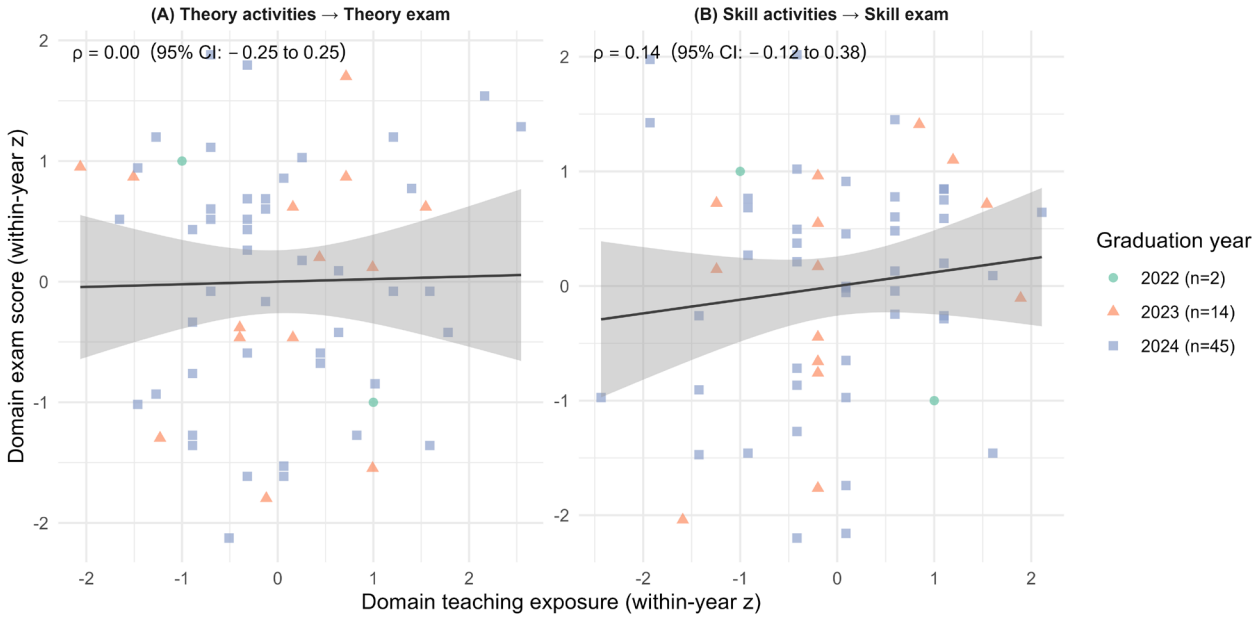


Table 3. Associations between captured exposure metrics and examination outcomes (Spearman ρ on within-cohort standardized scores; 95% CI by Fisher transformation).

Exposure metric	Outcome	ρ	95% CI	p
Domain-matched				
Theory-oriented activities	Theory score	0.00	−0.25 to 0.25	0.99
Skill-oriented activities	Skill score	0.14	−0.12 to 0.38	0.29
Theory-oriented activities (placebo)	Skill score	0.07	−0.18 to 0.32	0.59
Skill-oriented activities (placebo)	Theory score	0.05	−0.21 to 0.29	0.73
Rotation duration/breadth				
Total internship length (days)	Theory score	−0.01	−0.33 to 0.32	0.98
Activity-span (days)	Theory score	0.17	−0.09 to 0.41	0.20
Mean within-department span (days)	Theory score	0.06	−0.20 to 0.31	0.65
Departments rotated (breadth)	Theory score	0.39	0.16 to 0.59	0.002
Departments rotated (breadth)	Total score	0.23	−0.03 to 0.45	0.081

Within-cohort standardized scores were used because the 2022 skill component was on a different scale. The departments-rotated–theory association (Benjamini–Hochberg q = 0.021 within the 12-test rotation family) persisted after adjustment for overall participation volume (volume-adjusted regression coefficient 0.54, p < 0.001; partial Spearman ρ = 0.44; see text and Supplementary Table S5) but is reported as hypothesis-generating given the small observational sample.

Figure 3. Domain-matched exposure–outcome scatter (within-cohort standardized scores): (A) theory-oriented activity exposure versus theory examination score; (B) skill-oriented activity exposure versus skill examination score. The flat fits illustrate the absence of a detectable domain-matched dose–response relationship.



score ρ = 0.05). Domain-matched regressions of each examination component on the raw domain counts with cohort fixed effects explained little variance (theory model R^2 = 0.022; skill model R^2 = 0.034), with non-significant matched-exposure coefficients. Results were unchanged when case discussions were reclassified to the bedside/mixed category (theory-oriented exposure versus theory score ρ = −0.03) and when the small 2022 cohort was excluded (theory ρ = 0.03; skill ρ = 0.17; Supplementary Table S3).

Rotation duration and breadth

Three rotation-duration measures of differing quality were examined, together with rotation breadth (Table 3; the full rotation-family correlations are in Supplementary Table S2). Total internship length from the platform roster was available for 37 of 61 students (36 contributing to the within-cohort correlation) and was near-constant (median 335 days, IQR 332-359), consistent with a standard 11- to 12-month programme; its variance was further distorted by isolated implausible values. Two activity-timestamp proxies (each student's overall first-to-last activity span and the mean within-department activity span) were noisier surrogates that systematically understate true rotation length. None of the three duration measures was associated with any examination outcome (all 95% CI included zero; for example, total internship length versus theory score ρ = -0.01). The only association to cross the conventional significance threshold involved rotation breadth (the number of departments in which a student attended activities) and theory score (ρ = 0.39, 95% CI: 0.16 to 0.59, p = 0.002; Benjamini-Hochberg-adjusted q = 0.021 within the 12-test rotation family). Rotation breadth was itself correlated with overall participation volume (ρ = 0.41 with the total activity count) and overlapped the omnibus department-count metric (ρ = 0.23 with theory score). Contrary to the expectation that this breadth signal was merely an engagement artefact, however, it was not explained by participation volume: in a regression of standardised theory score on rotation breadth, the total activity count, and cohort fixed effects, the breadth coefficient remained significant and was, if anything, marginally larger than with breadth alone (0.54, 95% CI: 0.23 to 0.84, p < 0.001; partial Spearman ρ = 0.44, p < 0.001; Supplementary Table S5), because the total activity count was correlated with breadth yet unrelated to theory score and so acted as a mild suppressor. The association also persisted, and strengthened slightly, when the 2022 cohort was excluded (ρ = 0.43, p < 0.001). The breadth-theory signal is therefore not reducible to sheer participation volume; we nonetheless treat

Table 4. Activity-type taxonomy for the analytic cohort: platform activity types, domain assignment, and distinct-activity and participation-record counts.

Platform activity type	Domain	Distinct activities	Participation records	Excluded
Didactic lecture (小讲课)	Theory-oriented	407	814	No
Procedural-skill session (技能操作)	Skill-oriented	221	717	No
Teaching ward round (教学查房)	Bedside/mixed	196	388	No
Case discussion (病例讨论)	Theory-oriented	111	225	No
Collective lesson preparation (集体备课)	Excluded (faculty-facing)	8	17	Yes
Total		**943**	**2,161**	

Counts cover the distinct teaching activities attended by the analytic cohort (clinical-medicine interns of the three graduating cohorts), which is a subset of the platform-wide catalogue (3,216 activities). The classification was specified before examination outcomes were examined. Case discussions were assigned to the theory-oriented domain in the primary analysis and to bedside/mixed in a sensitivity analysis.

it as hypothesis-generating rather than as a dose-response effect, because rotation breadth may proxy unmeasured student attributes (such as ability or differences in rotation assignment), the linked sample is small and 2024-dominated, and the analysis is observational.

Discussion

In a newly established teaching hospital with an apparently rich teaching-participation dataset, we found that the data were, for the purpose of predicting graduation performance, of limited fitness for use on three counts. First, completeness and linkability at the individual-student level were low: of more than sixteen thousand participation records across more than three thousand activities, only 61 graduating students could be linked to their examination outcome, and that number climbed steeply with platform uptake rather than reflecting the true denominator of graduating interns. Second, among the students who could be linked, none of the routinely captured exposure metrics predicted examination performance after accounting for multiple comparisons. Third, sharpening the analysis (matching exposure to outcome by domain, and probing rotation duration) recovered only one candidate signal: rotation breadth crossed the significance threshold and persisted after adjustment for overall participation volume, but it rested on a small, single-cohort-dominated sample and so remained hypothesis-generating.

These findings are best read through the lens of data quality rather than pedagogy. Routinely captured educational data, like routinely captured clinical data, are generated for operational ends, and the dimensions developed to appraise health-record data (completeness, currency, plausibility, and fitness for a specific use) map onto what we observed, though the analogy is conceptual rather than domain-identical [4]. The capture gap and its year-on-year ramp are a completeness-and-currency problem: a platform introduced partway through these cohorts' training necessarily under-records the earlier ones, so a pooled dataset is dominated by the most recent, best-captured cohort. The isolated, implausible internship-length values are a plausibility problem. And the central result, that captured participation does not track summative competency, is a fitness-for-use problem: metrics built around attendance and compliance index presence, not learning [5], and attendance itself has been shown to relate only weakly and inconsistently to assessment performance in clinical rotations [6]. Their near-ceiling distributions (sign-in rate, activity-type and department counts) further leave little variance with which to discriminate between students, even if an underlying relationship existed.

This caution sits awkwardly beside the prevailing enthusiasm for educational big data and analytics [1], but it is the same caution that the health-informatics community arrived at once routinely collected data were put to analytic use [4]. It is also a validity argument: using a captured metric to infer something about a learner, and still more to act on that inference, requires evidence that the inference is defensible, and that evidence must be assembled, not assumed [7-8]. Our data provide no such evidence for these participation metrics in this setting; if anything, the domain-matched analysis, in which exposure and outcome were aligned to the same competency domain [10], makes the absence of a dose-response relationship harder to attribute to crude measurement alone. The lone breadth-theory association is an instructive case: it is consistent within its family and, when tested directly, was not explained by overall participation volume, yet it remains substantively

uncertain: rotation breadth may stand in for unmeasured student attributes rather than reflect a causal effect of rotational diversity, so reading it as an established dose-response effect would still outrun the evidence from this small, single-cohort-dominated, observational sample, even though breadth is the one captured metric that behaves like a genuine outcome-linked signal and would merit prospective validation.

The practical implication is not that teaching participation is unimportant, but that routinely captured participation data should be audited for linkability and predictive validity before they are used for high-stakes purposes such as remediation flags or predictive risk scores. In a maturing programme, an analytics dashboard built on this data would, on our evidence, risk reflecting how completely each cohort happened to be captured rather than a learning signal. Institutions investing in learning analytics infrastructure may gain more from improving the completeness and linkage of capture (and from validating a small number of outcome-linked metrics) than from expanding the catalogue of what is recorded. Concretely, capture and linkage could be strengthened by issuing a persistent student identifier shared across the teaching platform and the examination system (which would remove the reliance on name-based matching that constrained the present analysis), by making at-source capture mandatory and auditing linkage rates for each cohort, and by prospectively validating the few metrics that show an outcome-linked signal, such as rotation breadth, before they inform any dashboard or high-stakes decision.

This study has important limitations. The linked sample was small (N = 61) and dominated by one cohort, so the analysis was powered to detect only moderate associations, and the confidence intervals remained wide enough to span small-to-moderate effects in either direction; the null results therefore indicate the absence of a detectable relationship in this captured subset, not proof that teaching exposure is unrelated to performance. The data come from a single centre and a single platform; the linkable subset was smaller than the full graduating class and, although broadly representative of examination performance, not fully so: linked 2024 students scored modestly higher on the skill component (SMD = 0.38). This selective capture was, however, modest and confined to a single cohort and a single domain: differences from the unlinked remainder were small and non-significant for theory and total scores ($|SMD| \leq 0.26$), so the platform did not simply register a uniformly top-performing subset. Even so, because the captured students were not a random sample (clustering near the ceiling of attendance and, in 2024, slightly higher on skill), the usable range on several metrics was restricted, and range restriction of this kind tends to attenuate any underlying exposure-outcome association; our nulls therefore indicate the absence of a detectable association within this captured subset and should not be generalised to the non-captured majority. A more mature platform with higher and less cohort-skewed capture, or a multi-centre design, might achieve denser linkage and a different balance of signal and noise, so these results should not be read as a ceiling on what learning analytics can offer once capture matures. Attendance and participation counts capture presence rather than teaching quality or true instructional dose; the reflection count records whether notes were submitted, not their depth or quality; activity-type labels were platform-assigned; and the rotation-duration measures were noisy proxies in the absence of an authoritative rotation schedule. No examination blueprint, station-level scores, examiner information, or reliability indices were available, so a score-linked null cannot be fully separated from outcome measurement error, particularly for the skill component; and residual cross-cohort scale heterogeneity, only partly removed by within-cohort standardization, may also have attenuated associations. Finally, the analysis is observational, and even a clear association would not establish causation. These constraints are themselves part of the finding: they are the concrete ways in which routinely captured teaching-participation data fall short of fitness for outcome prediction in a newly established programme.

Conclusions

A clinical-teaching platform holding thousands of activities and tens of thousands of participation records yielded, for a graduating cohort, only a few dozen students who could be linked to their examination outcome; and among those, no captured participation metric predicted graduation performance, whether assessed overall, matched by competency domain, or by rotation duration. Routinely captured teaching-participation data in this newly established programme were thus of limited completeness, limited linkability, and limited predictive validity. Before such data are used to drive learning analytics dash-

boards or high-stakes decisions, their linkability and predictive validity should be audited, and capture and linkage improved, rather than assumed.

Practice Points

- A clinical-teaching platform may hold thousands of activity records yet link to only a few dozen students at the individual level; capture and link ability, not sheer data volume, determines analytic usefulness.
- Routinely captured participation metrics (attendance, activity counts, reflections) showed no detectable association with graduation examination performance, even when matched to the same competency domain.
- An apparent signal from such data, here rotation breadth, may persist even after adjustment for overall engagement, yet in a small, single-cohort-dominated sample, it should be validated prospectively rather than read as a dose-response effect.
- In a newly established program, participation data are dominated by the most recently and most completely captured cohort, which can masquerade as a substantive trend.
- Linkability and predictive validity should be audited before teaching-participation dashboards are used for high-stakes decisions such as remediation flags.

Declarations

Ethics approval and consent to participate. The study was approved by the Ethics Committee of the institution (approval number KY-2026-111-01) and used only de-identified, routinely collected records; the requirement for individual informed consent was waived.

Consent for publication. Not applicable.

Availability of data and materials. The de-identified, aggregated data and the analysis code supporting the findings are available from the corresponding author on reasonable request, subject to institutional data-governance approval.

Competing interests. The authors declare no competing interests.

Funding. None.

Authors' contributions. Chujie Chen: conceptualization, methodology, software, formal analysis, visualization, and writing – original draft. Zhen Zhang: data curation, validation, writing, review and editing. Peng Yun: supervision, project administration, and writing, review and editing. All authors read and approved the final manuscript.

Acknowledgements. None.

References

1. Ellaway RH, Pusic MV, Galbraith RM, Cameron T. **Developing the role of big data and analytics in health professional education**. *Med Teach*. 2014; 36(3):216-222. doi:10.3109/0142159X.2014.874553.
2. Gašević D, Dawson S, Siemens G. **Let's not forget: learning analytics are about learning**. *TechTrends*. 2015; 59(1):64-71. doi:10.1007/s11528-014-0822-x.
3. Conijn R, Snijders C, Kleingeld A, Matzat U. **Predicting student performance from LMS data: a comparison of 17 blended courses using Moodle LMS**. *IEEE Trans Learn Technol*. 2017; 10(1):17-29. doi:10.1109/TLT.2016.2616312.
4. Weiskopf NG, Weng C. **Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research**. *J Am Med Inform Assoc*. 2013; 20(1):144-151. doi:10.1136/amiajnl-2011-000681.
5. Watling CJ, Ginsburg S. **Assessment, feedback and the alchemy of learning**. *Med Educ*. 2019; 53(1):76-85. doi:10.1111/medu.13645.
6. Deane RP, Murphy DJ. **Student attendance and academic performance in undergraduate obstetrics/gynecology clinical rotations**. *JAMA*. 2013; 310(21):2282-2288. doi:10.1001/jama.2013.282228.
7. Cook DA, Brydges R, Ginsburg S, Hatala R. **A contemporary approach to validity arguments: a practical guide to Kane's framework**. *Med Educ*. 2015; 49(6):560-575. doi:10.1111/medu.12678.
8. Kane MT. **Validating the interpretations and uses of test scores**. *J Educ Meas*. 2013; 50(1):1-73. doi:10.1111/jedm.12000.

9. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. **The Reporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement**. *PLoS Med*. 2015; 12(10):e1001885. doi:10.1371/journal.pmed.1001885.
10. Miller GE. **The assessment of clinical skills/competence/performance**. *Acad Med*. 1990; 65(9 Suppl):S63-S67. doi:10.1097/00001888-199009000-00045.