Research Square

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

Percentage gain versus normalized gain for identifying training targets in medical record documentation education: a pre–post study

Chao Wang
The Affiliated Hospital of Southwest Jiaotong University & The Third People's Hospital of Chengdu
Qingxia Yang
Aba Prefecture Forestry Central Hospital
Ziyi Chen
Aba Prefecture Forestry Central Hospital
Guoyong Deng
Aba Prefecture Forestry Central Hospital
Zhihui Yu
Aba Prefecture Forestry Central Hospital
Yidong Liu
lyd2023@139.com
Aba Prefecture Forestry Central Hospital

Research Article

Keywords: Medical records, Clinical documentation, Continuing medical education, Program evaluation, Educational measurement, Normalized gain, Ceiling effect, Quality improvement

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11233839/v1

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

Additional Declarations: No competing interests reported.

Percentage gain versus normalized gain for identifying training targets in medical record documentation education: a pre–post study

Chao Wang[1,2], Qingxia Yang[3], Ziyi Chen[2], Guoyong Deng[2], Zhihui Yu[2], Yidong Liu[3*]

[1] International Medical Center of Geriatrics, The Affiliated Hospital of Southwest Jiaotong University & The Third People's Hospital of Chengdu, Chengdu 610031, Sichuan, China

[2] Department of Medical Record Management, Aba Prefecture Forestry Central Hospital, Dujiangyan 611830, Sichuan, China

[3] Department of Medical Affairs, Aba Prefecture Forestry Central Hospital, Dujiangyan 611830, Sichuan, China

***Corresponding author:** Yidong Liu, Department of Medical Affairs, Aba Prefecture Forestry Central Hospital, Dujiangyan 611830, Sichuan, China. Email: lyd2023@139.com

Email addresses: CW: wangchao6646203@163.com; QY: 869405420@qq.com; ZC: 1696629842@qq.com; GD: 1813285568@qq.com; ZY: 1107351108@qq.com; YL: lyd2023@139.com

Abstract

Background Record audits are widely used to evaluate documentation training, and improvement is often compared across documentation domains using the percentage change in scores to judge which components worked and where to focus next. When post-training scores approach the instrument ceiling, this metric may mainly reflect baseline room for improvement. We compared percentage score gain with defect burden reduction, a normalized gain, and examined whether the defect profile changed after one training cycle.

Methods We analysed blinded audit data from a self-paced micro-lecture program (53 lectures; 1327 min) for 49 physicians at a secondary-level hospital in China (December 2023 to July 2024). Five discharge records per physician were randomly sampled one month before and after training, de-identified, pooled, and scored on a 20-dimension instrument by four raters blinded to time point and author. Dimensions were ranked by each metric. Agreement between rankings, and between baseline and residual (post-training) defect burden, was assessed with Spearman correlation, with 95% uncertainty intervals (UIs) from 10 000 parametric simulations.

Results The composite score increased from 89.14 ± 2.73 to 96.82 ± 1.38 ($P < 0.001$), and post-training dimension scores converged within 96.04–97.97. Percentage score gain correlated with baseline score at $r = -0.960$ (95% UI −0.98 to −0.90), versus $r = -0.649$ (−0.79 to −0.35) for defect burden reduction. The two rankings were correlated ($r_s = 0.707$, 95% UI 0.52 to 0.90) but not interchangeable: only 2 of the top 5 dimensions overlapped, and 16 of 20 dimensions had rank intervals spanning more than 10 positions. Baseline and residual defect burden showed little agreement ($r_s = 0.144$, 95% UI −0.33 to 0.43); informed consent moved from the lightest baseline burden to the second heaviest residual burden.

Conclusions Near the instrument ceiling, percentage score gain largely reflects baseline room for improvement and should not be used alone to compare training effects across

documentation domains. Normalized gain depends less on baseline, but dimension ranks are imprecise and are better reported with uncertainty intervals and in tiers. Targets for the next training cycle should be set from the residual defect burden re-measured after each cycle.

Keywords: Medical records; Clinical documentation; Continuing medical education; Program evaluation; Educational measurement; Normalized gain; Ceiling effect; Quality improvement

Background

Medical records document the clinical reasoning and care delivered to patients and underpin continuity of care, quality assessment, reimbursement and medicolegal review [1, 2]. Deficient record keeping is common and has been described as an essential but overlooked aspect of the quality of care [3]. In China, national regulations specify the content and format of inpatient records [1], and the national action plan to improve medical quality for 2023–2025 designates the quality of medical record content as a priority and calls for strengthened documentation training [2]. Hospitals therefore run documentation training programs and evaluate them with record audits.

Evaluations of documentation training and related interventions typically report changes in mean audit scores, in the proportion of records graded Class A, or in defect rates before and after training [4, 5]. Because audit instruments are organised into domains that correspond to the components of the record, evaluators often go a step further and compare the size of improvement across domains, both to judge which parts of a curriculum worked best and to decide where the next round of training should focus. This comparison is usually made with the percentage change in domain scores. Two features of documentation audits make this practice problematic. First, domain scores on 100-point instruments usually exceed 85 points even before training, so the instrument operates close to its ceiling [6] and domains

that start higher have less room to improve. Second, percentage change from baseline is mathematically coupled to the baseline value and is a statistically inefficient and potentially misleading outcome [7–9]. Educational measurement has addressed the same problem with the normalized gain, which divides the observed gain by the maximum gain possible given the pre-test score [10, 11], although its own dependence on pre-test scores has been debated [12]. These ideas have rarely been applied to the evaluation of documentation training.

A second, related issue concerns how training targets are chosen. Documentation training is often planned as a one-off event whose content is set by a baseline audit. Whether the domains that most need attention after training are still those that were weakest before training is seldom examined, even though audit and feedback and plan–do–study–act approaches to quality improvement are explicitly iterative [13, 14]. Moreover, most evaluations of documentation-related training remain at the reaction and learning levels of Kirkpatrick's model [15, 16], whereas audits of records written in routine practice capture change at the level of behaviour.

Using blinded audit data from one cycle of a micro-lecture program on medical record documentation, we aimed (1) to compare percentage score gain with defect burden reduction, an application of the normalized gain to audit scores, in ranking improvement across documentation dimensions, and to quantify the uncertainty of those rankings; and (2) to examine whether the profile of documentation defects was reordered after a single training cycle. Because both aims concern alternative summaries of the same data, they do not depend on attributing the observed improvement to the training.

Methods

Study design

This was a single-group pre–post study using audit data generated during the routine evaluation of a hospital documentation training program. It is reported in accordance with the Standards for QUality Improvement Reporting Excellence (SQUIRE 2.0) [17].

Setting and participants

The study was conducted at Aba Prefecture Forestry Central Hospital, a secondary-level (Grade II, Class A) general hospital in Dujiangyan, Sichuan Province, China. All 56 frontline clinical physicians were eligible. Physicians were excluded if they had cumulative leave of more than one month during the training period, had missing audit data, or were transferred out of clinical posts during the study (n = 7). The remaining 49 physicians from seven clinical departments were included. Of these, 28 (57.1%) were women; mean age was 35.55 ± 7.09 years and mean length of clinical practice 10.41 ± 6.67 years. By professional title, 22 (44.9%) were resident physicians (the junior clinical rank in China, not trainees), 23 (46.9%) attending physicians, and 4 (8.2%) associate chief physicians or above.

Educational intervention

From January to June 2024, a series of 53 micro-lectures (15–30 min each; 1327 min in total) was released in batches through the learning module of the hospital's enterprise messaging platform (WeCom). The series was produced by an external organisation and covered the documentation requirements for each component of the inpatient record, completion of the medical record front sheet (the standardised summary page used for hospital statistics and reimbursement), and the legal role of medical records. Physicians studied at their own pace, and individual learning time was recorded by the platform. The content was not tailored to

the hospital's defect profile, and no other dedicated documentation quality improvement activities were undertaken during the training period.

Audit instrument

The audit instrument was derived from the Sichuan Provincial Inpatient Medical Record Quality Evaluation Standard (2023 edition), issued by the Sichuan Provincial Medical Record Quality Control Center for the evaluation of discharged records [18]. Revised from a 2016 trial version in line with national regulations on medical quality and record writing, the standard comprises nine sections totalling 100 points: front sheet, 20; progress notes, 25; admission record, surgery-related records, discharge or death record and informed consent, 10 each; and basic requirements, auxiliary examinations and medical orders, 5 each. In the 2023 revision, the weight of the front sheet was doubled from 10 to 20 points in view of its importance, and that of the admission record was halved from 20 to 10. The standard lists 178 defect items, each with a deduction value specified per item, per place or per occurrence. Twenty-three of these are single-item veto defects (for example, a falsified record, a missing first progress note, or no informed consent form for transfusion) that grade a record directly as Class C, and 29 are Class B defects (for example, a first progress note not completed within 8 h, or no attending physician's round within 48 h of admission), each of which deducts 11 points from the total score, so that a single Class B defect precludes a Class A record. Deductions within a section are capped at the section's points, and records are graded Class A (≥90 points), Class B (75 to <90 points) or Class C (<75 points).

For this study, the deduction items were grouped by record component and content into 20 dimensions (Additional file 1), and the grouping was finalised after two rounds of review by five medical record quality control experts with clinical backgrounds. Raters scored each record item by item, and deductions were accumulated within each dimension: defects specified per occurrence were deducted for every occurrence, other defects were deducted per

item, and Class B defects were counted as 11 points, as specified by the standard. Single-item veto defects were treated as equivalent to a Class C grade, consistent with the standard. Dimensions that did not apply to a record (for example, transfusion or resuscitation records when no transfusion or resuscitation took place) were treated as complete, with no deductions, and items without a dedicated dimension, including those for daily progress notes, surgery-related records and auxiliary examinations, were assessed within the progress-note dimensions. Each dimension was then scored from 0 to 100 as 100 − (accumulated deductions ÷ sum of the deduction values of the dimension's items) × 100. Because dimension scores were normalised to the deductible points of each dimension and averaged with equal weights, the composite score used in this study differs from the standard's weighted total score and was not used to assign record grades.

Record sampling and blinded rating

One month before training (December 2023) and one month after training (July 2024), the Department of Medical Record Management used a random number table to select five records per physician from patients discharged in that month. Patient names, inpatient numbers, bed numbers and physician signatures were removed, and the records from both periods were pooled and renumbered in random order. Four quality control staff who had received uniform training in the instrument scored each record independently without knowledge of the time point or the author. For each physician, the dimension score at each time point was the mean of 20 ratings (5 records × 4 raters), and the composite score was the mean of the 20 dimension scores. The study generated 39 200 dimension-level ratings from 490 records; only the physician-level means were retained, and these formed the unit of analysis.

Change metrics

The metrics compared are defined in Table 1. Percentage score gain was calculated as (post-training score − baseline score) ÷ baseline score × 100%. The defect burden L of a dimension was defined as 100 minus its score, that is, the mean number of points lost per record in that dimension. Defect burden reduction was calculated as $(L_{pre} - L_{post}) \div L_{pre} \times 100\%$. Its denominator is the baseline room for improvement, and the expression is algebraically identical to the normalized gain used in educational measurement [10, 11]. Percentage score gain and defect burden reduction were used to compare relative improvement across dimensions; the residual defect burden, L_{post}, was used to identify targets for the next training cycle.

Statistical analysis

Analyses were performed in Python 3.11 with SciPy. Continuous data are presented as mean ± standard deviation. Scores before and after training were compared with paired t tests, and effect sizes were expressed as Cohen's d_z (mean paired difference divided by the standard deviation of the differences). For the 20 dimension-level comparisons, a Bonferroni-corrected significance level of 0.0025 was used.

Dimensions were ranked in descending order of percentage score gain and of defect burden reduction. Agreement between the two rankings was assessed with Spearman's rank correlation coefficient (r_s) and Kendall's τ, and by counting the dimensions shared by the top five and by the bottom five of each ranking. Agreement between the rankings of baseline and residual defect burden was assessed with r_s. Associations between each metric and the baseline score across dimensions were expressed as Pearson's r.

To quantify uncertainty, we ran 10 000 parametric simulations in which each dimension's baseline and post-training means were drawn from normal distributions centred on the observed means, with standard deviations equal to their standard errors; dimensions

and time points were treated as independent. The 2.5th and 97.5th percentiles of each statistic, including each dimension's rank, formed its 95% uncertainty interval (UI). In a sensitivity analysis, the simulations were repeated assuming a pre–post correlation of 0.109, the observed correlation of the composite score. Between-dimension heterogeneity in residual defect burden was tested with an inverse-variance weighted Q statistic.

For the composite score, we calculated Cronbach's α and back-calculated the mean inter-dimension correlation with the Spearman–Brown formula. The association between the baseline composite score and its change was examined with Oldham's method to account for mathematical coupling and regression to the mean [8, 19]. All tests were two-sided with $\alpha = 0.05$.

Results

Scores before and after training

Recorded learning time exceeded 90% of the total course duration (1327 min) for all 49 physicians. Their composite score increased from 89.14 ± 2.73 before training to 96.82 ± 1.38 after training (mean difference 7.68 points; $t = 18.38$, $P < 0.001$; $d_z = 2.63$). All 20 dimensions improved significantly after Bonferroni correction (all $P < 0.001$). Baseline dimension scores ranged from 85.89 to 92.34, whereas post-training scores converged within 96.04 to 97.97, narrowing the between-dimension range from 6.45 to 1.93 points (Table 2).

Association of the change metrics with baseline score

Percentage score gain ranged from 4.1% to 12.0% across dimensions (coefficient of variation 23.6%). It ranked the dimensions identically to absolute gain ($r_s = 1.000$) and correlated strongly with baseline score ($r = −0.960$, 95% UI −0.98 to −0.90). Defect burden reduction ranged from 49.3% to 80.6% (coefficient of variation 11.0%) and correlated less strongly with baseline score ($r = −0.649$, 95% UI −0.79 to −0.35).

Agreement between improvement rankings

The rankings by the two metrics were correlated (r_s = 0.707, 95% UI 0.52 to 0.90; Kendall's τ = 0.547, $P < 0.001$) but differed substantively. The mean absolute rank shift was 3.50 positions (maximum 8); only 2 of the top 5 dimensions overlapped (95% UI 1 to 4), whereas 4 of the bottom 5 overlapped (Table 2, Fig. 1A). Front sheet completion, for example, ranked first by percentage score gain but eighth by defect burden reduction, whereas first progress note quality and consultation notes ranked first and second by defect burden reduction but seventh and ninth by percentage score gain. Rankings by defect burden reduction were imprecise: 16 of the 20 dimensions had rank intervals spanning more than 10 positions. Only first progress note quality (rank interval 1–10) was consistently in the upper part of the ranking, and only informed consent (15–20) and medical orders (11–20) were consistently in the lower part.

Baseline versus residual defect burden

Before training, the heaviest defect burdens were in front sheet completion, completeness of the past medical history, and consistency between the chief complaint and the history of present illness. After training, the heaviest residual burdens were in chief complaint quality, informed consent and front sheet completion. Informed consent moved from rank 20 (the lightest burden; UI 17–20) at baseline to rank 2 (UI 1–15) in residual burden, and medical orders from rank 19 (UI 14–20) to rank 4 (UI 1–17); the probabilities that these two dimensions were among the 10 heaviest residual burdens were 89% and 81%, respectively. Baseline and residual defect burden rankings showed little agreement (r_s = 0.144, 95% UI −0.33 to 0.43), and r_s exceeded 0.5 in only 1.0% of simulations (Table 2, Fig. 1B). Residual defect burden differed across dimensions (Q = 37.70, P = 0.006), although the between-dimension standard deviation (0.53 points) was only slightly larger than the mean standard

error (0.48 points). Uncertainty intervals obtained under an assumed pre–post correlation of 0.109 were essentially unchanged.

Composite score properties and individual-level change

Cronbach's α for the composite score was 0.65 before and 0.73 after training, corresponding to mean inter-dimension correlations of only 0.085 and 0.119. The composite effect size (d_z = 2.63) was much larger than the mean effect size of the individual dimensions (d_z = 0.96). At the physician level, the baseline composite score correlated with its change at $r = -0.883$; after Oldham's correction the correlation was −0.596 ($P < 0.001$). The correlation between physicians' composite scores before and after training was 0.109 ($P = 0.454$).

Discussion

In this evaluation of a documentation training program, post-training audit scores converged close to the ceiling of the instrument. Under these conditions, percentage score gain almost entirely reflected the room for improvement that each dimension had at baseline, and it identified a different set of most-improved dimensions from the normalized gain. Dimension ranks under either metric were imprecise. Moreover, the profile of residual defects after training bore little resemblance to the baseline profile, implying that training targets identified from a baseline audit may be outdated after a single cycle.

Percentage score gain mainly reflects baseline room for improvement

Rankings by percentage score gain were identical to those by absolute gain. Because baseline dimension scores varied only between 85.89 and 92.34, dividing by the baseline score barely changed their relative order, and the ranking was determined by the numerator. As post-training scores converged to 96–98 points in every dimension, the absolute gain approximately equalled this common post-training level minus the baseline score, which explains the correlation of −0.960 between percentage score gain and baseline score. Part of

this association is mathematical coupling, which is expected whenever change is related to its initial value [7, 8]. The practical point is that, in a real audit, the metric most commonly used to compare domains carried little information beyond the baseline score. Defect burden reduction places the baseline room for improvement in the denominator and weakened, but did not remove, the dependence on baseline ($r = -0.649$). Because residual burden varied only from 2.03 to 3.96 points, dimensions with a light baseline burden had lower reduction rates even when their residual burden was similar to that of other dimensions. This is consistent with methodological critiques of percentage change from baseline [7] and with evidence that the normalized gain is not independent of pre-test scores [11, 12]. We therefore recommend that evaluations report absolute change and defect burden reduction together rather than percentage score gain alone.

Improvement rankings depend on the metric and are imprecise

Front sheet completion ranked first by percentage score gain mainly because it had the lowest baseline score and hence the largest room for improvement; by defect burden reduction it ranked eighth. Conversely, first progress note quality and consultation notes ranked first and second by defect burden reduction but only seventh and ninth by percentage score gain. Judging the components of a curriculum by percentage score gain would therefore systematically favour the domains that started worst. Even with defect burden reduction, most dimension ranks spanned more than ten positions: with 49 physicians and five records per physician per period, the data could distinguish only the few dimensions with clearly higher or lower improvement. This mirrors the well-documented imprecision of league tables of institutional performance [20, 21]. When reporting domain-level effects of training, we recommend presenting absolute change and defect burden reduction with uncertainty intervals and grouping dimensions into tiers rather than ranking them one by one.

The residual defect profile differs from the baseline profile

The rank correlation between baseline and residual defect burden was only 0.144, and, allowing for sampling error, the weakest domains after training largely did not coincide with those before training. The heaviest baseline defects—omissions on the front sheet, incomplete past medical history and format errors—are mostly rule-based formal defects that uniform instruction can correct systematically. By contrast, the heaviest residual burdens—chief complaint quality, informed consent and medical orders—depend on the clinical synthesis of symptoms, communication with patients, coordination across departments and review by senior physicians. These defects do not necessarily arise from ignorance of the rules, and video lectures alone may do little to change them, in line with evidence that didactic continuing education has limited effects on complex professional behaviour [22, 23]. A similar pattern was reported for multi-round documentation training in a Chinese tertiary hospital, where errors in patient information declined across rounds while defects involving clinical content remained prominent [4]. The provincial standard also gives these residual domains particular weight: consent signed by an unauthorised person, missing consent for special examinations or treatments, a missing critical-illness notice and verbal orders not recorded after resuscitation are all Class B defects, any one of which precludes a Class A record, and the absence of a transfusion consent form or a long-term order sheet is a single-item veto [18]. Residual defects in these domains therefore matter more for record grading than their residual burden in points suggests.

Although the exact order of residual ranks near the ceiling is unreliable, the upward movement of informed consent and medical orders was consistent across simulations. If the next training cycle were designed from the baseline defect profile, resources would be directed at problems that had already been largely resolved. Consistent with the iterative logic of audit and feedback and of plan–do–study–act cycles [13, 14], training targets should be re-

derived from the residual defect burden after every cycle, and the format of training may need to match the type of defect; for defects that depend on clinical reasoning and communication, interactive formats such as case-based discussion and feedback on physicians' own records are plausible candidates.

Interpreting the composite effect size and individual scores

The low mean inter-dimension correlations indicate that the dimensions jointly constitute record quality rather than reflecting a common latent trait. The composite is therefore a formative index [24], for which internal consistency coefficients have little theoretical justification [25]. Averaging weakly correlated dimensions cancels much of their independent error variance, which explains why the composite effect size ($d_z = 2.63$) was far larger than that of a typical single dimension (0.96). Such composite effect sizes should not be compared directly with those derived from unidimensional scales or pooled with them in meta-analyses of educational interventions [26]. At the individual level, the negative association between baseline score and change persisted after Oldham's correction, but ceiling compression and regression to the mean can both produce such convergence [8, 27, 28]. The pre–post correlation of only 0.109 further suggests that a composite based on five records per period contains substantial measurement error for individual physicians. Judgements about individual competence based on such audits should therefore be made with caution, and educational and quality decisions are better made at the level of dimensions and departments.

Implications for practice

For educators and quality managers who evaluate documentation training with record audits, our findings suggest a simple sequence that requires no data beyond a routine audit: report absolute change together with defect burden reduction and their uncertainty intervals; group dimensions into tiers rather than publishing a league table; re-audit after each cycle and set the next targets from the residual defect burden, giving priority to residual defects that the

audit standard classifies as Class B or single-item veto; and choose training formats that fit the nature of the residual defects.

Strengths and limitations

The strengths of this study include the audit of records written in routine practice, which captures behaviour rather than knowledge; high and uniform completion of the course; the pooling, renumbering and blinded rating of records from both periods by four raters; and the explicit quantification of uncertainty in dimension rankings.

The core comparison of rankings under different metrics uses the same data and therefore holds within a single-group design, but estimates of the training effect itself are limited by the design. With a single-centre, single-group pre–post design, we cannot separate the educational effect from Hawthorne effects, regression to the mean, differences in case mix between December and July, and concurrent management factors [29]. Because dimensions that did not apply to a record were scored as complete, scores for rarely applicable dimensions such as transfusion and resuscitation records partly reflect how often those situations arose in each sampling month. The raters belonged to the department that organised the training, and dates embedded in the records could not be completely removed, so blinding to time point may have been incomplete. Rating-level data linking scores to raters and records were not retained; inter-rater reliability therefore could not be estimated [30], and mixed-effects models with random effects for records and raters could not be fitted. Validity evidence for the audit instrument was limited to expert review of its content [31]. Because all physicians completed more than 90% of the course, the relation between learning exposure and improvement could not be examined. The uncertainty intervals were derived from parametric simulation and relied on normal approximations and independence between dimensions; given mean inter-dimension correlations below 0.12 and essentially unchanged intervals under a pre–post correlation of 0.109, this assumption is unlikely to have materially

affected the conclusions. Finally, the study was conducted in one secondary-level hospital in western China with a single instrument, and the behaviour of defect burden reduction has not been examined in other institutions or instruments. Multicentre studies that retain rating-level data and include concurrent comparison groups are needed.

Conclusions

When post-training audit scores approach the ceiling of the instrument, percentage score gain and absolute gain mainly reflect baseline room for improvement and should not be used alone to compare training effects across documentation domains. Defect burden reduction, a normalized gain, depends less on baseline, but dimension ranks remain imprecise. Evaluations of documentation training should report these metrics together with their uncertainty intervals, and targets for each subsequent training cycle should be set from the residual defect burden measured after the preceding cycle.

Abbreviations

CC: Chief complaint

HPI: History of present illness

SQUIRE: Standards for QUality Improvement Reporting Excellence

UI: Uncertainty interval

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Aba Prefecture Forestry Central Hospital (approval No. 2024012) and was conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from all participating physicians. Records were de-identified before rating, with patient and physician identifiers removed, and the ethics

committee waived the requirement for informed consent from patients because the study used de-identified audit data generated during routine quality management.

Consent for publication

Not applicable.

Availability of data and materials

The de-identified physician-level dimension scores analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests. The micro-lecture series was produced by an external organisation; neither the hospital nor any author was involved in its development or promotion, or received any financial benefit from it.

Funding

This study received no funding.

Authors' contributions

CW designed the study, analysed the data and drafted the manuscript. QY designed and implemented the educational program. ZC, GD and ZY sampled the records, organised the ratings and collected the data. YL supervised the study design, critically revised the manuscript and takes overall responsibility for the work. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Clinical trial number

Not applicable.

References

1. Ministry of Health of the People's Republic of China. Basic norms for medical record writing (Wei Yi Zheng Fa [2010] No. 11). Beijing: Ministry of Health; 2010. (In Chinese).
2. National Health Commission of the People's Republic of China, National Administration of Traditional Chinese Medicine. Notice on launching the comprehensive action to improve medical quality (2023–2025) (Guo Wei Yi Zheng Fa [2023] No. 12). Beijing: National Health Commission; 2023. (In Chinese).
3. Pirkle CM, Dumont A, Zunzunegui MV. Medical recordkeeping, essential but overlooked aspect of quality of care in resource-limited settings. Int J Qual Health Care. 2012;24(6):564–7.
4. Ye Q, Lu X. Application of PDCA in medical records quality training. J Nanjing Med Univ (Soc Sci). 2021;21(4):383–7. (In Chinese).
5. Lorenzetti DL, Quan H, Lucyk K, Cunningham C, Hennessy D, Jiang J, et al. Strategies for improving physician documentation in the emergency department: a systematic review. BMC Emerg Med. 2018;18(1):36.
6. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42.
7. Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodol. 2001;1:6.
8. Tu YK, Gilthorpe MS. Revisiting the relation between change and initial value: a review and evaluation. Stat Med. 2007;26(2):443–57.
9. Cronbach LJ, Furby L. How we should measure “change”: or should we? Psychol Bull. 1970;74(1):68–80.

10. Hake RR. Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am J Phys. 1998;66(1):64–74.
11. Marx JD, Cummings K. Normalized change. Am J Phys. 2007;75(1):87–91.
12. Nissen JM, Talbot RM, Nasim Thompson A, Van Dusen B. Comparison of normalized gain and Cohen's d for analyzing gains on concept inventories. Phys Rev Phys Educ Res. 2018;14(1):010115.
13. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French SD, et al. Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane Database Syst Rev. 2012;(6):CD000259.
14. Taylor MJ, McNicholas C, Nicolay C, Darzi A, Bell D, Reed JE. Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. BMJ Qual Saf. 2014;23(4):290–8.
15. Kirkpatrick DL, Kirkpatrick JD. Evaluating training programs: the four levels. 3rd ed. San Francisco: Berrett-Koehler; 2006.
16. Smidt A, Balandin S, Sigafoos J, Reed VA. The Kirkpatrick model: a useful tool for evaluating training outcomes. J Intellect Dev Disabil. 2009;34(3):266–74.
17. Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process. BMJ Qual Saf. 2016;25(12):986–92.
18. Sichuan Provincial Medical Record Quality Control Center. Sichuan provincial inpatient medical record quality evaluation standard (2023 edition). 2023. (In Chinese).
19. Oldham PD. A note on the analysis of repeated measurements of the same subjects. J Chronic Dis. 1962;15:969–77.

20. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. J R Stat Soc Ser A Stat Soc. 1996;159(3):385–443.
21. Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. BMJ. 1998;316(7146):1701–4.
22. Davis D, O'Brien MA, Freemantle N, Wolf FM, Mazmanian P, Taylor-Vaisey A. Impact of formal continuing medical education: do conferences, workshops, rounds, and other traditional continuing education activities change physician behavior or health care outcomes? JAMA. 1999;282(9):867–74.
23. Forsetlund L, Bjørndal A, Rashidian A, Jamtvedt G, O'Brien MA, Wolf F, et al. Continuing education meetings and workshops: effects on professional practice and health care outcomes. Cochrane Database Syst Rev. 2009;(2):CD003030.
24. Bollen K, Lennox R. Conventional wisdom on measurement: a structural equation perspective. Psychol Bull. 1991;110(2):305–14.
25. Streiner DL. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. J Pers Assess. 2003;80(3):217–22.
26. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA. 2008;300(10):1181–96.
27. Bland JM, Altman DG. Regression towards the mean. BMJ. 1994;308(6942):1499.
28. Barnett AG, van der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. Int J Epidemiol. 2005;34(1):215–20.
29. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally; 1963.
30. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–8.

31. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):166.e7–16.

Figure legend

Fig. 1 Improvement ranks under two change metrics (A) and baseline versus residual defect burden (B)

A Each line connects the rank of one dimension by percentage score gain (left) with its rank by defect burden reduction (right); rank 1 denotes the largest improvement. Solid red lines, rank shift of 5 or more positions; dashed orange lines, 3–4 positions; dotted grey lines, fewer than 3 positions. **B** Each point represents one dimension; error bars show 95% confidence intervals of the means (n = 49 physicians). Defect burden is 100 minus the dimension score (mean points lost per record). CC, chief complaint; HPI, history of present illness; r_s, Spearman's rank correlation coefficient; UI, uncertainty interval

Additional files

Additional file 1 Composition of the 20 audit dimensions: source items from the Sichuan provincial standard and their deduction values (.docx)

Table 1 Definitions of the change metrics compared

Metric	Formula	Denominator	Use in this study
Absolute gain	$S_{post} - S_{pre}$	—	Reference metric
Percentage score gain	$(S_{post} - S_{pre}) \div S_{pre} \times 100\%$	Baseline score	Conventional comparison of relative improvement across dimensions
Defect burden, L	$100 - S$	—	Mean points lost per record in a dimension
Defect burden reduction (normalized gain)	$(L_{pre} - L_{post}) \div L_{pre} \times 100\% = (S_{post} - S_{pre}) \div (100 - S_{pre}) \times 100\%$	Baseline room for improvement	Comparison of relative improvement across dimensions, adjusted for room for improvement
Residual defect burden	$L_{post} = 100 - S_{post}$	—	Identification of targets for the next training cycle

S, dimension score on a 0–100 scale; pre, one month before training; post, one month after training

Table 2 Baseline and post-training scores, improvement metrics and residual defect burden for the 20 audit dimensions (n = 49 physicians)

Dimension	Baseline score, mean ± SD	Post-training score, mean ± SD	Percentage score gain, % (rank)	Defect burden reduction, % (rank; 95% UI)	Rank shift[a]	Residual defect burden, points (rank; 95% UI)
First progress note quality	89.51 ± 9.65	97.97 ± 1.77	9.5 (7)	80.6 (1; 1–10)	+6	2.03 (20; 17–20)
Consultation notes	89.56 ± 7.00	97.66 ± 2.01	9.0 (9)	77.6 (2; 1–12)	+7	2.34 (19; 13–20)
CC–HPI consistency	87.35 ± 7.69	97.07 ± 2.61	11.1 (3)	76.8 (3; 1–13)	0	2.93 (13; 6–19)
Physical examination	88.71 ± 8.29	97.18 ± 2.87	9.5 (6)	75.0 (4; 1–16)	+2	2.82 (15; 6–20)
Past medical history completeness	86.80 ± 11.33	96.60 ± 3.38	11.3 (2)	74.2 (5; 1–17)	−3	3.40 (9; 1–17)
Critical value handling	89.41 ± 8.56	97.22 ± 2.89	8.7 (11)	73.7 (6; 1–17)	+5	2.78 (17; 6–20)
Transfusion record completeness	89.41 ± 6.44	97.19 ± 2.96	8.7 (12)	73.5 (7; 1–17)	+5	2.81 (16; 6–20)
Front sheet completion	85.89 ± 6.94	96.18 ± 3.97	12.0 (1)	72.9 (8; 2–16)	−7	3.82 (3; 1–15)
Difficult-case discussion notes	89.34 ± 5.70	97.10 ± 3.64	8.7 (13)	72.8 (9; 1–17)	+4	2.90 (14; 4–20)
Diagnosis documentation	90.19 ± 8.07	97.33 ± 4.77	7.9 (14)	72.8 (10; 1–19)	+4	2.67 (18; 3–20)
Admission diagnosis accuracy	88.65 ± 7.23	96.89 ± 2.74	9.3 (8)	72.6 (11; 2–17)	−3	3.11 (11; 4–18)
Record format	87.49 ± 6.76	96.57 ± 3.39	10.4 (4)	72.6 (12; 2–17)	−8	3.43 (8; 1–17)
Continuity with nursing records	87.43 ± 9.53	96.31 ± 4.21	10.2 (5)	70.6 (13; 2–18)	−8	3.69 (5; 1–17)
HPI completeness	88.47 ± 5.99	96.37 ± 3.79	8.9 (10)	68.5 (14; 3–18)	−4	3.63 (6; 1–16)
Senior physician ward-round notes	90.36 ± 6.35	96.92 ± 3.79	7.3 (17)	68.0 (15; 3–19)	+2	3.08 (12; 2–19)
Discharge summary quality	89.73 ± 7.72	96.41 ± 3.43	7.4 (16)	65.0 (16; 6–19)	0	3.59 (7; 1–16)
Chief complaint quality	89.06 ± 5.77	96.04 ± 2.95	7.8 (15)	63.8 (17; 9–19)	−2	3.96 (1; 1–11)
Resuscitation records	91.37 ± 6.28	96.76 ± 3.53	5.9 (18)	62.5 (18; 7–20)	0	3.24 (10; 2–18)
Medical order completeness	91.78 ± 6.95	96.26 ± 4.43	4.9 (19)	54.5 (19; 11–20)	0	3.74 (4; 1–17)
Informed consent completeness	92.34 ± 4.98	96.12 ± 4.08	4.1 (20)	49.3 (20; 15–20)	0	3.88 (2; 1–15)

Dimensions are listed in descending order of defect burden reduction. Defect burden = 100 − dimension score; residual defect burden is the defect burden after training (rank 1 = heaviest). 95% UIs for ranks are the 2.5th–97.5th percentiles of 10 000 parametric simulations. All baseline versus post-training comparisons $P < 0.001$ after Bonferroni correction. [a]Rank by percentage score gain minus rank by defect burden reduction; positive values indicate a higher rank under defect burden reduction. CC, chief complaint; HPI, history of present illness; SD, standard deviation; UI, uncertainty interval

Figures

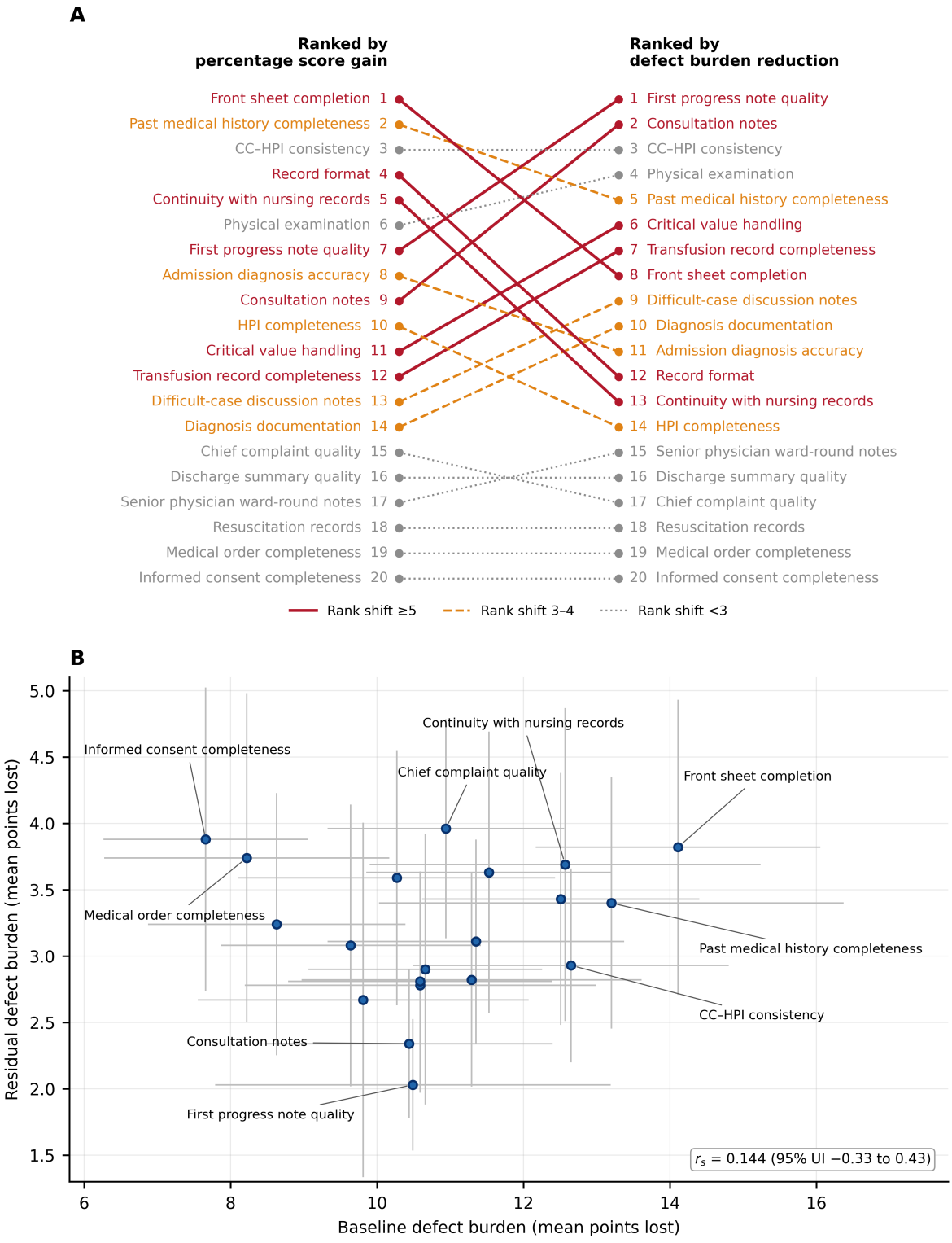


Figure 1

Improvement ranks under two change metrics (A) and baseline versus residual defect burden (B)

A Each line connects the rank of one dimension by percentage score gain (left) with its rank by defect burden reduction (right); rank 1 denotes the largest improvement. Solid red lines, rank shift of 5 or more positions; dashed orange lines, 3–4 positions; dotted grey lines, fewer than 3 positions. **B** Each point represents one dimension; error bars show 95% confidence intervals of the means (n = 49 physicians). Defect burden is 100 minus the dimension score (mean points lost per record). CC, chief complaint; HPI, history of present illness; r_s, Spearman's rank correlation coefficient; UI, uncertainty interval

Supplementary Files

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

- Additionalfile1.docx