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.

Residual learning for predicting revisions to hospital reimbursement standards

Yuyang Liang
Chinese University of Hong Kong, Shenzhen
Xiang Meng
Chinese University of Hong Kong, Shenzhen
Jinsong Han
Shandong First Medical University
Zhuodong Liu
Shenzhen Hospital (Futian) of Guangzhou University of Chinese Medicine
Shipei Zeng
Shenzhen Research Institute of Big Data
Danlan Li
Shenzhen Hospital (Futian) of Guangzhou University of Chinese Medicine
Zicheng Wang
wangzicheng@cuhk.edu.cn
Chinese University of Hong Kong, Shenzhen
Chenhao Ma
Chinese University of Hong Kong, Shenzhen
Hanli Lin
Shenzhen Hospital (Futian) of Guangzhou University of Chinese Medicine

Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11844387/v1

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

Additional Declarations: No competing interests reported.

Residual learning for predicting revisions to hospital reimbursement standards

Yuyang Liang[1†], Xiang Meng[1†], Jinsong Han[5†], Zhuodong Liu[3†], Shipei Zeng[2], Danlan Li[3], Zicheng Wang[1*], Chenhao Ma[1*], Hanli Lin[4*]

[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen, Shenzhen, Guangdong, China.
[2]Center for AI Large Foundation Models, Shenzhen Research Institute of Big Data, Shenzhen, Guangdong, China.
[3]Department of Operation Management, Shenzhen Hospital (Futian) of Guangzhou University of Chinese Medicine, Shenzhen, Guangdong, China.
[4]Department of Hospital Administration, Shenzhen Hospital (Futian) of Guangzhou University of Chinese Medicine, Shenzhen, Guangdong, China.
[5]Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan, Shandong, China.

*Corresponding author(s). E-mail(s): wangzicheng@cuhk.edu.cn; machenhao@cuhk.edu.cn; HLlin99@outlook.com;
Contributing authors: yuyangliang@link.cuhk.edu.cn; xiangmeng@link.cuhk.edu.cn; hanjs@sd-cancer.com; zhuodongliu@outlook.com; shipei.zeng@sribd.cn; dawnmed90@126.com;
†These authors contributed equally to this work.

Abstract

Hospital reimbursement standards may be revised after care is delivered, creating financial uncertainty. We developed ResPay, a neural framework that uses residual learning to predict corrections to current standards from a single hospital's clinical and cost records. A jointly trained classifier identifies groups likely to undergo large revisions and enables selective correction. We retrospectively

evaluated separate models at two Chinese hospitals on 219 diagnosis-intervention packet (DIP) and 69 diagnosis-related group (DRG) groups withheld from training within each cohort's payment transition. ResPay achieved the highest average precision (0.731 and 0.568, respectively) for identifying large revisions among evaluated methods. It also achieved the lowest mean absolute and root mean squared errors, reducing mean absolute error by 17.27% in DIP and 11.43% in DRG relative to retaining current standards. These exploratory findings support further evaluation of reimbursement forecasting for hospital planning.

Keywords: Hospital reimbursement; Reimbursement prediction; Diagnosis-related groups; Diagnosis-intervention packet; Residual learning

1 Introduction

Case-based payment is a common approach to hospital payment reform, whereby hospitals are reimbursed according to predefined standards for groups of inpatient cases [1]. Diagnosis-related group (DRG) and diagnosis-intervention packet (DIP) payment are two examples. DRG systems classify inpatient cases into groups with similar clinical characteristics and expected resource requirements, whereas DIP systems assign relative payment values to combinations of diagnoses and procedures [2, 3]. In some implementations, however, the payment parameters that determine final reimbursement are not fully specified when care is delivered. In China, for example, Shenzhen's DIP rules distinguish monthly provisional settlement from annual settlement based on an annual point value, while Jinan's DRG rules allow payment rates to be adjusted at year-end settlement [4, 5]. Such arrangements create financial uncertainty for hospitals, which may incur treatment costs before knowing the reimbursement they will receive. Predicting revised reimbursement standards for individual payment groups could help hospitals estimate expected revenue, prepare budgets, and allocate resources.

The financial implications of a revision depend on both its size and the number of hospital cases to which it applies. In our analysis of two hospitals in China, most payment groups underwent modest revisions, whereas a small proportion experienced large revisions that accounted for a disproportionate share of financial exposure (Fig. 1c–f). This uneven pattern may partly reflect adjustments to payment rules that affect particular groups more strongly than others. Chinese DRG guidance, for example, allows group weights to be adjusted to reflect resource use, treatment complexity and policy priorities [6]. The DRG 2.0 revision also targeted selected clinical specialties and refined groups with high resource use [7]. Such changes could produce substantial revisions for some groups even when standards for most groups remain relatively stable. A small number of affected groups may therefore warrant considerable attention, particularly when they contribute disproportionately to hospitals' financial exposure. Identifying groups likely to undergo large revisions can help hospitals recognize where current standards may provide a misleading basis for financial and operational planning.

However, predicting reimbursement revisions, particularly large revisions, is challenging when only a single hospital's records are available. Because clinical characteristics and cost composition vary across payment groups, each group is best characterized from its own cases, and groups with few patient cases provide limited information for distinguishing systematic patterns from individual variation. Moreover, regional reimbursement standards reflect clinical activity and resource use across all hospitals in a region, of which a single hospital observes only its own share. Large revisions are also relatively rare, which limits the examples available for learning what characterizes them. Research on related tasks nevertheless suggests that useful predictive information can be extracted from healthcare records [8–10]. Previous studies have used clinical notes to predict patients' DRG assignments [11] and obtain the corresponding cost weights [12], and insurance claims to predict patient expenditures [13] and identify large cost changes [14]. These studies establish the potential of healthcare records for related tasks but leave an important question unanswered: whether a hospital can use its own clinical and cost records, together with current standards, to predict revised reimbursement standards and identify groups likely to undergo large revisions.

To address this question, we develop ResPay, a neural framework for predicting revised reimbursement standards and identifying groups likely to undergo large revisions. ResPay combines learned record representations with summaries of clinical characteristics, cost composition and, where available, within-unit variation and record counts. Each prediction unit's representation is constructed from its own records, using encoder parameters shared across units. A residual predictor uses this representation to estimate signed corrections to current standards, while a classifier produces a group-level score for identifying large revisions. The representation network, residual predictor and classifier are trained jointly, with all training units contributing to numerical prediction regardless of revision magnitude. At prediction time, ResPay applies a correction when the classification score reaches a specified threshold and otherwise retains the current standard. In our evaluation, the selected thresholds applied corrections to all DIP groups and selectively to DRG groups, illustrating how users can adjust the threshold to suit their payment setting and planning needs.

We evaluated ResPay retrospectively in DIP and DRG cohorts from two Chinese hospitals, training a separate model using each hospital's own data. Evaluation included 219 held-out DIP groups and 69 held-out DRG groups, assessing generalization to unseen payment groups within the same hospital and payment transition. On average across five training runs, ResPay achieved the highest average precision for identifying large revisions and the lowest mean absolute and root mean squared errors among the evaluated methods in both cohorts. Numerical accuracy was assessed across all held-out groups, including those for which current standards were retained. Relative to retaining current standards for every group, ResPay reduced mean absolute error by 17.27% in DIP and 11.43% in DRG. These findings support the feasibility of using single-hospital data to identify large revisions and predict revised reimbursement standards.

2 Results

2.1 Study cohorts

We evaluated ResPay retrospectively using cohorts from two hospitals in China, training a separate model with each hospital's own data (Fig. 1a,b). The DIP cohort comprised 2023 discharge records from a Grade-A tertiary hospital in Shenzhen. Each disease–procedure group formed a prediction unit containing individual discharge records, with its 2023 payment standard as the current standard and its 2024 standard as the prediction target. The DRG cohort used clinical and cost records from version 1.1 cases treated between 2023 and the first quarter of 2025 at an oncology hospital in Jinan to predict version 2.0 payment standards. Each disease–insurance combination formed a prediction unit represented by one aggregated input record. For evaluation, insurance-specific current standards, revised standards and numerical predictions were aggregated to the disease-group level using the same insurance shares calculated from each group's source-period cases.

Within each cohort, a fixed random split assigned approximately 80% of payment groups to training and the remainder to testing. DIP groups were split at the disease–procedure level, keeping all records from each group together. DRG groups were split at the disease level, keeping all insurance-specific units from each disease group together. The DIP test set contained 219 groups and 3,956 individual discharge records. The DRG test set contained 69 disease groups, represented by 141 disease–insurance units summarizing 7,840 source cases. This design assessed generalization to payment groups excluded from training within the same hospital and payment transition. Appendices A, B and D provide details of cohort construction, preprocessing and evaluation, respectively.

2.2 Large revisions accounted for a disproportionate share of financial exposure

We defined groups with large reimbursement revisions (LR groups) as those whose observed absolute difference between revised and current standards was at least 0.45 in DIP or 0.25 in DRG, using the payment-standard scales described in Appendix B. These thresholds identified 24 of the 219 held-out DIP groups (11.0%) and 10 of the 69 held-out DRG groups (14.5%) as LR groups (Fig. 1c,d).

To assess their financial relevance, we calculated case-weighted absolute revision exposure by multiplying each group's absolute payment-standard revision by its source-period case count and summing across groups. Case counts and, for DRG, insurance shares were held fixed. DIP exposure was measured in case-weighted standardized fee units, while DRG exposure was calculated from payment amounts in RMB. Although LR groups accounted for only 5.18% of held-out DIP cases, they contributed 34.67% of total exposure. In DRG, LR groups accounted for 18.42% of cases and 79.41% of exposure (Fig. 1e,f).

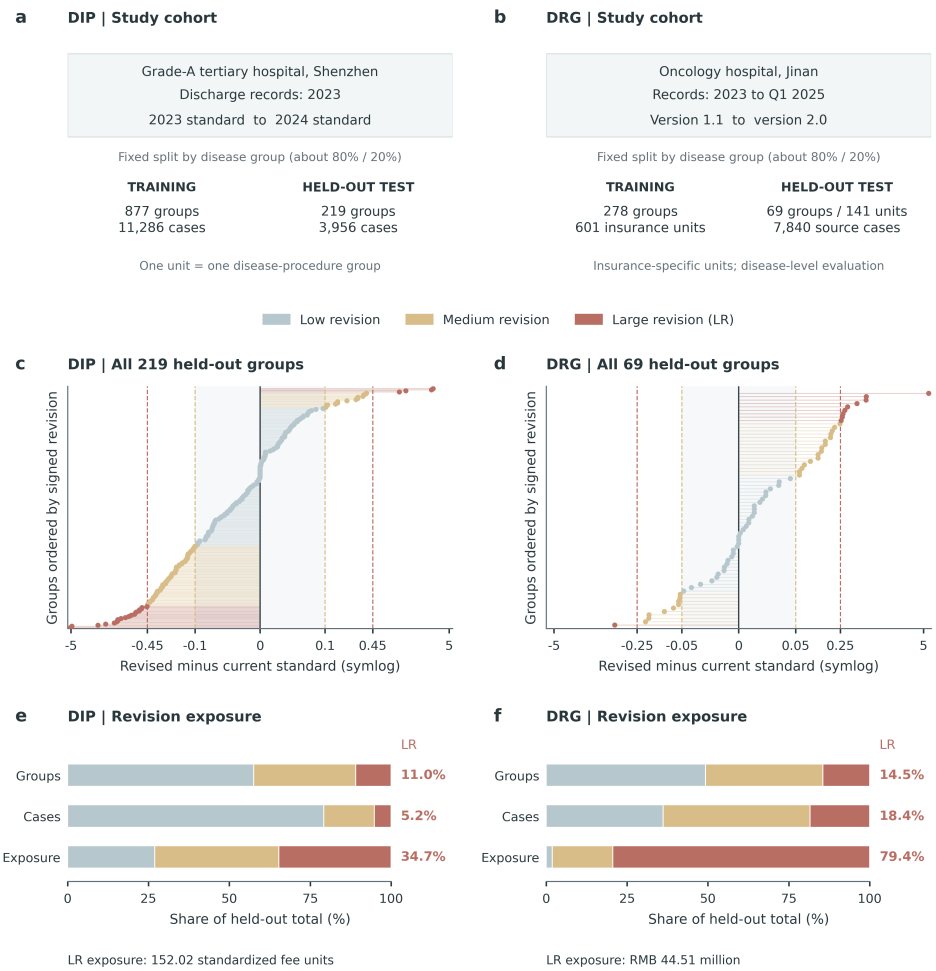


Fig. 1 Study cohorts and concentration of reimbursement revisions. a,b, Cohort construction, payment transitions and training–test splits for DIP and DRG, respectively. **c,d**, Observed revisions across all held-out DIP and DRG groups. Revisions are calculated as revised minus current standards; zero indicates no change. Horizontal axes use a symmetric-logarithmic scale. Dashed lines mark the absolute-revision boundaries separating small, moderate and large revisions. **e,f**, Shares of held-out groups, cases and case-weighted absolute revision exposure within each revision stratum. Exposure is calculated by summing each group's absolute revision multiplied by its source-period case count. DRG standards are first aggregated using source-period insurance shares. Case counts and insurance shares are held fixed. Labels show the shares attributable to LR groups.

2.3 Identification of payment groups with large revisions

We compared ResPay with Ridge regression, a multilayer perceptron (MLP) and XGBoost in identifying LR groups among all held-out payment groups. ResPay ranked groups by its group-level classification score, while the comparison methods ranked

them by the absolute predicted correction to the current standard. For DRG, correction magnitudes were calculated after aggregating insurance-specific corrections to the disease-group level. These rankings used all held-out groups, independently of the threshold controlling whether corrections were applied. Identification performance was evaluated against the observed LR labels using average precision (AP) and the area under the receiver operating characteristic curve (ROC-AUC).

ResPay achieved the highest AP among the evaluated methods in both cohorts (Fig. 2a–d), with values of 0.731 ± 0.010 in DIP and 0.568 ± 0.028 in DRG. These values are the means and sample standard deviations across five training runs using the same held-out split within each cohort. The strongest comparison methods by AP were XGBoost in DIP (0.635) and Ridge in DRG (0.399). ResPay also achieved the highest mean ROC-AUC in DRG (0.871). In DIP, Ridge's ROC-AUC was slightly higher than ResPay's mean ROC-AUC (0.922 versus 0.910).

2.4 Prediction of revised reimbursement standards

We evaluated numerical prediction accuracy on the same 219 held-out DIP groups and 69 held-out DRG groups for every method, giving each group equal weight. For ResPay, we used the classification-score thresholds selected from the sensitivity analysis: 0 for DIP and 0.75 for DRG. All DIP groups therefore received predicted corrections, while DRG groups with scores below 0.75 retained their current standards. MAE and RMSE were calculated across every held-out group, including those for which the current standard was retained.

ResPay achieved the lowest mean MAE and RMSE among the evaluated methods in both cohorts. Its MAE and RMSE were 0.2114 ± 0.0032 and 0.5354 ± 0.0097 in DIP, and 0.1702 ± 0.0036 and 0.6048 ± 0.0140 in DRG, respectively. These values are the means and sample standard deviations of metrics calculated separately for five training runs on the same held-out groups. Relative to retaining current standards for every group, ResPay reduced mean MAE by 17.27% in DIP and 11.43% in DRG (Fig. 2g,h).

Compared with Ridge, ResPay reduced mean MAE and RMSE by 8.41% and 5.54% in DIP, and by 14.83% and 5.26% in DRG, respectively (Fig. 2g,h). To examine how these improvements were distributed, we also compared each group's absolute error averaged across training runs. ResPay had lower mean absolute errors than Ridge in 132 of 219 DIP groups and 40 of 69 DRG groups (Fig. 2e,f). Improvements thus extended to more than half of the groups in each cohort, although ResPay produced larger errors for some groups.

2.5 Residual prediction improved accuracy

ResPay predicts corrections to current standards because these standards already provide an established reference for reimbursement. This design is motivated by the observed pattern that most groups undergo modest revisions, while a smaller proportion experience substantial changes. By adding a learned correction to the current standard, ResPay explicitly incorporates the existing payment level and focuses numerical learning on departures from it. This structure may be useful when hospital data

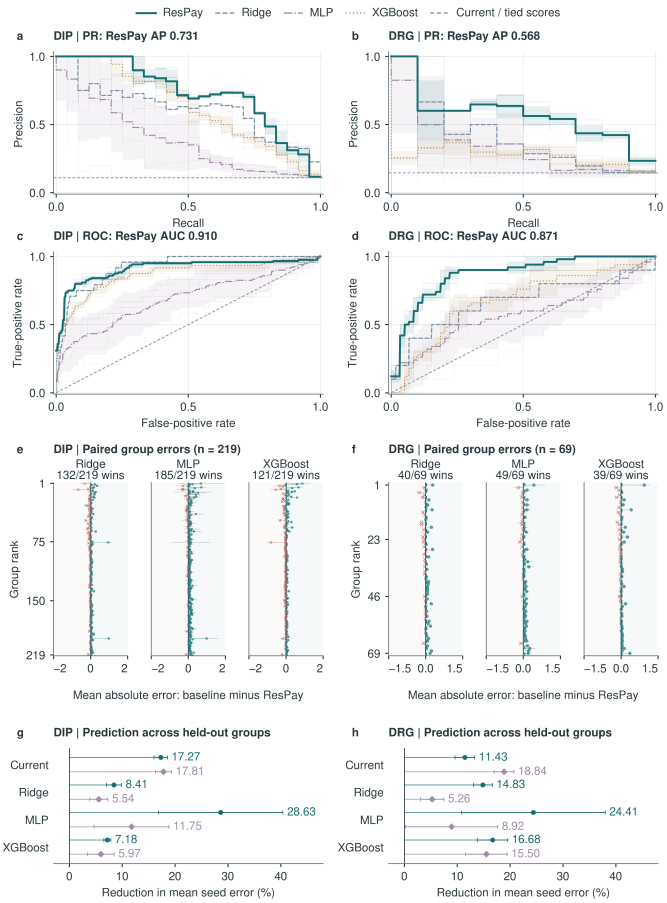


Fig. 2 Identification of large revisions and prediction of revised reimbursement standards. a–b, Precision–recall curves for identifying large-revision (LR) groups in DIP and DRG. **c–d**, Corresponding receiver operating characteristic (ROC) curves. Groups are ranked using ResPay's LR classification score or the absolute predicted correction for Ridge, multilayer perceptron (MLP) and XGBoost. Observed LR labels indicate absolute revisions of at least 0.45 in DIP or 0.25 in DRG. Reported average precision (AP) and ROC-AUC values are means across five training seeds. Gray reference lines indicate LR prevalence in the precision–recall plots and the no-discrimination diagonal in the ROC plots. **e,f**, Group-level differences in absolute error between each comparison method and ResPay, ordered by decreasing observed absolute revision. Points and whiskers show the mean and sample standard deviation of within-seed error differences across five training seeds. Positive values indicate lower errors for ResPay; annotations give the number of groups with lower mean absolute errors. **g,h**, Relative reductions in mean absolute error (MAE; teal circles) and root mean squared error (RMSE; mauve diamonds) compared with retaining current standards, Ridge, MLP and XGBoost. Reductions are calculated from mean metrics across seeds. Whiskers show the sample standard deviation of within-seed metric differences, expressed as a percentage of the corresponding baseline's mean metric. All evaluations use the same 219 DIP groups and 69 DRG groups across runs. Numerical predictions use classification score thresholds of 0 for DIP and 0.75 for DRG and are evaluated across all held-out groups, including those retaining current standards.

are limited, as the predictor does not need to learn how to reproduce the current standard as part of its output.

To assess the contribution of this design, we compared ResPay with a variant trained to predict revised standards directly. The variant retained current standards as inputs and used the same network architecture and training settings. Thus, both variants had access to the existing payment levels, but only the residual formulation included them as a fixed additive component of the numerical forecast. At the selected thresholds, the direct-prediction variant used its predicted standard for every DIP group and for DRG groups with classification scores of at least 0.75; the remaining DRG groups retained their current standards. All network components were retrained jointly, so the DRG groups selected by the classifier could differ between variants. Both variants were evaluated on the same full held-out cohorts across five training runs. Residual prediction yielded lower mean MAE and RMSE in both cohorts (Fig. 3a,b). In DIP, MAE was 0.2114 for residual prediction versus 0.2203 for direct prediction, and RMSE was 0.5354 versus 0.5520. In DRG, the corresponding MAE values were 0.1702 and 0.1784, and RMSE values were 0.6048 and 0.6194. These findings suggest that explicitly learning corrections can be more effective than supplying current standards as input features alone. The comparison supports the residual formulation as implemented, including its target transformation, within the jointly trained framework.

2.6 Classification score thresholds and selective correction

The LR classification score threshold determines how broadly ResPay applies predicted corrections. A threshold of zero applies corrections to every group, allowing the numerical predictor to estimate both modest and large revisions. Higher thresholds restrict corrections to groups with higher LR scores and retain current standards for the remaining groups. The threshold changes where corrections are applied without changing the definition of LR groups or the estimated correction for each group.

We varied the threshold from 0 to 1 in increments of 0.05, holding the trained model, numerical corrections and classification scores fixed for each cohort and training seed (Fig. 3c–f). Prediction errors were evaluated across all held-out groups and summarized across five training seeds. For clarity, the figure displays five thresholds: 0, 0.25, 0.5, 0.75 and 0.95. In DIP, low thresholds yielded the smallest errors, favoring broad application of corrections in this cohort. In DRG, increasing the threshold from 0 to 0.75 reduced mean MAE from 0.1780 to 0.1702, while mean RMSE was 0.6048 at both thresholds to four decimal places. Thus, withholding corrections from some lower-scoring DRG groups reduced MAE without materially changing RMSE. The thresholds used in the main prediction and residual-ablation comparisons, 0 for DIP and 0.75 for DRG, were selected from this exploratory sensitivity analysis of the held-out cohorts. ResPay maintained lower mean MAE and RMSE than all evaluated baselines across a broad range of thresholds in both cohorts.

These results suggest that the appropriate threshold depends on the prediction performance of ResPay relative to retaining current standards in each payment setting. Applying corrections to more groups allows ResPay to capture modest revisions, but may also introduce errors in groups whose current standards remain close to their

revised standards. Selective correction can avoid some of these adjustments, although corrections may also reduce errors for some groups assigned low LR scores. When applying ResPay to a new dataset, hospitals can assess this tradeoff using separate validation groups from completed payment transitions.

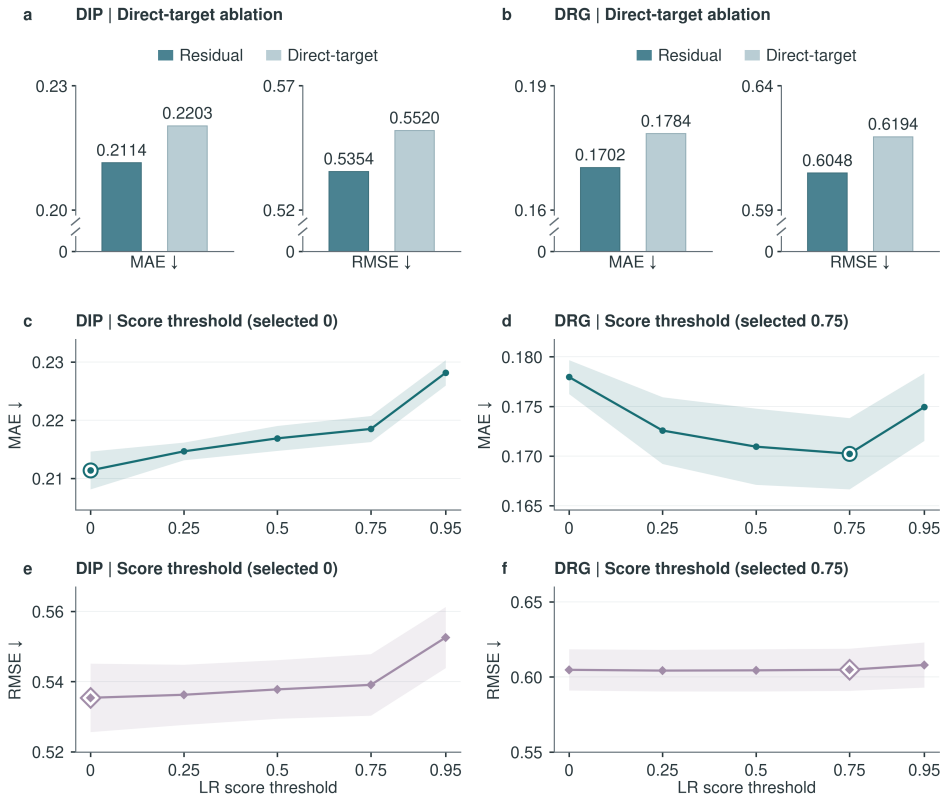


Fig. 3 Residual prediction and sensitivity to the LR classification score threshold. a,b, Comparison of residual prediction with direct prediction of revised reimbursement standards in DIP and DRG. Both variants retain current standards as inputs and use the same network architecture, training settings and classification score thresholds, with all model components retrained jointly. Bars show mean absolute error (MAE) and root mean squared error (RMSE), averaged across five training seeds, using thresholds of 0 for DIP and 0.75 for DRG. **c–f**, Changes in MAE (c,d; teal circles) and RMSE (e,f; mauve diamonds) as the LR classification score threshold varies. The trained model, numerical corrections and classification scores are held fixed within each run, so only the groups receiving corrections change. Thresholds were evaluated from 0 to 1 in increments of 0.05; the figure displays 0, 0.25, 0.5, 0.75 and 0.95. Points show mean errors across five training seeds. Outlined markers identify the thresholds used in the main comparisons, selected from this exploratory analysis of the held-out cohorts. All errors are evaluated across the full held-out cohorts of 219 DIP groups and 69 DRG groups.

3 Discussion

In this study, we found that an individual hospital's clinical and cost records, together with current standards, could support both the prediction of revised reimbursement standards and the identification of groups with large revisions. In retrospective DIP and DRG evaluations, ResPay achieved the highest average precision and the lowest mean MAE and RMSE among the evaluated methods, with models trained separately for each hospital. Relative to retaining current standards, mean MAE decreased by 17.27% in DIP and 11.43% in DRG. These findings suggest that routinely collected hospital data contain information that could inform financial and operational planning under reimbursement uncertainty. Hospitals could combine predicted standards with their anticipated case mix to estimate expected reimbursement, while LR scores could identify groups for which plans based on current standards require closer review. By accounting for potential reimbursement changes, hospitals may be better positioned to prepare budgets and allocate resources while revised standards remain uncertain.

The residual ablation supports the use of current standards as an explicit reference for predicting reimbursement revisions. By learning changes from existing payment levels, ResPay incorporates information already available to hospitals and focuses numerical learning on how those levels may change. This design may be particularly relevant when local data are limited and most groups undergo modest revisions. Predicting corrections yielded lower errors than predicting revised standards directly, even though both variants received current standards as inputs. LR identification complements this approach by providing a score trained specifically to distinguish groups with large revisions, while all training groups contribute to numerical learning. This distinction allows ResPay to address the identification of relatively uncommon large revisions without restricting numerical training to those groups. The threshold analysis further shows that the benefits of selective correction differed between cohorts: applying corrections to all groups performed best in DIP, whereas retaining current standards for some lower-scoring groups reduced MAE in DRG. These findings support flexibility in the application of corrections. However, the LR score identifies potential large revisions rather than the accuracy of their estimated corrections, so threshold selection should consider the errors of the final predictions.

Our study has several limitations. The evaluation involved two hospitals and one payment transition per cohort, with training and testing conducted on different groups within the same transition. Performance for future payment transitions therefore remains to be established. When data from successive transitions become available, models should be trained on earlier transitions and evaluated on later ones using only information available at the intended prediction date. The classification-score thresholds were selected using the same held-out cohorts used for performance evaluation, which may yield optimistic estimates; future evaluations should select thresholds on separate validation data. Another direction is to examine whether combining data from multiple hospitals improves prediction relative to using a single hospital's records. Additional records may better characterize clinical activity and resource use, particularly for groups with few local cases, although differences in case mix and payment rules must be considered. Further work should also assess uncertainty in individual

corrections and examine case-weighted errors and aggregate reimbursement predictions. Prospective studies could then determine whether improved forecasts lead to better reimbursement estimates, budget preparation and resource allocation.

4 Methods

4.1 Study design and prediction task

We evaluated ResPay retrospectively using clinical and cost records from two Chinese hospitals, training a separate model for each hospital. The DIP task used 2023 discharge records from a Grade-A tertiary hospital in Shenzhen to predict 2024 payment standards from 2023 standards. The DRG task used records collected between 2023 and the first quarter of 2025 at an oncology hospital in Jinan to predict the transition from version 1.1 to version 2.0 standards.

Input records, prediction units and payment groups differed between cohorts. In DIP, each disease–procedure group was both a prediction unit and an evaluation payment group, containing individual discharge records. In DRG, each disease–insurance combination was a prediction unit represented by one aggregated record; evaluation was at the disease-group level. DIP and DRG inputs comprised 16 and 24 features, respectively, describing clinical and cost information together with current standards. Revised standards were supervised targets and were excluded from predictors.

Let p_u and y_u denote the current and revised standards for prediction unit u on the cohort's analytical scale. The analytical scale refers to the scale on which payment standards are modeled and evaluated: the supplied processed standard-score fees for DIP and payment amounts standardized using training disease-group statistics for DRG (Appendix B). For each payment group g, let U_g denote its set of prediction units and w_{gu} the aggregation weight assigned to unit $u \in U_g$, with $w_{gu} \geq 0$ and $\sum_{u \in U_g} w_{gu} = 1$. The corresponding group-level current and revised standards, p_g and y_g, are defined as

$$p_g = \sum_{u \in U_g} w_{gu} p_u, \qquad y_g = \sum_{u \in U_g} w_{gu} y_u. \tag{1}$$

For DIP, each group contains a single prediction unit with weight one. For DRG, U_g contains the insurance-specific units of disease group g, and their weights are the corresponding source-period insurance shares. These weights are held fixed when aggregating current standards, revised standards and predictions.

Within each cohort, a fixed random split assigned approximately 80% of payment groups to training and the remainder to testing, keeping all records and insurance-specific units belonging to the same group in one partition. This design assessed prediction performance for groups excluded from training within the same hospital and payment transition. Details of cohort construction, feature definitions and preprocessing are provided in Appendices A and B.

4.2 ResPay architecture and prediction rule

ResPay combines a residual predictor with a classifier for large reimbursement revisions (LR groups), sharing a learned representation of each prediction unit (Fig. 4).

Residual learning uses the current standard as a reference, while classification provides a separate score for identifying groups with large revisions and controlling whether a correction enters the final prediction.

A record encoder maps each input record to a learned vector. Attention pooling combines these vectors within each prediction unit, allowing records to contribute unequally [15]. In parallel, a summary encoder processes feature means, population standard deviations and the logarithm of one plus the input-record count. The pooled and summary vectors are fused into a shared unit representation. For DIP, this captures information across individual cases. Each DRG unit contains one aggregated record, so attention pooling returns that record's encoding and the variation and record-count summaries are constant; underlying case counts and insurance shares remain separate input features.

The numerical predictor estimates the signed residual $r_u = y_u - p_u$. A record-level head receives each record encoding and its unit representation; its outputs are averaged equally within the unit and added to a unit-level head's output. Attention weights therefore affect representation pooling rather than the numerical average. DIP residual targets are standardized using training-unit statistics and restored before prediction; DRG residual targets require no additional standardization. Unit corrections are aggregated to $\widehat{r}_g$ using the weights in Eq. (1).

The LR classifier reads the shared unit representation and is supervised by

$$H_g = \mathbf{1}\{|y_g - p_g| \geq \tau\}, \tag{2}$$

where $\tau = 0.45$ for DIP and 0.25 for DRG on their respective analytical scales. A classification head produces unit logits; for DRG, these are averaged using source-period insurance shares before applying the logistic sigmoid. For DIP, the sigmoid is applied to the single unit's logit. The resulting group score q_g is an uncalibrated classification score.

The final prediction applies the estimated correction only when the score reaches threshold t:

$$\widehat{y}_g = p_g + \mathbf{1}\{q_g \geq t\}\widehat{r}_g. \tag{3}$$

The parameter sensitivity experiment selected $t = 0$ for DIP and $t = 0.75$ for DRG. Thus, all DIP groups receive corrections, whereas DRG groups below 0.75 retain current standards. The label threshold τ defines observed large revisions; t controls correction application. The gate operates only at prediction time, and all training units contribute to numerical learning. Complete representation and prediction equations and layer configurations are provided in Appendix C.

4.3 Joint learning and training

The shared representation network, numerical heads and LR classifier are trained jointly using four loss components. The unit-level loss $\mathcal{L}_{\text{unit}}$ combines absolute and squared errors in residual predictions, while the auxiliary record-level loss $\mathcal{L}_{\text{rec}}$ compares each record-head output with its unit's residual target. The DRG aggregation loss $\mathcal{L}_{\text{agg}}$ compares insurance-weighted predictions with observed disease-level revisions. The classification loss $\mathcal{L}_{\text{LR}}$ uses weighted binary cross-entropy to account for

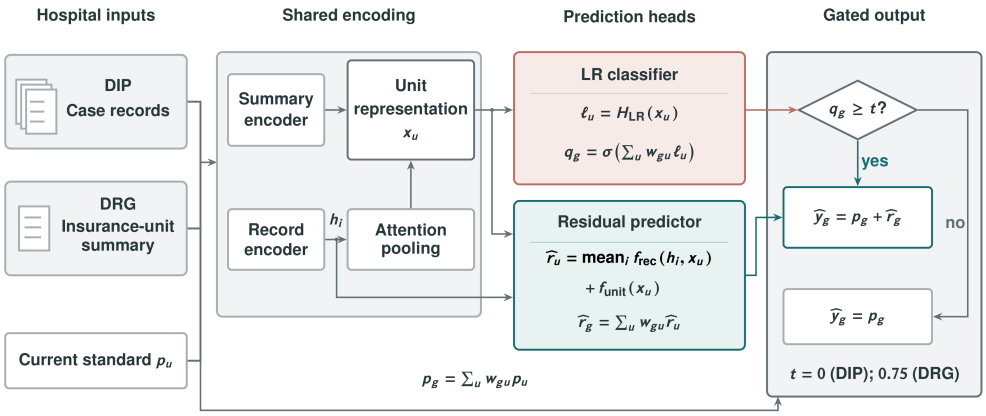


Fig. 4 ResPay architecture for predicting revised reimbursement standards. Within each prediction unit, attention-pooled record encodings are combined with encoded feature summaries to form a representation shared by the residual predictor and the large-revision (LR) classifier. Feature summaries include within-unit means, standard deviations and the logarithm of one plus the input-record count. The residual predictor combines the equally weighted mean of record-level outputs with a unit-level output. DIP outputs are transformed back to the payment-standard scale before aggregation. For DRG, insurance-specific corrections and classification logits are aggregated using source-period insurance shares; the logistic sigmoid is then applied to the aggregated logit to obtain the group-level LR score. The final prediction adds the estimated correction to the current standard when the LR score meets the classification score threshold and retains the current standard otherwise. The thresholds used in this study are 0 for DIP and 0.75 for DRG.

the imbalance between LR and non-LR groups. Setting the aggregation weight λ_{agg} to one for DRG and zero for DIP, the total loss is

$$\mathcal{L} = \mathcal{L}_{\text{unit}} + \mathcal{L}_{\text{rec}} + \lambda_{\text{agg}} \mathcal{L}_{\text{agg}} + 0.5 \mathcal{L}_{\text{LR}}. \tag{4}$$

The unit-level loss assigns equal weight to prediction units, the auxiliary record-level loss assigns equal weight to input records, and the aggregation and classification losses assign equal weight to payment groups. DIP units containing more cases therefore contribute more terms to the auxiliary loss, while each aggregated DRG record contributes one term. Each record is supervised using its unit's residual target, which represents a revision to the payment standard rather than an individual patient's cost.

For each hospital, we trained the model using the AdamW optimizer [16] with five random seeds, keeping the group-based training and test partitions fixed across runs. Each epoch processed all training records in a single forward pass followed by one parameter update. The classification score threshold was applied only at prediction time. Detailed loss definitions, coefficients, class weights and numerical operations are provided in Appendix C.

4.4 Evaluation and ablation

We compared ResPay with retaining current standards, Ridge regression [17], a multilayer perceptron (MLP) and XGBoost [18] in the held-out DIP and DRG cohorts. The unchanged-standard baseline predicted $\widehat{y}_g = p_g$, while the learned baselines estimated corrections to current standards. LR identification was evaluated using average precision (AP) and the area under the receiver operating characteristic curve (ROC-AUC). ResPay ranked groups by its LR classification score q_g, whereas the learned baselines ranked groups by the absolute predicted correction, calculated after insurance-weighted aggregation for DRG. Numerical prediction was evaluated using mean absolute error (MAE) and root mean squared error (RMSE). For ResPay, these metrics used the final predictions after applying the classification-score threshold. All methods were evaluated on the same analytical scale within each cohort.

To assess residual prediction, we compared ResPay with a variant that predicted revised standards directly. This variant retained current standards as inputs and used the same architecture, training settings, LR labels and classification-score thresholds. All components of each variant were retrained jointly, and both variants were evaluated using MAE, RMSE and the coefficient of determination (R^2). Joint retraining could change which DRG groups received corrections; the comparison therefore assessed the residual formulation, including its associated target transformation, within the complete framework.

For threshold sensitivity analysis, we held each trained model's numerical corrections and classification scores fixed and varied the classification-score threshold t from 0 to 1 in increments of 0.05. Prediction errors were evaluated across all held-out groups. The thresholds used in the main numerical prediction and ablation comparisons, 0 for DIP and 0.75 for DRG, were selected from this exploratory analysis of the same held-out cohorts used for performance evaluation.

4.5 Statistics and reproducibility

Statistical analyses were conducted at the payment-group level using 219 held-out DIP disease–procedure groups and 69 held-out DRG disease groups. All methods were evaluated on the same groups. MAE and RMSE were calculated over every held-out group with equal weight, including groups for which ResPay retained the current standard.

ResPay and its direct-prediction variant were each trained using five corresponding random seeds, with the training and test partitions held fixed across runs. Their performance metrics were calculated separately for each run and summarized using the mean and sample standard deviation. The reported standard deviations therefore describe variation across repeated model training on the same evaluation cohort.

Relative reductions in MAE and RMSE were calculated from the mean metrics across runs. Error bars for these reductions were calculated as the sample standard deviation of within-run metric differences, divided by the baseline's mean metric and expressed as a percentage. Group-level comparisons used the difference between each baseline's absolute error and ResPay's absolute error within each run, summarized by the mean and sample standard deviation across runs. Positive differences indicate

lower errors for ResPay. For comparisons involving fixed baseline predictions, the same baseline predictions were used across runs.

Descriptive analyses stratified groups by observed revision magnitude. Case-weighted absolute revision exposure was calculated by summing each group's absolute revision multiplied by its source-period case count. Case counts and DRG insurance shares were held fixed, with exposure measured on the analytical scale for DIP and in RMB for DRG. Appendix D provides further details of the statistical calculations, revision strata, exposure measures, direct-prediction ablation and threshold sensitivity analysis.

Model training and statistical analyses were implemented in Python. The verified reproducibility environment used Python 3.12.2, PyTorch 2.9.1, NumPy 1.26.4, pandas 2.2.2, SciPy 1.13.1, scikit-learn 1.6.1, XGBoost 3.2.0 and openpyxl 3.1.5.

4.6 Use of large language models

The authors used large language models (GPT-6.1 Sol) to assist with language editing of selected sections of the manuscript and to generate portions of the code used in this study. All LLM-assisted text and code were reviewed by the authors and revised where necessary. The authors take full responsibility for the accuracy and integrity of the manuscript, code and reported results.

5 Data availability

The hospital-derived datasets analyzed in this study are not publicly available owing to confidentiality restrictions on medical insurance and hospital operational records. Fully synthetic example inputs for the DIP and DRG cohorts are publicly available in the code repository at https://github.com/LeoMenh/ResPay-payment-standard-prediction, in the `respay/data/samples/` directory.

6 Code availability

The ResPay implementation is publicly available at https://github.com/LeoMenh/ResPay-payment-standard-prediction.

7 Acknowledgements

Zicheng Wang is supported in part by the Key Program of the National Natural Science Foundation of China (NSFC) under Grant No. 72495131 and the Guangdong (China) Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence [2023B1212010001]. Chenhao Ma, Yuyang Liang and Xiang Meng are supported in part by the NSFC under Grant No. 62672430, the Basic and Applied Basic Research Fund in Guangdong Province under Grant No. 2025A1515010439, Ant Group through the CCF-Ant Research Fund, the 1+1+1 CUHK-CUHK(SZ)-GDSTC Joint Collaboration Fund (including Grant No. 2025A0505000079), and the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen. This work is also supported by the Monitoring and Statistical Research

Program of the National Administration of Traditional Chinese Medicine under Grant No. 2025JCTJD3. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

8 Author contributions

Y.L., X.M. and S.Z. designed the study. Y.L. and X.M. developed the methodology, implemented the models, conducted the experiments, analyzed the results and prepared the figures. J.H. led the acquisition and curation of the DRG cohort data. Z.L., H.L., D.L., and S.Z. led the acquisition and curation of the DIP cohort data. J.H., H.L., Z.L., and D.L. contributed expertise in healthcare reimbursement and hospital management to the interpretation of the findings. Y.L., X.M., J.H., Z.L., H.L., Z.W. and C.M. drafted the manuscript. C.M. and Z.W. supervised the study and revised the manuscript. All authors reviewed the manuscript and approved the final version.

9 Competing interests

The authors declare no competing interests.

10 Ethics review

This study used retrospectively collected hospital data that had been anonymized before analysis. The study was exempt from ethics review under Article 32(2) of China's *Measures for Ethical Review of Life Sciences and Medical Research Involving Humans* (2023).

Appendix A Cohort construction and feature definitions

A.1 Prediction units and input features

Let I_u denote the input records belonging to prediction unit u, with $n_u = |I_u|$. DIP units contain individual discharge records, while DRG units contain one aggregated disease-insurance record ($n_u = 1$), which can represent multiple hospital cases. Each DIP record contains 16 features: ten expenditure proportions, class-A and class-B drug costs and utilization rates, length of stay and the current payment standard. Each DRG record contains 24 features describing current insurance-specific and disease-level standards, insurance category and share, source case count, current weight and base rate, cost composition and length of stay. Current standards are predictors; revised standards are targets.

A.2 DIP record construction

The DIP cohort used 2023 cases with valid discharge dates and disease-procedure codes and an available 2024 standard-score fee. Current and revised standards were defined by the modal standard-score fee within each year and group. Each case supplied the 16 numerical inputs in Table A1, including the current standard p_u; the following-year target y_u was excluded from the inputs. Disease–procedure codes were used only for grouping and partitioning. All eligible cases were retained without a minimum group-size requirement or exclusion based on revision magnitude.

Table A1 Record-level input features for the DIP cohort ($d = 16$). Features are listed in model input order. Expenditure shares and drug utilization rates were taken directly from the supplied dataset, without reconstructing their underlying denominators. No additional feature scaling was applied.

Position	Feature
1–10	Expenditure shares, in order: drugs, consumables, examinations, laboratory tests, bed fees, consultations, nursing, treatment, surgery and other fees.
11–12	Class-A drug expenditure and class-A drug utilization rate.
13–14	Class-B drug expenditure and class-B drug utilization rate.
15	Length of stay.
16	Current year–group modal standard-score fee (p_u).

A.3 DRG case eligibility and payment targets

The DRG cohort combined annual case records and version-specific payment schedules. Source predictors came from version 1.1 cases in 2023, 2024 and January–March 2025. Cases without a valid disease code, numerical weight or payment version were

excluded, and only disease codes represented under both versions were retained. Version 2.0 records contributed to code eligibility and fallback target weights; their clinical and cost features were excluded from model inputs.

Current and target amounts were obtained from disease- and insurance-specific schedules. Missing or non-positive amounts were replaced by the corresponding modal weight multiplied by the insurance-specific base rate estimated from the schedule; a still-missing current amount used the recorded standard fee. Eligible source cases with recognized insurance categories were aggregated into disease–insurance units, using modal payment labels and mean numerical covariates. Source-period case counts determined insurance shares, which were held fixed when aggregating standards and predictions to the disease level. The resulting dataset contained 601 training and 141 test insurance units.

Appendix B Preprocessing and analytical scales

Records lacking a required group identifier or target standard were excluded during retrospective dataset construction. Missing feature values were imputed using training-partition medians, with zero used when a median was unavailable. DRG feature-standardization and target-transformation parameters were estimated from training data; DIP inputs retained their supplied processed values without additional feature scaling.

B.1 DIP analytical scale and residual transformation

DIP used the supplied extract's processed standard-score fees as its analytical payment scale, without additional scaling of input features or payment standards. Reported errors and the LR boundary 0.45 refer to this scale; conversion to raw currency requires the original preprocessing parameters. For neural training, residuals were standardized as $(y_u - p_u - \mu_r)/s_r$, with $\mu_r = -0.08800495$ and $s_r = 0.46773300$ estimated from training prediction units. Both numerical heads used these standardized targets. Outputs were transformed back before adding the current standard and calculating evaluation metrics.

B.2 DRG payment and covariate transformations

DRG payment amounts were standardized using statistics estimated across training disease groups:

$$T(a) = \frac{a - \mu_A}{s_A}, \tag{B1}$$

where μ_A and s_A denote the mean and population standard deviation of revised amounts, respectively. The same transformation was applied to current and revised disease- and insurance-specific amounts and the current insurance-specific base-rate feature. Predictions can be returned to the original scale as $\mu_A + s_A\widehat{y}_g$; MAE and RMSE are multiplied by s_A, while R^2 is unchanged.

Weight, expenditure and length-of-stay covariates were standardized using training-case statistics before within-insurance averaging. After unit construction and

training-median imputation, all 24 input columns were standardized again using the 601 training units, with the same transformation applied to test units. These input transformations did not alter payment priors or targets on the analytical scale. DRG residual targets used $y_u - p_u$ without further residual standardization.

B.3 Numerical training targets and inverse transformation

The numerical target uses the training-unit residual mean μ_r and population standard deviation s_r for DIP, held fixed for test units:

$$z_u = \begin{cases} (y_u - p_u - \mu_r)/s_r, & \text{DIP}, \\ y_u - p_u, & \text{DRG}. \end{cases} \tag{B2}$$

DRG residuals receive no additional residual standardization. Numerical predictions are converted to analytical-scale corrections and aggregated as

$$\widehat{r}_u = \begin{cases} \mu_r + s_r \widehat{z}_u, & \text{DIP}, \\ \widehat{z}_u, & \text{DRG}, \end{cases} \qquad \widehat{r}_g = \sum_{u \in U_g} w_{gu} \widehat{r}_u. \tag{B3}$$

The weights are the same as those defining current and revised standards in Eq. (1).

Appendix C ResPay implementation and training

C.1 Representation and prediction equations

Let $\mathbf{f}_i \in \mathbb{R}^d$ denote the preprocessed feature vector of record i. A record encoder E maps each record to a vector $\mathbf{h}_i = E(\mathbf{f}_i)$. An attention network A assigns a scalar score to each record representation. These scores are normalized within each unit to obtain attention weights and a pooled representation:

$$\alpha_{iu} = \frac{\exp A(\mathbf{h}_i)}{\sum_{j \in I_u} \exp A(\mathbf{h}_j)}, \qquad \overline{\mathbf{h}}_u = \sum_{i \in I_u} \alpha_{iu} \mathbf{h}_i. \tag{C4}$$

The weights sum to one within each unit and determine each record's contribution to the pooled representation.

In parallel, ResPay summarizes the preprocessed input features using their means and population standard deviations across records in the unit, together with $\log(1 + n_u)$. Let $\mathbf{s}_u$ collect these summaries. A summary encoder E_s maps $\mathbf{s}_u$ to a learned vector. This vector is concatenated with the attention-pooled representation and passed through a linear transformation followed by a rectified linear unit (ReLU) activation:

$$\mathbf{x}_u = \mathrm{ReLU}\big(W_f[\overline{\mathbf{h}}_u, E_s(\mathbf{s}_u)] + \mathbf{b}_f\big)\,. \tag{C5}$$

Here, brackets denote concatenation, and W_f and $\mathbf{b}_f$ are learned fusion weights and bias. The transformation produces the unit representation $\mathbf{x}_u$, which is used by both

the numerical predictor and the LR classifier. All representation-network parameters are shared across prediction units within each hospital's model, while each unit representation is constructed from that unit's own records.

A record head H_r takes the record encoding and its unit representation, while a unit head H_u takes the unit representation alone:

$$\widehat{z}_u = \frac{1}{n_u} \sum_{i \in I_u} H_r([\mathbf{h}_i, \mathbf{x}_u]) + H_u(\mathbf{x}_u). \tag{C6}$$

The record outputs are averaged equally; attention is used only in representation pooling. Outputs are restored and aggregated using Eq. (B3).

A separate head produces a unit-level classification logit $\ell_u = H_{\mathrm{LR}}(\mathbf{x}_u)$. The group-level score is

$$q_g = \sigma\left(\sum_{u \in U_g} w_{gu} \ell_u \right), \tag{C7}$$

where σ is the logistic sigmoid. DRG insurance-specific logits are aggregated before the sigmoid; DIP groups contain a single unit. The head shares the representation network with the numerical predictor but has its own prediction parameters.

C.2 Network dimensions and numerical operations

Table C2 gives the complete layer configuration. Each arrow denotes an affine layer with bias; ReLU is applied after every hidden layer, with the attention hidden layer using tanh instead. Numerical outputs and classification logits are linear scalar outputs. There are 40,836 trainable parameters for DIP and 42,372 for DRG. Both prediction tasks use the same architecture with their respective input dimensions.

Table C2 Complete ResPay architecture. DIP has $d = 16$ inputs and DRG has $d = 24$. Unit summaries contain feature means, population standard deviations and $\log(1 + n_u)$.

Component	Layer widths and activation
Record encoder	$d \to 64 \to 64$; ReLU after both layers.
Attention scorer	$64 \to 64 \to 1$; tanh after the hidden layer.
Summary encoder	$(2d + 1) \to 64 \to 64$; ReLU after both layers.
Fusion layer	$128 \to 64$; ReLU.
Record numerical head	$128 \to 64 \to 1$; ReLU after the hidden layer.
Unit numerical head	$64 \to 64 \to 1$; ReLU after the hidden layer.
LR classification head	$64 \to 64 \to 1$; ReLU after the hidden layer.

Attention scores are normalized only over records within the same prediction unit. For stability, the maximum unit score is subtracted before exponentiation, and the denominator is bounded below by 10^{-12}. Population feature variance is bounded below by 10^{-8} before taking its square root. Consequently, a constant feature contributes

a standard-deviation summary of 10^{-4}. For a DRG unit containing one aggregate row, the attention weight is one, all standard-deviation summaries are 10^{-4} and the count summary is $\log 2$. The source case count represented by that aggregate is supplied separately as feature 4. The final numerical output is the equally weighted mean of record-head outputs plus the unit-head output; record attention is used for representation pooling rather than this numerical mean.

C.3 Loss definitions

Let M denote the number of training prediction units, G the number of training payment groups, and $N = \sum_{u=1}^{M} n_u$ the total number of training input records. The main numerical loss combines absolute and squared errors on the training target scale defined in Eq. (B2):

$$\mathcal{L}_{\text{unit}} = \frac{1}{M} \sum_{u=1}^{M} \left[|\widehat{z}_u - z_u| + 0.2\,(\widehat{z}_u - z_u)^2 \right]. \tag{C8}$$

Each prediction unit receives equal weight in this loss.

An auxiliary loss provides direct supervision for the record-level numerical head by comparing each record's output with its unit's residual target:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N} \sum_{u=1}^{M} \sum_{i \in I_u} \rho_{0.5}(H_r([\mathbf{h}_i, \mathbf{x}_u]) - z_u)\,. \tag{C9}$$

Here, ρ_β is the Smooth L1 loss, defined as $\rho_\beta(e) = e^2/(2\beta)$ when $|e| < \beta$ and $\rho_\beta(e) = |e| - \beta/2$ otherwise, with $\beta = 0.5$. Each record inherits the target of its prediction unit, so this auxiliary supervision concerns revisions to unit-level standards. The loss is averaged over individual discharge records in DIP and aggregated disease–insurance records in DRG. Consequently, DIP units with more cases contribute more terms to this auxiliary loss.

For DRG, an additional aggregation loss aligns insurance-weighted numerical predictions with the observed disease-level revisions:

$$\mathcal{L}_{\text{agg}} = \frac{1}{G} \sum_{g=1}^{G} \rho_{0.5} \left(\sum_{u \in U_g} w_{gu} \widehat{z}_u - (y_g - p_g) \right). \tag{C10}$$

The weights w_{gu} are the source-period insurance shares used to define the group-level standards. Because DRG residual targets do not undergo additional residual standardization, the aggregated predictions and observed group-level revisions are expressed on the same analytical scale. Each disease group receives equal weight in this loss. The aggregation loss is omitted for DIP.

The LR classifier is trained using weighted binary cross-entropy:

$$\mathcal{L}_{\mathrm{LR}} = -\frac{1}{G}\sum_{g=1}^{G}\left[w_{+}H_g \log q_g + (1-H_g)\log(1-q_g)\right], \tag{C11}$$

where H_g is the observed LR label defined in Eq. (2). The positive-class weight is calculated from the training partition as $w_{+} = \sum_{g=1}^{G}(1-H_g)/\sum_{g=1}^{G} H_g$, the ratio of non-LR to LR payment groups. For numerical stability, the implementation evaluates this loss directly from the group-level logits underlying q_g.

The losses are combined as in Eq. (4), with $\lambda_{\mathrm{agg}} = 1$ for DRG and zero for DIP. Numerical supervision includes all training units, irrespective of LR labels or classification scores.

C.4 Optimization and weighting

Each epoch performs one forward pass over all training rows, computes the joint objective in Eq. (4), backpropagates once and applies one AdamW update. Unit regression assigns equal weight to prediction units; the record auxiliary loss assigns equal weight to input rows; DRG aggregation and LR classification assign equal weight to disease groups. No sampling by case count or LR status is used. The LR positive-class weight is $(G-G_{+})/G_{+}$, giving $793/84 = 9.440476$ for DIP and $250/28 = 8.928571$ for DRG. For DRG, source insurance shares aggregate logits before the weighted binary cross-entropy, and aggregate numerical outputs before the disease-level Smooth L1 loss.

Appendix D Evaluation calculations and experimental protocols

D.1 Repeated training and error comparisons

For ResPay, identification and numerical prediction metrics were calculated separately for each of the five training seeds and summarized using their mean and sample standard deviation. The held-out groups remained fixed across runs, so the reported variability describes variation across repeated model training on the same evaluation cohort. Relative reductions in MAE and RMSE were calculated from the mean metrics across seeds. For either error metric, let $\overline{m}_{\mathrm{base}}$ and $\overline{m}_{\mathrm{ResPay}}$ denote the baseline and ResPay means, respectively. The relative reduction was

$$\text{Relative reduction (\%)} = 100\frac{\overline{m}_{\mathrm{base}} - \overline{m}_{\mathrm{ResPay}}}{\overline{m}_{\mathrm{base}}}. \tag{D12}$$

Error bars for these comparisons represent the sample standard deviation of within-seed metric differences, divided by the baseline's mean metric and multiplied by 100.

To examine how prediction differences were distributed across payment groups, we compared absolute errors within each seed and then averaged these differences across

seeds. Let $\widehat{y}_{g,s}^{\text{base}}$ denote the baseline prediction used in the comparison with ResPay run s. For a baseline represented by a fixed prediction, this value is the same across runs. The mean difference for group g was

$$d_g = \frac{1}{5}\sum_{s=1}^{5}\left[\left|\widehat{y}_{g,s}^{\text{base}} - y_g\right| - \left|\widehat{y}_{g,s}^{\text{ResPay}} - y_g\right|\right]. \tag{D13}$$

Positive values indicate lower mean absolute error for ResPay. A group was counted as improved when $d_g > 10^{-6}$. Group-level error bars show the sample standard deviation of the corresponding within-seed absolute-error differences.

D.2 Revision strata and exposure calculations

For descriptive analyses, held-out groups were stratified according to their observed absolute revisions. Small revisions had magnitudes below 0.10 in DIP or 0.05 in DRG. Moderate revisions were at least these lower boundaries but below the corresponding LR thresholds defined in Eq. (2). Large revisions met or exceeded the LR thresholds. These strata were used to describe the distribution of payment groups, hospital cases and revision exposure.

Case-weighted absolute revision exposure was calculated by multiplying each group's absolute revision by its source-period case count and summing across the relevant groups. Case counts refer to the underlying hospital cases, including those represented by aggregated DRG input records. DIP exposure was calculated on the analytical payment-standard scale and reported in case-weighted standardized fee units. For DRG, current and revised insurance-specific amounts on the original RMB scale were first aggregated using the same source-period insurance shares, after which the absolute disease-level revision was multiplied by the disease group's case count. Case counts and insurance shares were held fixed when comparing current and revised standards. The LR exposure share was calculated as exposure summed over observed LR groups divided by exposure summed over all held-out groups.

D.3 Direct-prediction ablation

The direct-prediction variant retained current standards as inputs, the network architecture, training settings, LR labels and gating threshold. All model components were retrained jointly.

Let $\widehat{y}_g^{\text{direct,raw}}$ denote the directly predicted revised standard before gating, expressed on the analytical payment-standard scale. For DRG, this prediction was obtained by aggregating insurance-specific predictions using the same source-period insurance shares as ResPay. Let q_g^{direct} denote the classification score produced by the retrained variant, using the logit aggregation in Eq. (C7). Its final prediction was

$$\widehat{y}_g^{\text{direct}} = p_g + \mathbf{1}\{q_g^{\text{direct}} \geq t\}\left(\widehat{y}_g^{\text{direct,raw}} - p_g\right). \tag{D14}$$

Thus, groups selected by the classifier received the directly predicted revised standard, while the remaining groups retained their current standards. Because the representation network and classifier were jointly retrained, the groups selected by the gate could differ between variants. The comparison therefore assessed the residual formulation as implemented, including its associated target transformation, within the jointly trained framework.

Both variants were evaluated on the same held-out payment groups across five corresponding training seeds. MAE, RMSE and R^2 were calculated separately for each seed using the final gated predictions across all held-out groups. Figure 3a,b compares the mean MAE and RMSE of the residual and direct-target variants.

D.4 Classification-score threshold sensitivity

For each cohort and training seed, the trained model, numerical corrections and classification scores were held fixed while the gate threshold t varied from 0 to 1 in increments of 0.05. Final gated predictions were evaluated across all held-out payment groups using MAE and RMSE and summarized across five seeds. Figure 3c–f displays thresholds 0, 0.25, 0.5, 0.75 and 0.95. The empirical sensitivity experiment selected $t = 0$ for DIP and $t = 0.75$ for DRG; these settings were used in the main numerical comparisons and direct-prediction ablation. DIP therefore applied all corrections, while DRG retained current standards below 0.75. In DRG, the threshold was chosen on the basis of MAE; RMSE was identical at thresholds 0 and 0.75 to four decimal places.

References

[1] Zou, K., Li, H.-Y., Zhou, D. & Liao, Z.-J. The effects of diagnosis-related groups payment on hospital healthcare in China: a systematic review. *BMC Health Serv. Res.* **20**, 112 (2020).

[2] Fetter, R. B., Shin, Y., Freeman, J. L., Averill, R. F. & Thompson, J. D. Case mix definition by diagnosis-related groups. *Med. Care* **18**, iii, 1–53 (1980).

[3] Qian, M., Zhang, X., Chen, Y., Xu, S. & Ying, X. The pilot of a new patient classification-based payment system in China: the impact on costs, length of stay and quality. *Soc. Sci. Med.* **289**, 114415 (2021).

[4] Shenzhen Municipal Healthcare Security Bureau. Implementation rules for diagnosis-intervention packet (DIP) payment of inpatient medical expenses in Shenzhen (trial) (2023). URL https://www.sz.gov.cn/cn/xxgk/zfxxgj/zcfg/content/post_10821050.html. Issued August 25, 2023. Articles 31–32. In Chinese; title translated.

[5] Jinan Municipal Healthcare Security Bureau. Interim measures for diagnosis-related group (DRG) payment in Jinan (2021). URL https://ybj.jinan.gov.cn/col38569/art/2021/art_38569_4760874.html. Issued December 24, 2021. Document No. 15 [2021]. Articles 17–20. In Chinese; title translated.

[6] National Healthcare Security Administration. CHS-DRG grouping and payment technical specifications. Tech. Rep., National Healthcare Security Administration (2019). URL https://www.nhsa.gov.cn/module/download/downfile.jsp?filename=a3cbb51dc6354dd4b6a5ab09bec18121.pdf. In Chinese; title translated. See Sections 4.3–4.4.

[7] National Healthcare Security Administration. Press conference on the DRG/DIP version 2.0 grouping schemes (2024). URL https://www.nhsa.gov.cn/art/2024/7/23/art_14_13318.html. 23 July 2024. In Chinese; title translated.

[8] Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 18 (2018).

[9] Mani, K. A. *et al.* Multimodal machine learning for risk-stratified bundled payments in spinal surgery. *npj Digit. Med.* **8**, 512 (2025).

[10] Andriola, C. *et al.* A novel machine learning algorithm for creating risk-adjusted payment formulas. *JAMA Health Forum* **5**, e240625 (2024).

[11] Wang, H., Gao, C., Dantona, C., Hull, B. & Sun, J. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *npj Digit. Med.* **7**, 16 (2024).

[12] Liu, J., Capurro, D., Nguyen, A. & Verspoor, K. Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. *npj Digit. Med.* **4**, 103 (2021).

[13] Bertsimas, D. *et al.* Algorithmic prediction of health-care costs. *Oper. Res.* **56**, 1382–1392 (2008).

[14] Drewe-Boss, P., Enders, D., Walker, J. & Ohler, U. Deep learning for prediction of population health costs. *BMC Med. Inform. Decis. Mak.* **22**, 32 (2022).

[15] Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80, 2127–2136 (2018). URL https://proceedings.mlr.press/v80/ilse18a.html.

[16] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (2019). URL https://openreview.net/forum?id=Bkg6RiCqY7.

[17] Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).

[18] Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).