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.

Clinical coding capacity and hospital-reported quality across hospital tiers during China's transition to DRG-based payment: a cross-sectional survey of 299 hospitals in Anhui Province

Lijuan Ma
The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital of Wannan Medical University)

Zhengyu Zhang
The First Affiliated Hospital, Zhejiang University School of Medicine

Yichao Wu
The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital of Wannan Medical University)

Dandan Ge
The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital of Wannan Medical University)

Yuanying Yao
The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital of Wannan Medical University)

Xiaogan Jiang
yjsicu@126.com

The First Affiliated Hospital of Wannan Medical University (Yijishan Hospital of Wannan Medical University)

Research Article

Keywords: Clinical coding, ICD-10, Diagnosis-related groups, Case-mix payment, Health information management, Health workforce, Hospital tier, Health equity, China

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11199977/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background Diagnosis-related group (DRG) payment relies on coded medical records to assign cases and determine reimbursement. China has rapidly expanded case-based payment, but the coding workforce and infrastructure available across hospital tiers remain poorly described.

Methods We conducted a province-wide cross-sectional survey of hospital medical record departments in Anhui Province, China, in June and July 2026; quantitative responses referred to the year ending 31 December 2025. Participation was required for secondary and higher-level hospitals and voluntary for primary and unclassified institutions. Of 311 returned questionnaires, 299 were retained. Measures covered coding workforce, certification and training, workload, coding software, coding dictionaries, physician participation, hospital-reported coding accuracy, and reported error types. Hospitals were grouped by official tier, and tier comparisons used chi-square tests.

Results The 299 hospitals employed 1,470 coders. All measured capacity indicators declined across descending hospital tiers. National coder certification was present in 93.5% of Tier 3A hospitals and 21.6% of Tier 1 and unclassified institutions; a regularly updated coding dictionary in 74.2% and 21.6%, respectively; and coding assistance software rated as having good functionality in 58.1% and 5.4% (all $P < 0.001$). Overall, 54.8% of hospitals had no coding assistance software and 24.7% had no coding dictionary. Hospital-reported primary diagnosis accuracy of at least 95% was reported by 83.9% of Tier 3A hospitals and 48.6% of Tier 1 and unclassified institutions. The most frequently reported problems were principal diagnosis selection errors (72.2%), imprecise or non-specific coding (68.9%), and omission of complications, comorbidities, or other secondary diagnoses (60.2%).

Conclusions Clinical coding capacity showed a marked tier gradient, whereas the gradient in hospital-reported performance was shallower. Because coding quality and reimbursement were not independently audited, the study cannot establish payment effects. Nevertheless, uneven coding capacity may disadvantage lower-tier institutions under uniform case-based payment rules. Coding readiness should be monitored during payment reform, with targeted training, shared technical infrastructure, and proportionate audit support.

Background

Case-based hospital payment changes what a medical record is for. Under diagnosis-related group (DRG) systems, the codes assigned to a discharge determine the group the case falls into, the weight attached to that group, and therefore the revenue the hospital receives. A record that was previously a clinical and legal document becomes, additionally, a financial instrument. This shift makes the accuracy and completeness of clinical coding a structural property of the payment system rather than an administrative detail [1, 2].

China has moved through this transition unusually quickly. Beginning with parallel pilots of DRG and the diagnosis-intervention packet (DIP) in 101 cities in 2020, case-based payment was extended to more

than 200 cities within two years [3]. By 2025, DRG/DIP payment had achieved near-complete coverage of pooled insurance regions nationwide [16]. He characterised the reform as "scaling-up through piloting" and noted that the dual-track design partly reflected the higher demands that DRG places on data quality and technical capacity [3]. Capability was therefore recognised as heterogeneous at the design stage, but its distribution across hospital tiers has not been established.

Coding is imperfect in every health system that has been measured. Systematic reviews of discharge coding accuracy report principal diagnosis accuracy typically between 65% and 96% with wide inter-institutional variation, and identify omission of secondary diagnoses and use of non-specific codes as recurrent problems [4, 5]. Studies using dually coded or re-abstracted records confirm that accuracy varies substantially by condition and by the specificity demanded of the coder [6–8]. Coding error is therefore not an anomaly to be eliminated but a quantity to be managed, and the relevant policy question is what determines its magnitude.

The evidence on that question points consistently away from individual coder ability and towards organisational conditions. In an Australian multi-hospital study, Santos and colleagues found that coding accuracy was shaped by staffing levels, throughput pressure, time available per record, coder career structure and the quality of collaboration with clinicians, rather than by hospital location or specialty [9]. An international survey of coding training found no common standard for coder qualification across countries, with substantial variation in the duration, content and mandatory status of training [10]. Evidence from China is consistent with this organisational reading: a meta-analysis of 11 studies covering 1783 hospitals estimated a mean workload of approximately 11 300 records per coder per year, which the authors judged incompatible with careful coding [11]; and a prospective study in a Chinese tertiary hospital found that structured physician–coder collaboration raised coding accuracy from 78.9% to 87.1%, while also showing that 89.9% of coding errors originated in poor-quality front-page documentation rather than in the coding step itself [12].

Two further features of the Chinese context raise the stakes. First, coding personnel are frequently drawn from nursing, statistics or administrative backgrounds rather than from health information management, so certification and in-service training carry more of the burden of competence than they would in systems with a dedicated coding profession. Second, the financial incentives created by case-based payment cut in both directions. Zhou and colleagues, auditing 5157 records from 28 hospitals in a DRG pilot, found that 12.9% had been upcoded, and identified tertiary hospital status as a risk factor [13]. Coding distortion under DRG is therefore not solely a capability problem at the bottom of the system; it is also an incentive problem at the top, and the two have opposite signs.

Despite this, systematic data on coding capacity in China at scale are scarce. Published work consists predominantly of single-institution audits or quality-improvement reports, which can characterise accuracy in one setting but cannot describe how capability is distributed across a health system [11, 12]. A recent dual-centre study documented increasing coding depth and specificity in tertiary hospitals during DRG reform but did not assess workforce or infrastructure across hospital tiers [15]. Studies that

examine accuracy also rarely measure the certification, training, workload, software, and reference tools that may explain institutional variation. The distribution of these inputs is relevant to whether a uniform payment rule can be implemented consistently across a heterogeneous hospital system.

We therefore surveyed the coding workforce and coding infrastructure of participating hospitals in one Chinese province, alongside hospital-reported coding performance. Our objectives were: (1) to describe the coding workforce and institutional support available across hospital tiers; (2) to describe hospital-reported coding accuracy and reported error types; and (3) to compare the tier gradients in structural capacity and reported performance. We further considered the implications of these descriptive patterns for implementation of case-based payment, while recognising that the survey did not measure reimbursement outcomes.

Methods

Study design and setting

This was a cross-sectional survey conducted as part of the 2026 Anhui Provincial Baseline Survey of Medical Record Management, organised by the Anhui Provincial Medical Record Quality Control Centre. Anhui is an eastern Chinese province of approximately 61 million people comprising 16 prefecture-level cities, with tertiary referral centres, secondary district hospitals, primary institutions, and specialty facilities. Reporting follows the STROBE statement for cross-sectional studies.

Participants

Participation was required for all secondary and higher-level hospitals in the province and encouraged for primary and unclassified institutions. Eligible institutions included general hospitals, specialty hospitals, maternal and child health hospitals, and other clinical facilities. Hospitals of traditional Chinese medicine were administered through a separate quality-control system and were not included. A total of 311 questionnaires were returned. After logical validation, duplicate submissions and records with irreconcilable internal inconsistencies were removed, leaving 299 hospitals (96.1% of returned questionnaires) for analysis. Because participation by primary and unclassified institutions was voluntary, this proportion is not a population response rate. No a priori sample-size calculation was performed because the survey sought participation across the provincial hospital system.

Hospitals were stratified into four groups using the official Chinese hospital classification: Tier 3A (the highest grade, n = 31), Tier 3 comprising Tier 3B and unclassified tertiary hospitals (n = 74), Tier 2 comprising Tier 2A, 2B and unclassified secondary hospitals (n = 157), and Tier 1 and unclassified institutions (n = 37).

Survey instrument and variables

The coding module of the survey instrument was developed by the provincial quality control centre with reference to national coding guidance and the ICD-10 and ICD-9-CM-3 classifications. It captured four domains.

Coding workforce. Total and full-time coder headcount; coder age band; years of coding experience; educational and professional background; current professional title; ICD-10 coding competency certification at provincial and national level; ICD coding trainer certification; number of ICD training sessions attended; and annual coding volume per coder.

Hospital-reported coding quality. Three indicators were reported in closed ordinal bands: principal diagnosis selection accuracy and principal procedure coding accuracy (≥ 95%, 85–94%, 70–84%, < 70%), and secondary diagnosis omission rate (≤ 3%, 4–8%, 9–15%, > 15%). These institutional self-assessments drew on each hospital's internal audit records and national front-page data-quality standards. Audit sampling and verification procedures were not standardised across hospitals, and the study team did not independently re-abstract records.

Reported error types. A multiple-select item covered principal diagnosis selection errors, principal procedure coding errors, omission of complications, comorbidities, or other secondary diagnoses, imprecise or non-specific coding, mixing of old and new code versions, and an explicit "no obvious coding problems" option.

Institutional coding support. Availability and functionality of coding assistance or intelligent coding software (available with good, moderate or poor functionality, or not available); status of the institutional coding dictionary (established and regularly updated, established but rarely updated, simple summary table only, or none); and whether physicians participate in the coding process.

Data collection and quality control

Data were collected between June and July 2026 through the Wenjuanxing online survey platform. The questionnaire was completed by the head of the medical record department or the staff member with the most complete knowledge of coding operations at each institution. All quantitative responses referred to the calendar year ending 31 December 2025. Municipal quality control centres monitored completion within their jurisdictions. The provincial centre reviewed every submission and conducted telephone follow-up where responses were internally inconsistent or contained implausible values.

Statistical analysis

Analyses were performed in Excel 2021 and SPSS 26.0 (IBM Corp., Armonk, NY, USA). Categorical variables are summarised as counts and percentages and skewed continuous variables as medians with interquartile ranges (IQR). Differences across the four hospital groups were tested with Pearson chi-square tests applied to the highest response category versus all other categories, thereby avoiding sparse cells in full ordinal cross-tabulations. Tests were two-sided, with $P < 0.05$ considered statistically

significant. Because the analysis was descriptive, the comparisons do not estimate independent or causal effects of individual capacity indicators.

To compare how steeply structural capacity and reported performance vary across the hospital system, we calculated for each indicator the ratio of the Tier 3A value to the Tier 1 value. This is a descriptive contrast of gradient steepness across indicators measured on a common percentage scale; it is not a statistical test and no inference is drawn from the ratio itself.

Ethical considerations

The survey was conducted as an institution-level administrative quality-improvement activity under the authority of the provincial quality-control centre. It collected operational information from authorised institutional representatives and did not collect identifiable information about patients or individual respondents. Under this administrative framework, the need for ethics approval and informed consent was waived by the Ethics Committee of The First Affiliated Hospital of Wannan Medical University in accordance with Article 32 of the Measures for Ethical Review of Life Sciences and Medical Research Involving Humans (Order No. 4 [2023] of the National Health Commission Science and Education Department).

Results

Coding workforce

The 299 hospitals employed 1470 coding personnel, of whom 1063 (72.3%) worked full time and 407 (27.7%) part time. The median number of full-time coders per hospital was 3 (IQR 1–5), ranging from a median of 5 (IQR 3–8) in Tier 3A hospitals to 1 (IQR 0–2) in Tier 1 and unclassified institutions.

Workforce composition is shown in Table 1. Coders were concentrated in early and mid career: 704 (47.9%) were aged 31–40 years and 418 (28.4%) were under 30, while 478 (32.5%) had two years or less of coding experience and only 256 (17.4%) had more than ten years. Professional origins were heterogeneous, with nursing the single largest background (387, 26.3%), followed by health information management and clinical medicine (249, 16.9% each); 556 coders (37.8%) held a professional title in health information technology, the remainder holding titles in nursing, statistics, hospital administration or clinical medicine.

Table 1
Characteristics of the clinical coding workforce (n = 1470 coders in 299 hospitals)

Characteristic	Coders (n)	Proportion (%)
Employment status		
Full-time	1063	72.3
Part-time	407	27.7
Age group (years)		
20–30	418	28.4
31–40	704	47.9
41–50	155	10.5
≥51	193	13.1
Coding experience		
≤2 years	478	32.5
3–5 years	421	28.6
>5–10 years	315	21.4
>10 years	256	17.4
Educational/professional background		
Health information management	249	16.9
Clinical medicine	249	16.9
Nursing	387	26.3
Health statistics	178	12.1
Hospital administration	120	8.2
Other	287	19.5
Current professional title		
Health information technology	556	37.8
Nursing	334	22.7
Statistics	237	16.1
Hospital administration	182	12.4
Clinical medicine	107	7.3

Percentages are of the 1470 coders and may not sum to exactly 100.0 because of rounding.

Characteristic	Coders (n)	Proportion (%)
Other	54	3.7
Percentages are of the 1470 coders and may not sum to exactly 100.0 because of rounding.		

The median annual coding volume was 5,200 records per coder (IQR 1,500 – 10,000); the arithmetic mean was 7,900. The median was 11,000 in Tier 3A hospitals and 4,600 in Tier 2 hospitals.

Certification and training across hospital tiers

Certification and training are summarised in Table 2. Nationally certified ICD-10 coders were present in 164 hospitals (54.8%) and provincially certified coders in 206 (68.9%); 90 hospitals (30.1%) had a certified coding trainer, and 223 (74.6%) had at least one coder who had attended a formal ICD training session.

Table 2
Coder certification and training by hospital tier, % (n)

Indicator	All hospitals (n = 299)	Tier 3A (n = 31)	Tier 3 (n = 74)	Tier 2 (n = 157)	Tier 1 (n = 37)	χ^2	*P*
National ICD-10 coding certification	54.8 (164)	93.5 (29)	77.0 (57)	44.6 (70)	21.6 (8)	56.62	< 0.001
Provincial ICD-10 coding certification	68.9 (206)	96.8 (30)	89.2 (66)	62.4 (98)	32.4 (12)	51.49	< 0.001
ICD coding trainer certification	30.1 (90)	67.7 (21)	44.6 (33)	20.4 (32)	10.8 (4)	41.86	< 0.001
≥ 1 coder attended ICD training	74.6 (223)	96.8 (30)	89.2 (66)	70.1 (110)	45.9 (17)	34.08	< 0.001
Values are the percentage (and number) of hospitals in which the indicator was present. χ^2 tests compare the four tiers. Tier 3 comprises Tier 3B and unclassified tertiary hospitals; Tier 2 comprises Tier 2A, 2B and unclassified secondary hospitals; Tier 1 comprises primary and unclassified institutions.							

Every certification and training indicator declined monotonically across descending tiers. National certification was present in 93.5% of Tier 3A hospitals, 77.0% of Tier 3, 44.6% of Tier 2 and 21.6% of Tier 1 institutions ($\chi^2 = 56.62$, $P < 0.001$). The gradient was steepest for trainer certification, present in 67.7% of Tier 3A but 10.8% of Tier 1 hospitals ($\chi^2 = 41.86$, $P < 0.001$) — that is, the institutions least able to send staff away for training were also those least able to train staff internally.

Institutional coding support across hospital tiers

Institutional coding support by hospital tier are shown in Table 3. Coding assistance software rated as having good functionality was available in 67 hospitals (22.4%), while 164 (54.8%) had no coding assistance software. A regularly updated institutional coding dictionary was maintained by 131 hospitals (43.8%), whereas 74 (24.7%) had no coding dictionary. Physicians participated in coding in 183 hospitals (61.2%).

Table 3
Institutional coding support by hospital tier, % (n)

Support system	**All hospitals (n = 299)**	**Tier 3A (n = 31)**	**Tier 3 (n = 74)**	**Tier 2 (n = 157)**	**Tier 1 (n = 37)**
Coding assistance software					
Available, good functionality	22.4 (67)	58.1 (18)	28.4 (21)	16.6 (26)	5.4 (2)
Available, moderate functionality	16.7 (50)	16.1 (5)	18.9 (14)	16.6 (26)	13.5 (5)
Available, poor functionality	6.0 (18)	3.2 (1)	6.8 (5)	6.4 (10)	5.4 (2)
Not available	54.8 (164)	22.6 (7)	45.9 (34)	60.5 (95)	75.7 (28)
Institutional coding dictionary					
Established, regularly updated	43.8 (131)	74.2 (23)	52.7 (39)	38.9 (61)	21.6 (8)
Established, rarely updated	17.1 (51)	12.9 (4)	17.6 (13)	17.8 (28)	16.2 (6)
Simple coding summary table only	14.4 (43)	6.5 (2)	10.8 (8)	14.6 (23)	27.0 (10)
No dictionary	24.7 (74)	6.5 (2)	18.9 (14)	28.7 (45)	35.1 (13)
Physician participation in coding	61.2 (183)	87.1 (27)	74.3 (55)	54.1 (85)	43.2 (16)
Comparison across tiers of the highest category: coding assistance software rated as having good functionality, chi-square = 33.42, P < 0.001; coding dictionary established and regularly updated, chi-square = 22.97, P < 0.001; physician participation, chi-square = 22.44, P < 0.001.					

The tier gradient was again monotonic and was steepest for software rated as having good functionality: 58.1% of Tier 3A hospitals, 28.4% of Tier 3 hospitals, 16.6% of Tier 2 hospitals, and 5.4% of Tier 1 and unclassified institutions (chi-square = 33.42, P < 0.001). Regularly updated dictionaries followed the same ordering (74.2%, 52.7%, 38.9%, and 21.6%; chi-square = 22.97, P < 0.001), as did physician participation (87.1%, 74.3%, 54.1%, and 43.2%; chi-square = 22.44, P < 0.001).

Human and technical deficits were concentrated in the same hospital tiers. Among Tier 1 and unclassified institutions, 75.7% had no coding software and 35.1% had no coding dictionary; among Tier 2 hospitals, the corresponding proportions were 60.5% and 28.7%. These tier-level marginal distributions

do not establish that deficits co-occurred within individual hospitals, but they show that lower-tier groups had less access to both human and technical support.

Hospital-reported coding quality

Self-reported quality indicators are shown in Table 4. Across all hospitals, 183 (61.2%) reported primary diagnosis selection accuracy of at least 95%, 86 (28.8%) reported 85–94%, 22 (7.4%) reported 70–84% and 8 (2.7%) reported below 70%. Primary procedure coding accuracy was similarly distributed, with 176 hospitals (58.9%) reporting at least 95%. Secondary diagnosis omission was reported as ≤ 3% by 161 hospitals (53.8%), 4–8% by 96 (32.1%), 9–15% by 29 (9.7%) and above 15% by 13 (4.3%); that is, 42 hospitals (14.0%) reported omitting more than one in eleven secondary diagnoses.

Table 4
Hospital-reported coding quality indicators by hospital tier, % (n)

Quality indicator	All hospitals (n = 299)	Tier 3A (n = 31)	Tier 3 (n = 74)	Tier 2 (n = 157)	Tier 1 (n = 37)
Primary diagnosis selection accuracy					
≥95%	61.2 (183)	83.9 (26)	70.3 (52)	55.4 (87)	48.6 (18)
85–94%	28.8 (86)	12.9 (4)	25.7 (19)	31.2 (49)	37.8 (14)
70–84%	7.4 (22)	3.2 (1)	2.7 (2)	10.2 (16)	8.1 (3)
<70%	2.7 (8)	0.0 (0)	1.4 (1)	3.2 (5)	5.4 (2)
Primary procedure coding accuracy					
≥95%	58.9 (176)	80.6 (25)	67.6 (50)	53.5 (84)	45.9 (17)
85–94%	30.1 (90)	16.1 (5)	27.0 (20)	32.5 (51)	37.8 (14)
70–84%	8.0 (24)	3.2 (1)	4.1 (3)	10.2 (16)	10.8 (4)
<70%	3.0 (9)	0.0 (0)	1.4 (1)	3.8 (6)	5.4 (2)
Secondary diagnosis omission rate					
≤3%	53.8 (161)	71.0 (22)	60.8 (45)	50.3 (79)	40.5 (15)
4–8%	32.1 (96)	22.6 (7)	29.7 (22)	33.8 (53)	37.8 (14)
9–15%	9.7 (29)	6.5 (2)	5.4 (4)	12.1 (19)	10.8 (4)
>15%	4.3 (13)	0.0 (0)	4.1 (3)	3.8 (6)	10.8 (4)

Indicators are hospital self-assessments against national front-page data-quality standards, not independent re-abstraction. Chi-square tests compare the highest category across the four groups: principal diagnosis accuracy ≥ 95%, chi-square = 13.94, $P < 0.01$; principal procedure accuracy ≥ 95%, chi-square = 12.80, $P < 0.01$; secondary diagnosis omission ≤ 3%, chi-square = 8.52, $P < 0.05$.

Reported quality also fell across tiers. Primary diagnosis accuracy of at least 95% was reported by 83.9% of Tier 3A, 70.3% of Tier 3, 55.4% of Tier 2 and 48.6% of Tier 1 hospitals ($\chi^2 = 13.94$, $P < 0.01$), with comparable gradients for procedure accuracy ($\chi^2 = 12.8$, $P < 0.01$) and secondary diagnosis omission ($\chi^2 = 8.52$, $P < 0.05$).

Reported error types

The distribution of reported error types is shown in Fig. 1. Principal diagnosis selection errors were reported by 216 hospitals (72.2%), imprecise or non-specific coding by 206 (68.9%), omission of complications, comorbidities and other secondary diagnoses by 180 (60.2%) and principal procedure

coding errors by 165 (55.2%). Mixing of old and new code versions was reported by 65 hospitals (21.7%). Only 36 hospitals (12.0%) reported no obvious coding problems.

The three most frequently reported problems have direct implications for DRG grouping. The principal diagnosis influences the major diagnostic category and base group; complications and comorbidities can affect severity variants; and non-specific codes may prevent assignment to a more specific group. Depending on the coding error and grouping rules, these problems can alter the recorded case mix. The survey did not link individual errors to DRG assignment or reimbursement.

Capacity gradient versus reported performance gradient

Figure 2 places selected capacity and reported-performance indicators on a common percentage scale. From Tier 3A to Tier 1 and unclassified institutions, the ratio was 10.8 for software rated as having good functionality, 4.3 for national coder certification, and 3.4 for a regularly updated dictionary. The corresponding ratios were 1.7 for reported principal diagnosis accuracy, 1.8 for reported principal procedure accuracy, and 1.8 for reported control of secondary diagnosis omission. These ratios are descriptive and are sensitive to the value in the lowest-tier group.

The shallower gradient in reported performance could reflect diminishing returns to additional capacity, but it could also reflect differences in how hospitals audit and detect errors. Because audit procedures were not standardised and no independent re-abstraction was performed, the survey cannot distinguish these explanations.

Hospitals' own improvement priorities

Asked to select priorities for improvement, hospitals most often chose strengthening physician training in coding principles (245, 81.9%), expanding provincial-level coding training (229, 76.6%), requiring national or provincial coder certification (220, 73.6%) and conducting routine coding quality audits (204, 68.2%). Only 54 hospitals (18.1%) selected "no suggestions". Notably, the two most-selected priorities concern the supply of training from outside the institution rather than actions the hospital can take alone, which is consistent with the capacity constraints documented above.

Discussion

Principal findings

In this survey of 299 hospitals spanning all hospital tiers in Anhui Province, the resources associated with accurate coding followed a steep and consistent tier gradient. Certification, training, trainer capacity, coding software, coding dictionaries, and physician engagement all declined from Tier 3A to Tier 1 and unclassified institutions. Hospital-reported coding performance declined across the same gradient but less steeply. The most frequently reported errors-principal diagnosis selection, non-specific coding, and omission of comorbidities-can affect DRG assignment and severity weighting, although their payment consequences were not measured here.

A capacity gradient, not simply a quality problem

The distribution of capacity is more informative than a provincial average. Tier 2 hospitals accounted for 157 of the 299 institutions surveyed; fewer than half reported a nationally certified coder and three in five had no coding assistance software. The results therefore identify lower-tier groups in which capacity-building efforts could be prioritised, while not establishing that every deficit co-occurred in the same hospital.

This is consistent with what is known about the determinants of coding quality elsewhere. Santos and colleagues concluded that coding accuracy is primarily a function of organisational conditions — staffing, time, career structure and clinician collaboration — rather than of individual aptitude [9]; the international training survey found no common standard for coder qualification across health systems [10]; and the Chinese workload meta-analysis found volumes incompatible with careful coding [11]. Our contribution is to show that within a single health system operating under a single payment rule, these organisational conditions are not merely variable but systematically stratified.

We note one apparent discrepancy with the national picture. The pooled Chinese estimate of coder workload is approximately 11 300 records per coder per year, with about 7400 in secondary hospitals [11], whereas our Tier 3A median of 11 000 is comparable but our Tier 2 median of 4600 is considerably lower. This may reflect genuinely lower throughput in the secondary hospitals of this province, or it may reflect the different denominators used when part-time coders are counted. It should not be read as evidence that workload pressure is absent in lower-tier institutions, since a low absolute volume per coder can coexist with severe time pressure where coders hold multiple roles — a pattern consistent with the 27.7% of the workforce in our sample working part time.

Interpreting the reported-performance gradient

The contrast between a 3.4- to 10.8-fold spread in selected capacity indicators and a less than 1.8-fold spread in reported performance warrants cautious interpretation.

First, the performance outcome was self-assessed against local audit records, and audit procedures may differ between hospitals. Second, the ability to detect coding errors may itself depend on certified staff, software, and reference tools. Hospitals with less audit capacity may therefore identify fewer errors, potentially compressing the reported gradient. This mechanism is plausible but was not directly tested.

External evidence is consistent with caution about self-reported accuracy. In one well-resourced Chinese tertiary hospital, independent review of 323,320 records found coding accuracy increased from 78.9% to 87.1% after structured physician-coder collaboration [12]. In the present survey, 83.9% of Tier 3A hospitals reported accuracy of at least 95%. The studies used different definitions and methods and are not directly comparable, but the contrast illustrates why hospital-reported thresholds should not be treated as independently validated accuracy estimates.

Accordingly, the reported quality figures describe hospitals' own audit assessments rather than independently verified performance. The capacity gradient is more directly supported than any estimate of the absolute level of coding accuracy.

Implications for the equity of case-based payment

DRG payment applies one set of grouping rules, one weight schedule and one set of documentation requirements to every hospital. The capacity to satisfy those requirements, as shown here, is distributed very unevenly. The consequences of that mismatch are not symmetric.

Under-coding is one plausible consequence of limited coding capacity. An omitted comorbidity can prevent assignment to a severity-adjusted variant; a non-specific principal diagnosis may result in a residual group; and an unrecorded procedure can alter procedural grouping. These pathways could reduce reimbursement in some cases. However, this study did not compare clinically equivalent patients, independently audit coding errors, or observe payments, so it cannot estimate the direction or magnitude of any reimbursement effect.

If future record-level analyses confirm that lower coding capacity leads to systematic under-representation of case complexity, the resulting payment differences could reinforce existing resource inequalities. Longitudinal studies linking audited records, DRG reassignment, and reimbursement are needed to test this mechanism.

A different risk may arise at the well-resourced end of the gradient. Zhou and colleagues found that 12.9% of audited records in a Chinese DRG pilot were upcoded and that tertiary status was a risk factor [13]. Coding capability may support both completeness and strategic coding; its effect therefore depends on the audit and incentive regime. Capacity-building in lower-tier hospitals and effective audit across all tiers should be developed together.

Implications for health systems adopting case-mix payment

Case-mix payment reforms are typically evaluated on cost containment, length of stay and service utilisation. This study suggests that coding capacity deserves to be treated as a measured precondition of such reforms rather than as an assumed input, and that it is straightforward to measure. The indicators used here — the proportion of hospitals with a certified coder, with functional coding software, and with a maintained coding dictionary — are inexpensive to collect through routine administrative reporting, discriminate sharply between institutions, and are directly actionable.

Three operational implications merit evaluation. First, readiness indicators could be assessed by hospital tier as payment reforms are implemented; the dual-track DRG/DIP design already acknowledges variation in institutional capability [3]. Second, shared or provincial provision of coding software and maintained dictionaries may reduce fixed-cost barriers for small hospitals. Human-AI collaborative coding has reduced coding time and improved detection of missing codes in a pilot study, but wider

deployment requires validation across hospital tiers and depends on clinical documentation quality [14]. Third, expanding coder and trainer certification in lower-tier institutions is a plausible target, but its effect should be evaluated with independently audited coding outcomes.

Hospitals' own stated priorities align with this analysis: the two most frequently selected were expanding externally supplied training and physician education, both of which require action above the level of the individual institution.

Comparison with international evidence

The reported error profile was consistent with international literature in which principal diagnosis selection errors and non-specific coding recur as common problems [4, 5, 7]. Direct comparison of accuracy levels is not possible because international estimates generally derive from re-abstraction audits, whereas this survey used institutional self-assessment in ordinal bands. The contribution of this study is therefore the description of within-system differences in workforce and infrastructure across hospital tiers, rather than a comparison of coding accuracy between countries.

Limitations

This study has important limitations. First, coding quality was reported by hospitals in ordinal bands rather than measured through independent re-abstraction. Audit methods may have differed across institutions, and both performance and structural capacity were reported by institutional representatives. Absolute accuracy figures should therefore not be taken at face value, and reporting or social-desirability bias is possible.

Second, participation was required for secondary and higher-level hospitals but voluntary for primary and unclassified institutions. Lower-capacity facilities may therefore be under-represented. The survey was also organised through the provincial quality-control system, which may have influenced how respondents described institutional performance.

Third, the study is cross-sectional and cannot establish temporal or causal relationships. Fourth, it covers one province, and absolute capacity levels may differ in wealthier coastal or poorer western provinces. Fifth, hospitals of traditional Chinese medicine were excluded. Sixth, coder characteristics were aggregated to the institutional level, preventing analysis of individual coder attributes and outcomes. Seventh, certification and training were used as proxies for competence and coder knowledge was not tested directly. Finally, the study did not measure DRG reassignment or reimbursement, so the proposed payment pathways require direct evaluation.

Strengths

The study also has notable strengths. It is a large province-wide assessment spanning every hospital tier, and 299 of 311 returned questionnaires were retained after validation. Measuring workforce, technology, reference infrastructure, error profiles, and reported performance in one instrument allowed their tier

distributions to be examined together. The survey also provides a defined provincial baseline against which future capacity investments and independently audited outcomes can be evaluated.

Conclusions

Across surveyed hospitals in Anhui Province, the workforce and infrastructure associated with accurate clinical coding followed a marked hospital-tier gradient. Tier 1 and unclassified institutions were less likely to report certified coders, coding software rated as having good functionality, and regularly updated coding dictionaries. Hospital-reported coding performance showed a shallower gradient, but these self-assessments were not independently validated.

Uniform case-based payment rules are therefore being implemented across institutions with unequal coding capacity. The study did not measure reimbursement effects, but limited capacity may increase the risk that case complexity is incompletely represented. Coding readiness should be monitored alongside payment reform, and targeted training, shared technical infrastructure, and proportionate audit support should be evaluated using record-level accuracy and payment outcomes.

Abbreviations

CC/MCC
complications and comorbidities / major complications and comorbidities
DIP
diagnosis-intervention packet
DRG
diagnosis-related group
ICD-9-CM-3
International Classification of Diseases, Ninth Revision, Clinical Modification, Volume 3
ICD-10
International Classification of Diseases, Tenth Revision
IQR
interquartile range
STROBE
Strengthening the Reporting of Observational Studies in Epidemiology

Declarations

Ethics approval and consent to participate

This study analysed institution-level operational data collected through a provincial administrative quality-improvement survey. No identifiable patient or respondent data were collected. The need for ethics approval and informed consent was waived by the Ethics Committee of The First Affiliated Hospital of Wannan Medical University in accordance with Article 32 of the Measures for Ethical Review

of Life Sciences and Medical Research Involving Humans (Order No. 4 [2023] of the National Health Commission Science and Education Department), because the study used anonymised information data, did not harm human subjects, and did not involve sensitive personal information or commercial interests. All procedures were performed in accordance with relevant guidelines and regulations, including the Declaration of Helsinki.

Consent for publication

Not applicable.

Availability of data and materials

The aggregated datasets 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.

Funding

This study was supported by the Key Project of Philosophy and Social Science Research in Universities of Anhui Province (Grant No. 2023AH051725). The funding body contributed to study design, data analysis, and manuscript preparation but was not involved in data collection or interpretation of the findings.

Authors' contributions

LM, XJ, and DG conceived and designed the study. LM, XJ, DG, and YY coordinated data collection and performed data validation. LM, ZZ, and YW analysed the data and drafted the manuscript. LM and ZZ interpreted the results and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the Anhui Provincial Medical Record Quality Control Centre for organising the survey, the municipal quality control centres for coordinating data collection, and the medical record departments of all participating hospitals.

References

1. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res. 2005;40(5 Pt 2):1620–39.

2. Jetté N, Quan H, Hemmelgarn B, Drosler S, Maass C, Moskal L, et al. The development, evolution, and modifications of ICD-10: challenges to the international comparability of morbidity data. Med Care. 2010;48(12):1105–10.
3. He AJ. Scaling-up through piloting: dual-track provider payment reforms in China's health system. Health Policy Plan. 2023;38(2):218–27.
4. Campbell SE, Campbell MK, Grimshaw JM, Walker AE. A systematic review of discharge coding accuracy. J Public Health Med. 2001;23(3):205–11.
5. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, et al. Systematic review of discharge coding accuracy. J Public Health (Oxf). 2012;34(1):138–48.
6. Henderson T, Shepheard J, Sundararajan V. Quality of diagnosis and procedure coding in ICD-10 administrative data. Med Care. 2006;44(11):1011–9.
7. Mikkelsen G, Aasly J. Reliability of diagnoses coding with ICD-10. Int J Med Inf. 2007;76(11–12):781–8.
8. Quan H, Li B, Saunders LD, Parsons GA, Nilsson CI, Alibhai A, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health Serv Res. 2008;43(4):1424–41.
9. Santos S, Murphy G, Baxter K, Robinson KM. Organisational factors affecting the quality of hospital clinical coding. Health Inf Manag J. 2008;37(1):25–37.
10. Otero Varela L, Doktorchik C, Wiebe N, Southern DA, Knudsen S, Mathur P, et al. International Classification of Diseases clinical coding training: an international survey. Health Inf Manag J. 2024;53(2):68–75.
11. Liu Y, Wu C, Cao M, Lei C, Zhou Z, Ou W. Systematic review and meta-analysis of workload among medical records coders in China. BMC Med Inf Decis Mak. 2024;24:335.
12. Xu Y, Zhu H, Xu Z, Jin F, Chen J, Pan X, et al. Physician–clinical coder collaboration effectively improves coding accuracy: a single-centre prospective study in China. Health Inf Manag J. 2025;54(3):268–78.
13. Zhou W, Xu C, Zhang L, Fu H, Jian W. Behaviours and drivers of diagnosis-related group upcoding in China: a mixed-methods study. Soc Sci Med. 2025;366:117660.
14. Gao Y, Chen Y, Wang M, Wu J, Kim Y, Zhou K, et al. Optimising the paradigms of human AI collaborative clinical coding. NPJ Digit Med. 2024;7:368.
15. Wu H, Zhao L, Chen J, Fan W. Improvement in secondary diagnosis-related coding under DRG reform: a retrospective dual-center study in China. Br J Hosp Med. 2026;87(3):52079.
16. National Healthcare Security Administration. Press conference on safeguarding population health and supporting economic development. 2025. https://www.nhsa.gov.cn/art/2025/1/17/art_52_15492.html. Accessed 29 Sep 2026.

Figures

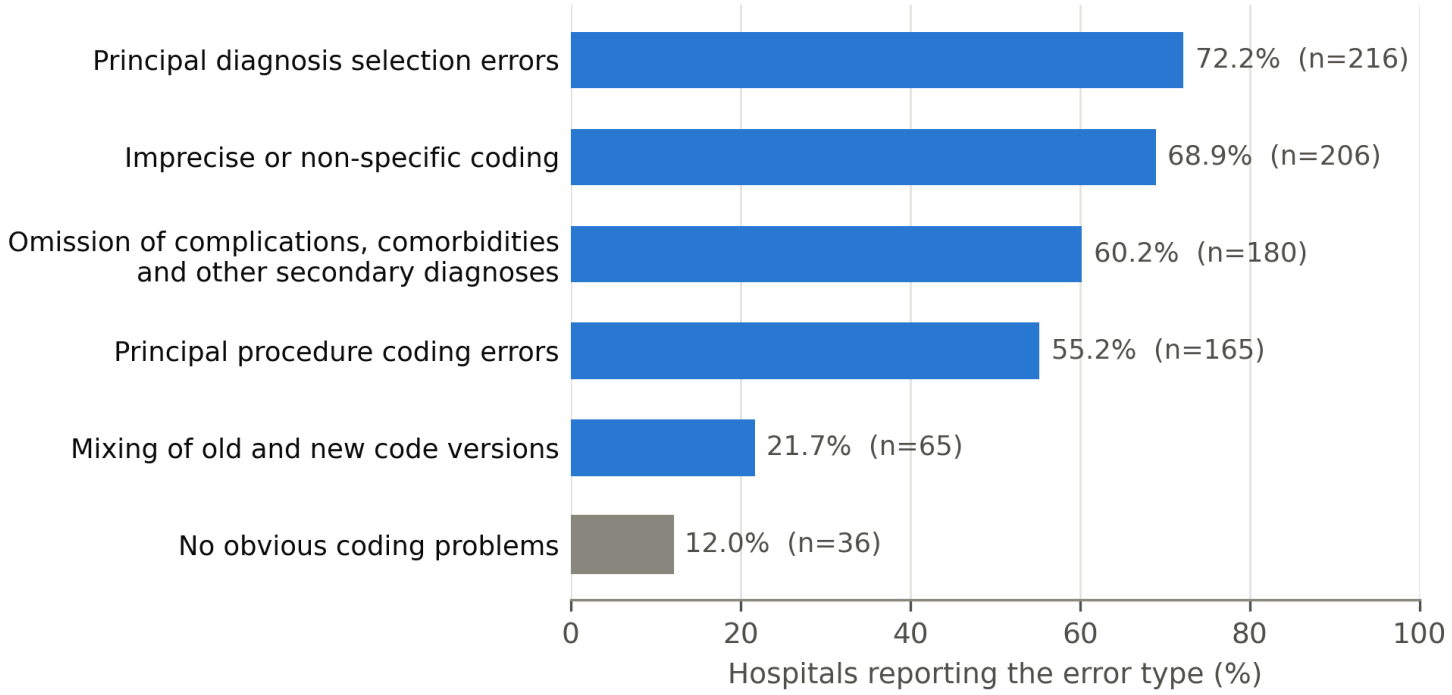


Figure 1

Coding error types reported by hospitals (N = 299). Respondents could select more than one option, so percentages do not sum to 100. Principal diagnosis selection errors, non-specific coding, and omission of complications, comorbidities, or other secondary diagnoses can affect DRG group assignment or severity weighting; payment effects were not measured in this survey.

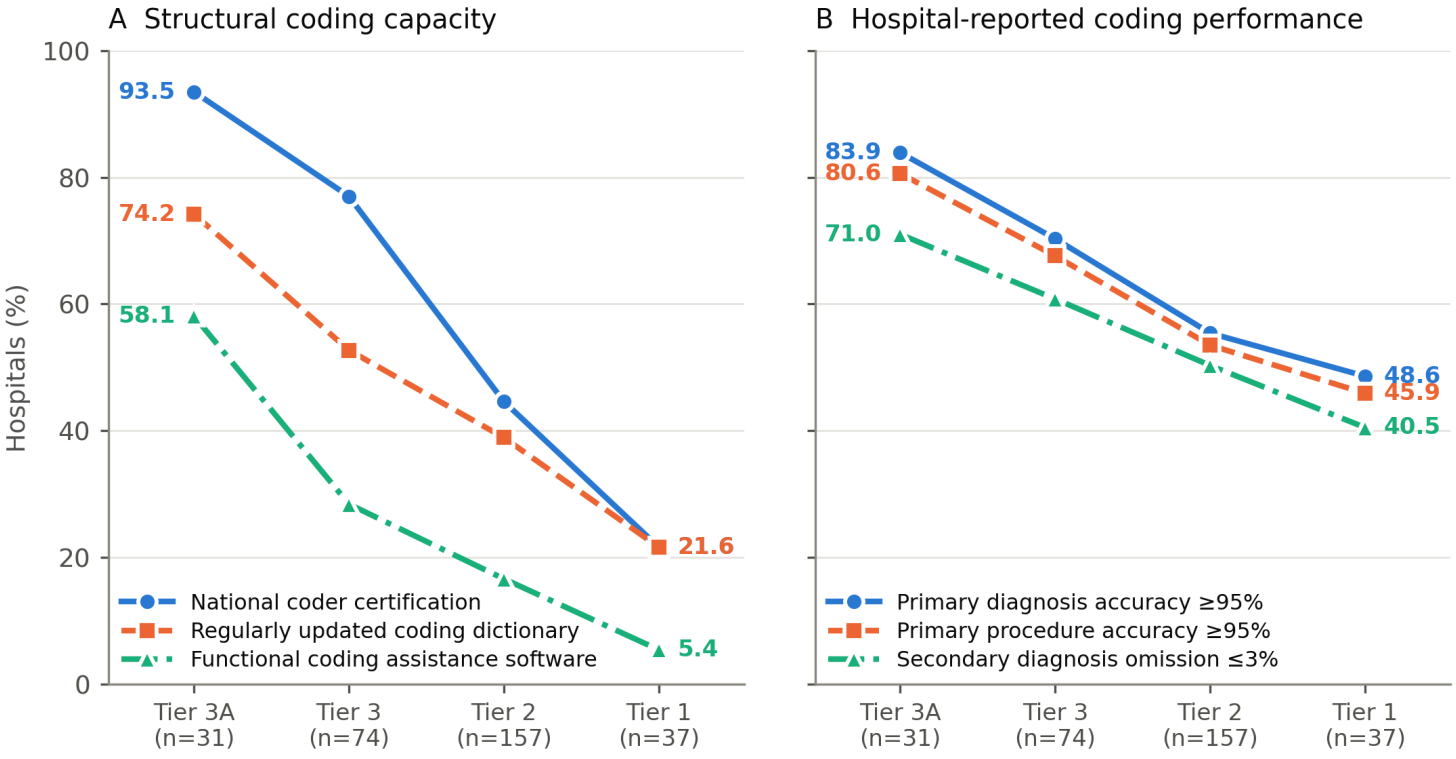


Figure 2

Selected structural coding capacity indicators (a) and hospital-reported coding performance indicators (b) across hospital tiers. Both panels use a common percentage scale. The selected capacity indicators vary 3.4- to 10.8-fold between Tier 3A and Tier 1 and unclassified institutions, whereas the selected reported-performance indicators vary less than 1.8-fold. Ratios are descriptive and sensitive to the lowest-tier value. End-point values are labelled. Tier 3 comprises Tier 3B and unclassified tertiary hospitals; Tier 2 comprises Tier 2A, Tier 2B, and unclassified secondary hospitals.

Supplementary Files

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

- Additionalfile.docx