Natural Resources for Human Health
eISSN: 2583-1194

Original Research
Received 03 Aug 2026
Revised 05 Sep 2026
Accepted 25 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (23s): 815-828

KEYWORDS:
Clinical documentation; Medical coding accuracy; Clinical documentation improvement; Data quality; Administrative efficiency; Healthcare outcomes

Clinical Documentation and Medical Coding Accuracy in Healthcare Organizations: A Systematic Review

Ruwida Khalid Qashqari[1*], Wed Abdullah Mahdaly[2], Hanouf Abdulmonem Alqurashi[3], Bshair Abdulaziz Bawhab[4], Abrar Ahmed Asiri[5], Mohammed Mwafaq Aloraini[6], Shumukh Hamed Alfakhri[7], Afnan Saleem Alfahmi[8], Sarah Sanhut T Alruqi[9], Hussam Zainaldeen Abdullah Siam[10]

[1] Health Services Administration Specialist, King Abdullah Medical City, Saudi Arabia
[2] Medical Transcription Technician, King Abdullah Medical City, Saudi Arabia
[3] Medical Transcription Technician, King Abdullah Medical City, Saudi Arabia
[4] Specialist, King Abdullah Medical City, Saudi Arabia
[5] Health Services Administration Specialist, King Abdullah Medical City, Saudi Arabia
[6] Health Administration Specialist, King Abdullah Medical City, Saudi Arabia
[7] Medical Transcription Technician, King Abdullah Medical City, Saudi Arabia
[8] Medical Transcription Technician, King Abdullah Medical City, Saudi Arabia
[9] Administrative Support Specialist, King Abdullah Medical City, Saudi Arabia
[10] Health Information Technician, King Abdullah Medical City, Makkah, Saudi Arabia
* Corresponding author: Ruwida Khalid Qashqari, King Abdullah Medical City, Saudi Arabia · Qashqari.r@kamc.med.sa

ABSTRACT: Clinical documentation is the source from which diagnoses and procedures are translated into codes that drive reimbursement, quality benchmarking, and research. Evidence linking documentation and coding accuracy to organizational performance has not been synthesized. To synthesize evidence on the relationship between clinical documentation and medical coding accuracy and administrative efficiency, data quality, and healthcare outcomes. This review followed PRISMA 2020. Studies published in English between January 2009 and 15 December 2025 were eligible if they were randomized trials, cohort, case-control, cross-sectional, or case series designs. Risk of bias was appraised with Joanna Briggs Institute tools, and certainty of evidence was graded using GRADE. Narrative synthesis was used because outcome heterogeneity precluded meta-analysis. Eighteen reports representing 17 studies from five countries met the criteria: 10 before–after evaluations and seven audits, coding-method comparisons, or surveys. Audits changed 13% to 17% of primary diagnoses and up to 55% of discharges when any code was considered. Large multispecialty audit data showed substantial coding inaccuracy, and targeted clinician and coder education improved initial coding accuracy. Documentation education, concurrent review, and structured templates improved reported severity or case-mix indices across studies measuring them. Financial effects varied. Length of stay was unchanged where measured, and effects on mortality remained uncertain. Documentation quality was associated with coding accuracy. Education and concurrent clinician–coder review were associated with improved data quality and more complete capture of reimbursable activity. Certainty was low to very low.

Natural Resources for Human Health
eISSN: 2583-1194

1. INTRODUCTION

The clinical record is the primary means by which clinicians communicate a patient's condition, the care delivered, and the reasoning behind it. Beyond its clinical function, the record is the raw material from which trained coders assign diagnosis and procedure codes under classification systems such as the International Classification of Diseases (Rangachari, 2007; Burns et al., 2012). Coded data therefore represent a translation of documentation, and the fidelity of that translation depends on what clinicians have written and how it is interpreted.

Coded data now serve purposes far beyond billing. They determine hospital funding under casemix and diagnosis-related group (DRG) payment models, support epidemiological surveillance and research, and feed benchmarking and public reporting of institutional performance (Cheng et al., 2009; Nouraei et al., 2015). As casemix systems expand into middle-income health systems, the dependence of financing and planning on coded data has become a structural feature of health service management rather than an administrative detail (Zafirah et al., 2018).

Coding accuracy is not determined by coders alone. It reflects a chain that begins with clinician documentation, passes through the transfer of information from clinicians to coders, and ends with the application of coding rules that do not allow coders to infer diagnoses that are not stated (Nouraei et al., 2016; Zafirah et al., 2018). Organizational conditions, including staffing, training, information technology, and internal audit arrangements, shape each step of that chain and vary substantially between hospitals and countries (Pongpirul et al., 2011).

Errors in this chain carry consequences in several directions. Under-documentation and under-coding can understate patient complexity and reduce the funding that follows the patient, whereas over-coding exposes organizations to audit and recovery of payments. Misclassified severity also distorts risk-adjusted indicators used to compare hospitals and surgeons, because expected outcomes are calculated from the diagnoses that were documented and coded (Rangachari, 2007; DeCicco et al., 2022).

In response, many organizations have introduced clinical documentation improvement (CDI) programs. These include clinician education on classification and payment rules, problem-based or structured note templates, concurrent review by clinical documentation specialists, and redesign of electronic documentation platforms (Spurgeon et al., 2011; Reyes et al., 2017; Aiello et al., 2018; Elkbuli et al., 2018). Evaluations are typically conducted in single departments or hospitals and report different metrics, which makes it difficult for managers and policy makers to judge what such programs achieve.

Existing syntheses have concentrated on the accuracy of discharge coding in routinely collected hospital data, principally in Great Britain, and on primary diagnoses and main procedures (Campbell et al., 2001; Burns et al., 2012). To our knowledge, no review has examined documentation and coding accuracy alongside administrative efficiency, data quality, and healthcare outcomes, nor compared evidence from audit studies with evidence from improvement interventions. Such a synthesis is needed to inform investment in coder training, documentation programs, and audit infrastructure, particularly in settings where coding capacity is still developing.

This systematic review therefore aimed to synthesize the evidence from 2009 to 2025 on clinical documentation and medical coding accuracy in healthcare organizations and its relationship with administrative efficiency, data quality, and healthcare outcomes. The review question was: in healthcare organizations, how do the accuracy of clinical documentation and medical coding, and interventions to improve them, affect administrative efficiency, data quality, and healthcare outcomes compared with baseline or reference-standard coding?

2. MATERIALS AND METHODS

2.1. Protocol and reporting

This review was designed and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). The completed flow diagram is shown in **Figure 1**. The review was not prospectively registered. The a priori review protocol is available from the corresponding author.

2.2. Eligibility criteria

The review question was structured using the PICO framework:

Population: healthcare organizations (hospitals, departments, and services), their clinicians, clinical coders, and the patient records they produce.

Intervention/exposure: clinical documentation practices, medical coding practices, or interventions to improve either (clinician education, structured templates, concurrent documentation review, electronic health record redesign, coding audit, or alternative coding methods).

Comparator: pre-intervention baseline, independent reference-standard re-coding or audit, or comparison between coding methods or settings.

Outcomes: coding accuracy or error; administrative efficiency (reimbursement, revenue, tariff, documentation time, productivity); data quality (severity of illness, risk of mortality, case mix index, completeness); and healthcare outcomes (mortality, length of stay).

Eligible designs were randomized controlled trials, cohort studies, case-control studies, cross-sectional studies, and case series, published in English from 1 January 2009 through 15 December 2025, inclusive. Systematic and narrative reviews, qualitative-only studies, editorials, commentaries, conference abstracts, and studies without original quantitative data were excluded.

2.3. Information sources and search strategy

PubMed/MEDLINE, Scopus, Web of Science, and CINAHL were searched for eligible publications from January 2009 through 15 December 2025, supplemented by backward and forward citation searching of included studies. The core PubMed strategy combined three concept blocks: ("clinical documentation" OR "clinical documentation improvement" OR "medical record documentation" OR "physician documentation") AND ("clinical coding" OR "medical coding" OR "diagnosis coding" OR "ICD-10" OR "diagnosis-related group*") AND (accuracy OR error* OR audit OR "data quality" OR reimbursement OR "case mix" OR "severity of illness" OR mortality). The strategy was adapted to the syntax of each database. The final search was executed on 15 December 2025. **Figure 1** gives the retrieval counts by database.

2.4. Study selection and data extraction

Records were deduplicated and screened by title and abstract, and full texts of potentially eligible reports were assessed against the criteria by two reviewers independently, with disagreements resolved by discussion. Data were extracted into a standardized form: country and setting, design, sample size and period, documentation or coding focus, outcome domain, quantitative results, and authors' conclusions.

2.5. Risk of bias and certainty of evidence

Methodological quality was appraised independently by two reviewers using the Joanna Briggs Institute (JBI) checklists matched to design: analytical cross-sectional, cohort, and quasi-experimental (non-randomized) tools. Disagreements were resolved by consensus. Overall judgments reflected concern on a continuum from low to high. Certainty of evidence for each outcome domain was rated independently by two reviewers using GRADE. Final ratings and the principal reasons for downgrading are presented in Table 3.

2.6. Synthesis

A meta-analysis was planned if data allowed. Pooling was not appropriate because studies differed in design, setting, and outcome metrics (error proportions, index scores, and currency-denominated revenue), and several reported no measure of variance. Results were therefore synthesized narratively, structured by outcome domain and by study design (audit versus intervention), following the Synthesis Without Meta-analysis (SWiM) guidance.

3. RESULTS AND DISCUSSION

3.1. Study selection

The database searches identified 2,846 records. After removal of 934 duplicates, 1,912 records were screened by title and abstract. Seventy-eight full-text reports were sought and assessed, and 60 were excluded (ineligible design, n = 21; outside the 2009–2025 window, n = 12; irrelevant outcomes, n = 19; non-English, n = 8). Eighteen reports representing 17 studies were included in the synthesis (**Figure 1**).

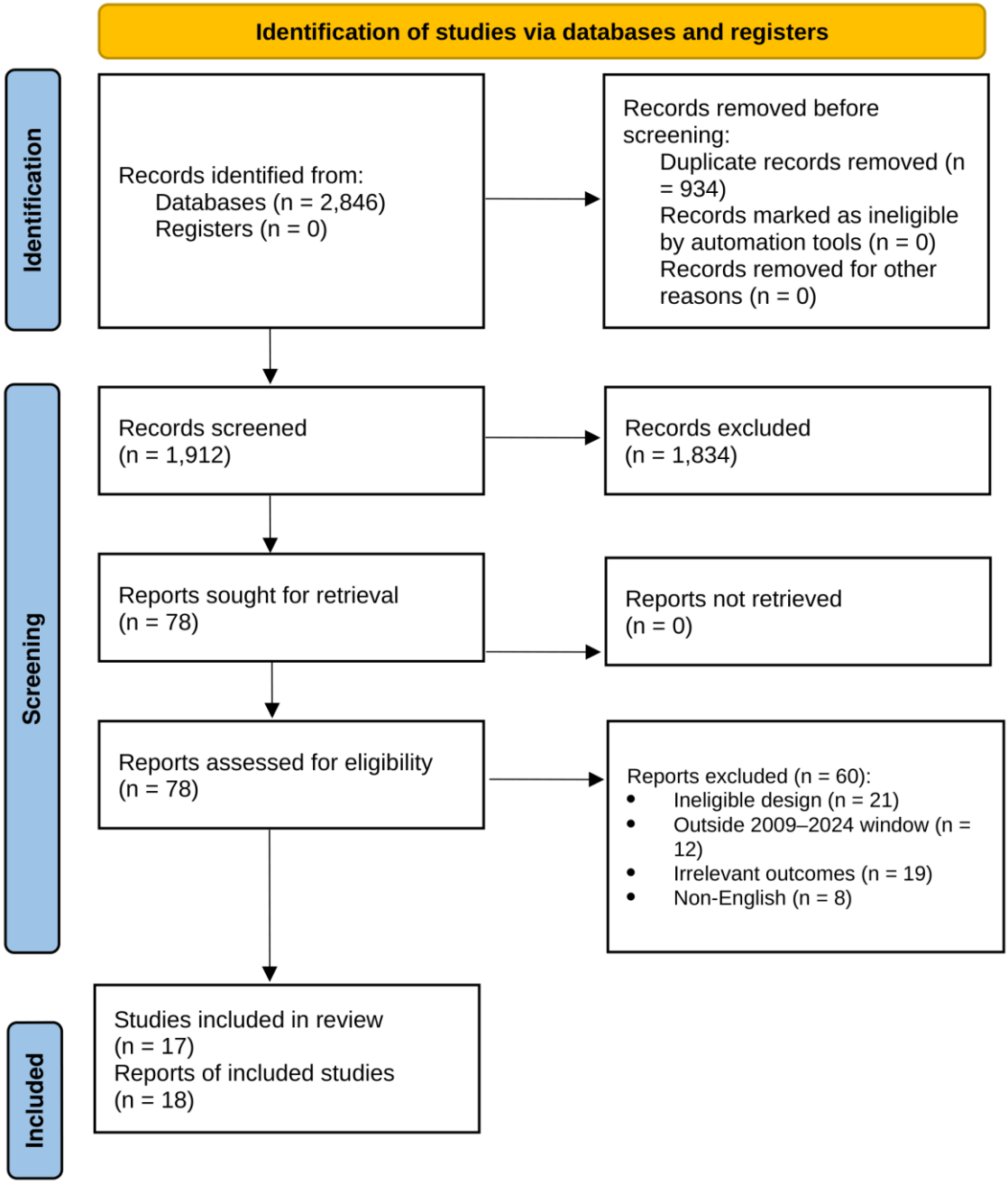


Figure 1. PRISMA 2020 flow diagram of study identification, screening, and inclusion. Database records comprised PubMed/MEDLINE (n = 1,020), Scopus (n = 742), Web of Science (n = 596), and CINAHL (n = 488). Adapted from Page et al. (2021), CC BY 4.0.

3.2. Characteristics of included studies

The 18 included reports were published between 2009 and 2025 and represented 17 studies in five countries: the United States (n = 9), the United Kingdom (n = 5), and Australia, Malaysia, and Thailand (n = 1 each). No randomized trials, case-control studies, or case series were identified. Ten studies used before–after designs and seven were coding audits, method comparisons, or surveys. The two reports by Tsopra et al. (2018, 2019) analyzed the same 107 discharges and were counted once in study totals. The Thai nationwide survey provided comparative evidence on hospital coding practices by audit status and was used to

contextualize the direct assessments of coding accuracy. Settings ranged from single departments to whole-hospital programs and a national hospital survey (Table 1).

Table 1. Characteristics of the included studies

Study	Country / setting	Design	Sample / period	Documentation or coding focus
Cheng et al. (2009)	Australia; Victorian casemix-funded hospital	Retrospective coding audit (cross-sectional case study)	752 audited cases	Miscoding due to inadequate medical documentation and its effect on funding
Pongpirul et al. (2011)	Thailand; hospitals under the Universal Coverage scheme	Cross-sectional nationwide questionnaire survey	465 responding hospitals (370 audited; 95 non-audited); 920 audited and 390 non-audited hospitals approached	Hospital coding practice and DRG coding quality
Nouraei et al. (2013)	United Kingdom; NHS otolaryngology services	Prospective multidisciplinary coding audit (cross-sectional)	3,131 randomly selected patients, 2010–2012	Subjectivity, variability, and error in coding; education plus audit
Rosenbaum et al. (2014)	United States; Cleveland Clinic neurosurgery	Retrospective before–after cohort	3,323 discharges before vs. 3,334 after (1 year each)	Clinician education on MS-DRG with CDI and coding collaboration
Frazee et al. (2015)	United States; trauma/acute care surgery	Retrospective before–after cohort	584 trauma admissions in the 2009 baseline year; comparison with annual cohorts in 2010–2012	Concurrent physician-panel documentation review
Fox et al. (2016)	United States; Level 1 trauma center	Retrospective before–after cohort	9 surgeons; 776 Medicare patients, 2014 (411 charts reviewed)	Focused documentation improvement: 3-h training plus specialist concurrent review
Nouraei et al. (2016)	United Kingdom; teaching and district hospitals	Clinician–auditor coding audit (cross-sectional)	8,888 emergency medical discharges over 3 years	Clinician–coder information handover and coding accuracy
Tsopra et al. (2018)	United Kingdom; NHS hospital	Comparative coding-method study (cross-sectional)	107 consecutive discharges from three respiratory wards, Leeds, March 2015 (same cohort as ref. 12)	Coding from discharge summary alone, with case notes, or with medical input

Study	Country / setting	Design	Sample / period	Documentation or coding focus
Liu et al. (2018)	United States; neonatal intensive care unit	Before–after quality improvement cohort	Oct 2015–Oct 2016 vs. Nov 2016–May 2017	Problem-based neonatal EHR documentation redesign and provider education
Elsamadicy et al. (2018)	United States; academic neurosurgery service	Cohort (prospectively collected data)	5,434 patients: 2,793 pre vs. 2,641 post; Jul 2014–Jun 2016	Multifactorial quality initiative with documentation education
Zafirah et al. (2018)	Malaysia; teaching hospital (MY-DRG casemix)	Cross-sectional descriptive audit	464 coded records, 2013	Coding errors versus an independent senior coder and tariff impact
Tsopra et al. (2019)	United Kingdom; three respiratory wards	Prospective cohort	107 of 110 consecutive discharges, Leeds, March 2015 (same cohort as ref. 8)	Accuracy of diagnoses recorded in discharge summaries
DeCicco et al. (2022)	United States; academic referral center	Retrospective cohort (PDSA quality improvement)	26 hospitalists; 7,987 admission and 51,583 subsequent encounters over 2.5 years	Hospital-wide multidisciplinary billing and coding documentation intervention
Nouraei et al. (2015)	United Kingdom; two hospital centres across surgical specialties	Cross-sectional clinician–coder audit	30,127 patients; independent re-audit of 400 randomly selected cases	Agreement between original and audited diagnoses, procedures, HRGs, morbidity and complication coding
Reyes et al. (2017)	United States; University of New Mexico academic surgery department	Before–after evaluation	Department-wide monthly metrics before and after September 2014 rollout	CDI and ICD-10 education with provider training and documentation queries
Aiello et al. (2018)	United States; University of Massachusetts vascular and endovascular surgery service	Retrospective before–after evaluation	January 2013–July 2016; 458 carotid endarterectomies (205 pre, 253 post) and 504 open infrainguinal procedures (277 pre, 227 post)	Physician-led education, specialty documentation guide and concurrent review with coding specialists
Elkbuli et al. (2018)	United States; Kendall Regional Medical Center trauma service, Miami	Before–after comparison of prospectively collected data	Two-year period; 1,419 trauma patients pre-CDI and 1,454 post-CDI	Effects of a CDI program on coded severity and risk-adjusted trauma outcomes

Study	Country / setting	Design	Sample / period	Documentation or coding focus
Radhakrishnan et al. (2024)	United Kingdom; single inpatient ward, Imperial College Healthcare NHS Trust	Before–after quality improvement evaluation	94 admissions before and 73 after education; one month per phase	Targeted clinician and coder education followed by review of HRGs and individual codes

CDI, clinical documentation improvement; DRG, diagnosis-related group; EHR, electronic health record; MS-DRG, Medicare Severity DRG; PDSA, plan–do–study–act.

3.3. Key findings

Findings are grouped by outcome domain (Table 2 and Table 3). Audit studies consistently showed substantial discordance between original and audited codes. Interventions that educated clinicians and added concurrent review improved severity indices, and financial effects were reported in both directions.

Nouraei et al. (2015) documented substantial coding inaccuracy in a large surgical audit. Reyes et al. (2017), Aiello et al. (2018), and Elkbuli et al. (2018) reported improvements in documentation completeness, coded case mix, estimated charges or contribution margin, or risk-adjusted reporting after surgical CDI. Radhakrishnan et al. (2024) reported improved initial coding accuracy after targeted education.

Table 2. Key findings and authors' conclusions of the included studies

Study	Key quantitative findings	Domain	Authors' conclusions
Cheng et al. (2009)	About 16% of 752 audited cases changed DRG; net revenue effect of nearly AU$575,300; 56% of DRG changes were due to documentation issues	ACC, EFF	Inadequate documentation creates a risk of miscoding with measurable funding consequences
Pongpirul et al. (2011)	Audited and non-audited hospitals differed in numbers of medical statisticians, experience, and certified coders; factor analysis produced a 3-factor, 20-item coding-practice scale	Context	High-quality DRG coding should not be assumed in resource-limited settings
Nouraei et al. (2013)	Primary diagnosis changed in 13% (420) and primary procedure in 13% (417) of 3,131 patients; surgeon "code in theatre" without quality assurance gave an HRG error rate of 45%	ACC	Coding is complex and subject to variability and error; education plus audit improves but does not eliminate it
Rosenbaum et al. (2014)	Normalized CMI +28% ($p < 0.0001$); CMI +5.3% ($p = 0.0074$); MCC capture +39%; ROM 3–4 +24%; SOI 3–4 +11%; ICD-9 codes per stay +12%; actual LOS +1.2% (NS)	DQ, OUT	Education improved documentation; no adverse effect on quality metrics was observed
Frazee et al. (2015)	SOI, ROM, and CMI higher than 2009 baseline in each subsequent year ($p < 0.05$); LOS unchanged; Injury Severity Score lower in 2011 and 2012	DQ, OUT	Concurrent review increased scores through more accurate keyword documentation, not higher acuity

Study	Key quantitative findings	Domain	Authors' conclusions
Fox et al. (2016)	Opportunities for improvement in 43% (177/411) of reviewed charts; 100% physician query response; CMI 1.80 (SD 0.15) to 2.11 (SD 0.19), $p < 0.001$	DQ, EFF	Focused documentation improvement engages physicians and yields a significant return on investment
Nouraei et al. (2016)	At least one code changed in 55% (4,889) of 8,888 discharges; primary diagnosis changed in 16.8% (1,496 spells); Charlson score rose in 8.2% and fell in 2.3%; income variance £816,977 (+5.0%), £91.92 per patient	ACC, DQ, EFF	Accuracy depends on clinician–coder information transfer; greater clinician involvement recommended
Tsopra et al. (2018)	Incorrect HRGs: 42% (summary alone), 31% (with case notes), 35% (with medical input); estimated annual remuneration loss £1.8 M–£16.5 M across methods	ACC, EFF	Discharge summary quality needs to be improved
Liu et al. (2018)	Monthly SOI +11.1% ($p = 0.008$); ROM +13.5% ($p = 0.007$); CMI improved; expected payment +US$14,020 per patient per month ($p < 0.001$); H&P time +47%, progress note time +91%, discharge summary time −41%	DQ, EFF	Problem-based EHR documentation improved documentation without provider dissatisfaction
Elsamadicy et al. (2018)	Monthly mortality 3.08 vs. 4.17 (trend); mortality index 0.73 vs. 0.98; CMI 3.14 vs. 2.96 ($p = 0.02$); discharge SOI 2.10 vs. 1.95 ($p < 0.0001$); discharge ROM 1.69 vs. 1.58 ($p = 0.003$)	DQ, OUT	Initiative improved documentation and was associated with lower observed mortality
Zafirah et al. (2018)	Errors in 89.4% (415/464) of records; secondary diagnoses 81.3%, secondary procedures 58.2%, principal procedures 50.9%, primary diagnoses 49.8%; MY-DRG changed in 74.0% (307/415), of which 52.1% (160/307) received a lower tariff; potential loss RM654,303.91	ACC, EFF	Coding quality is crucial for casemix; retraining and monitoring of coders recommended
Tsopra et al. (2019)	Mean inaccuracy 55% per discharge summary (95% CI 52–58%); primary diagnosis wrong, inaccurate, missing, or mis-recorded in about half; inaccuracy correlated with disease type but not with number of diagnoses or LOS	ACC	Diagnoses were not accurately recorded; discharge summary quality should be measured and improved
DeCicco et al. (2022)	Level 3 admission visits +76% and subsequent visits +112% (both $p < 0.001$); wRVUs +13.6% and +14.3%; revenue +US$233,988.79 per 10,000 encounters; 58% of level 2 notes met level 3 criteria	ACC, EFF	Multidisciplinary intervention aligned coding with true complexity of care and raised revenue
Nouraei et al. (2015)	Any code changed in 15,402/30,127 patients (51%); primary diagnosis changed in 3,911 (13%) and HRG in 5,183 (17%); audited income variance +£3,974,544 (+6.2%)	ACC, EFF	Coding variability can affect financial estimates and comparisons of surgical performance

Study	Key quantitative findings	Domain	Authors' conclusions
Reyes et al. (2017)	Delinquent documentation fell 85%; surgical process measure compliance rose from 85% to 97%; SOI, ROM, CMI and APR-DRG metrics increased (all $p < 0.01$); estimated charges increased by US$4,672,786	DQ, EFF	A department-wide curriculum improved documentation and reporting metrics
Aiello et al. (2018)	CMI rose from 1.25 to 1.36 after carotid endarterectomy ($p = 0.03$) and from 2.05 to 2.23 after open infrainguinal procedures ($p < 0.01$). Contribution margin rose from US$6,650 to US$7,859 and from US$8,840 to US$12,165, respectively. LOS was unchanged.	DQ, EFF, OUT	Focused documentation review improved coded case mix and contribution margin
Elkbuli et al. (2018)	CMI rose from 2.48 to 2.87 ($p < 0.001$); SOI rose from 2.12 to 2.91 ($p < 0.018$); observed-to-expected mortality ratio fell from 1.36 to 0.70 ($p < 0.001$). Crude mortality fell from 3.45% to 2.61% (non-significant).	DQ, OUT	Documentation changed reported severity and risk-adjusted mortality
Radhakrishnan et al. (2024)	HRG reassignment fell from 55/94 (58.5%) to 15/73 (20.5%; $p < 0.0001$). Primary diagnosis changes fell from 17.3% to 5.2%; secondary diagnosis changes fell from 37.6% to 27.8%.	ACC	Targeted education was associated with improved initial inpatient coding accuracy

ACC, coding accuracy; DQ, data quality; EFF, administrative efficiency (financial or time); OUT, healthcare outcome; CMI, case mix index; HRG, healthcare resource group; LOS, length of stay; MCC, major complication or comorbidity; NS, not significant; ROM, risk of mortality; SOI, severity of illness; wRVU, work relative value unit.

Table 3. Synthesis of findings by outcome domain

Outcome domain	Studies (n)	Summary of quantified findings	GRADE certainty
Coding accuracy	7 studies, 8 reports (refs 1, 3, 7, 8, 11, 12, 14, 18)	Primary diagnosis changed on audit in 13% to 16.8% of episodes. At least one code changed in 51% of surgical cases (ref. 14) and 55% of acute medical discharges (ref. 7). In the education evaluation, HRG reassignment fell from 58.5% to 20.5% (ref. 18).	Low (observational designs and heterogeneous reference standards)
Administrative efficiency	10 (refs 1, 6, 7, 8, 9, 11, 13, 14, 15, 16)	Audit-related income variance included +£3.97 million (+6.2%) in the surgical audit (ref. 14). Post-CDI estimates showed higher charges (ref. 15) and contribution margin (ref. 16). Note-writing time increased while discharge summary time fell in the EHR redesign (ref. 9).	Low (uncontrolled comparisons and heterogeneous financial and time metrics)
Data quality (SOI, ROM, CMI)	8 (refs 4, 5, 6, 9, 10, 15, 16, 17)	Coded severity or case mix improved after intervention. Examples include CMI increases from 1.25 to 1.36 and 2.05 to 2.23 across two vascular procedures (ref. 16), and from 2.48 to 2.87 in trauma (ref. 17).	Low (uncontrolled before–after designs and heterogeneous documentation measures)

Outcome domain	Studies (n)	Summary of quantified findings	GRADE certainty
Healthcare outcomes	5 (refs 4, 5, 10, 16, 17)	Length of stay was unchanged in three studies (refs 4, 5, 16). Observed mortality trended lower after a multifactorial initiative (ref. 10). In the trauma CDI study, crude mortality fell from 3.45% to 2.61% without statistical significance (ref. 17).	Very low (non-randomized comparisons, imprecision, and confounding by co-interventions)

GRADE, Grading of Recommendations Assessment, Development and Evaluation. Studies may contribute to more than one domain.

3.4. Risk of bias

Risk of bias was moderate or higher in most studies (Table 4). Audit studies generally used independent re-coding but came from single institutions, whereas before–after studies lacked concurrent control groups and often introduced several changes simultaneously. No study was judged to have a low overall concern in every domain. Assessments were made independently by two reviewers, with disagreements resolved by consensus.

Table 4. Risk-of-bias appraisal (JBI critical appraisal tools)

Study	JBI tool	Main methodological concerns	Overall concern
Cheng et al. (2009)	Analytical cross-sectional	Single jurisdiction; audit sampling frame and auditor blinding not clearly described	Moderate
Pongpirul et al. (2011)	Analytical cross-sectional	Self-reported practice data; possible response and social-desirability bias; no accuracy measure	Moderate
Nouraei et al. (2013)	Analytical cross-sectional	Random sample and double-reading with arbitration; single service and trust	Low–moderate
Rosenbaum et al. (2014)	Quasi-experimental	No control group; retrospective; large samples and multiple-comparison adjustment reduce but do not remove confounding	Moderate–high
Frazee et al. (2015)	Quasi-experimental	No control group; single baseline year; secular changes in coding practice possible	Moderate–high
Fox et al. (2016)	Quasi-experimental	Few surgeons; only 57% of charts reviewed; no concurrent control	High
Nouraei et al. (2016)	Analytical cross-sectional	Large audit across two hospitals; audit performed by one team; limited to acute medicine	Low–moderate
Tsopra et al. (2018)	Analytical cross-sectional	Single hospital and three respiratory wards; reference standard based on clinician–coder judgment; different coders used across methods	Moderate
Liu et al. (2018)	Quasi-experimental	Short post-intervention period; single unit; no control; multiple concurrent changes	High

Study	JBI tool	Main methodological concerns	Overall concern
Elsamadicy et al. (2018)	Cohort	Multifactorial intervention prevents attribution; 12-month windows; prospective data collection	Moderate–high
Zafirah et al. (2018)	Analytical cross-sectional	Single hospital; independent senior coder as reference but also the trainer of the in-house coders	Moderate
Tsopra et al. (2019)	Cohort	Small sample from three wards; reference list based on consultant judgment	Moderate
DeCicco et al. (2022)	Cohort / quasi-experimental	No standard control group; pandemic affected volumes; paired comparison of same hospitalists	Moderate–high
Nouraei et al. (2015)	Analytical cross-sectional	Audit reference depended on clinician–coder judgment. Independent re-audit covered a subset and found residual disagreement.	Moderate
Reyes et al. (2017)	Quasi-experimental	Single department; no concurrent control; bundled intervention; charge estimates based on query responses.	Moderate–high
Aiello et al. (2018)	Quasi-experimental	Single specialty service; no concurrent control; retrospective comparison and possible secular changes.	Moderate–high
Elkbuli et al. (2018)	Quasi-experimental	Single trauma centre; no concurrent control; changes in case mix or care could affect comparisons of mortality.	Moderate–high
Radhakrishnan et al. (2024)	Quasi-experimental	One ward and one month per phase; no concurrent control; reviewer judgment informed the audit reference.	Moderate

JBI, Joanna Briggs Institute. Overall concern: low to high, based on the domains of each checklist.

3.5. DISCUSSION

Clinical documentation and medical coding sit at the junction of clinical care, financing, and information governance. Accurate documentation allows coders to translate care into a classification that determines payment, populates quality indicators, and supports research. This review shows that accuracy at that junction is far from uniform and that it can be improved by deliberate organizational effort.

The audit studies produced a consistent picture across health systems with different payment models. Substantial proportions of episodes were re-coded when independent auditors reviewed the record, and the discordance was greatest for secondary diagnoses and procedures, which are the elements that determine severity and complexity. Studies from Australia, the United Kingdom, and Malaysia reached the same conclusion, indicating that the problem is not confined to one classification or payment system. These findings partly diverge from the systematic review by Burns et al. (2012), which concluded that routinely collected data had become sufficiently robust for research and managerial use. The difference is probably explained by scope: that review examined accuracy for the primary diagnosis and main procedure, where agreement is highest, whereas the audits included here also examined secondary codes. The earlier review by Campbell et al. (2001) likewise documented wide variation in accuracy between studies, supporting the heterogeneity observed in the present synthesis. Nouraei et al. (2015), included in this review, reported a similar pattern in a very large multispecialty surgical audit, which corroborates the audit evidence synthesized here.

The mechanism linking documentation to coding accuracy is largely one of information transfer. Coders are bound by rules that prevent them from inferring clinical meaning, so a diagnosis that is clinically obvious but not explicitly documented in the required terms will not be captured. Spurgeon et al. (2011) tested this mechanism in neurosurgery with an in-service, a progress-note template, and concurrent coder feedback, and the later studies by Rosenbaum et al. and Frazee et al. followed the same logic. Education and concurrent review increased the capture of complications, comorbidities, and severity markers, while patient age, length of stay, and case type did not change, which suggests that the gains reflected better documentation rather than sicker patients. Fox et al. showed that a clinical documentation specialist embedded in the service achieved comparable effects, and the included studies by Reyes et al. (2017), Aiello et al. (2018), and Elkbuli et al. (2018) reported improvements in quality metrics, estimated charges, case mix, or risk-adjusted reporting in surgical departments. The study by Liu et al. extends the mechanism to platform design: a problem-based electronic note prompted clinicians to record diagnoses in a structured way, and it redistributed time between documents rather than simply adding workload.

Comparison between groups showed that the source of information matters as much as the intervention. The comparison of coding methods by Tsopra et al. indicated that coding from discharge summaries alone was the least accurate approach, and that access to case notes improved accuracy more than medical input did, which implies that the underlying record, and not only the coding process, is the limiting factor. The qualitative work of Pongpirul et al. (2011) in Thailand explains why: coding practice involves many professional groups and steps, and it is influenced by internal dynamics, resources, and financial dependency. Resource-limited settings therefore face a double burden of fewer certified coders and less standardized documentation, a pattern echoed by the Thai survey and the Malaysian audit included here.

The financial consequences were bidirectional, but under-capture predominated. Correction of codes raised income in the large British audit and in the Australian audit, most re-grouped cases in the Malaysian audit moved to a lower tariff because secondary diagnoses had been omitted, and the hospital-wide intervention by DeCicco et al. found that downcoding was far more prevalent than upcoding. Revenue gains after improvement should nevertheless be interpreted cautiously. Nouraei et al. (2013) found that education improved coding variability only when supported by an audit programme, DeCicco et al. committed to continued internal audit so that improvement would not drift into over-coding, and Rosenbaum et al. warned against a singular focus on revenue. Efficiency also has a time dimension that financial metrics overlook, and the redesign described by Liu et al. suggests that documentation burden can be rebalanced rather than increased.

The relationship with healthcare outcomes is the least secure. Improved severity documentation raises expected mortality and expected length of stay, which makes observed outcomes look better against risk-adjusted benchmarks without any change in care, an effect that Elsamadicy et al. recognized when interpreting their mortality findings. The observed decline in mortality in that cohort coincided with standardized care protocols, so documentation cannot be credited with a clinical effect. Rosenbaum et al. and Frazee et al. at least showed that actual length of stay was unchanged, which supports the view that documentation programs mainly improve the validity of risk adjustment rather than clinical results. The included study by Radhakrishnan et al. (2024) showed that targeted education of clinicians and coders improved initial coding accuracy, which reinforces the value of upstream investment.

Several limitations qualify these conclusions. All included studies were observational, most were single-center, and none used randomization or concurrent controls, so confounding by secular change and co-interventions cannot be excluded. Outcome metrics differed widely and could not be pooled, and publication bias favoring successful programs is likely. Evidence came mainly from the United States and the United Kingdom, which limits generalizability to systems with different classification, payment, and workforce arrangements. Future research should use cluster-randomized or stepped-wedge designs, adopt a core set of outcome measures that includes documentation time and patient outcomes, and evaluate computer-assisted coding and natural-language tools against the same standards.

4. CONCLUSIONS

Evidence from 17 observational studies indicates that incomplete or nonspecific clinical documentation is associated with coding errors, particularly in the capture of secondary diagnoses and patient complexity. Clinician and coder education, concurrent review, and structured documentation were associated with more complete coding and improved measures of data quality, while financial effects varied across settings. The available evidence does not establish that these improvements reduce mortality, patient harm, or length of stay. Low to very low certainty, heterogeneous outcome measures, and potential confounding limit causal interpretation. Healthcare organizations should evaluate documentation improvement using both

coding accuracy and independently measured clinical outcomes. Controlled multicenter studies with standardized measures of accuracy, workload, financial effects, and patient safety are needed.

FUNDING

No external funding was received for this work.

ACKNOWLEDGEMENTS

None.

AUTHOR CONTRIBUTIONS

All authors equally contributed to this work (Conceptualization, Methodology, Data Curation, Formal Analysis, Writing – Original Draft, Writing – Review & Editing, and Supervision).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Ethics Statement

This study is a systematic review of previously published, de-identified data and did not involve new human or animal subjects; formal ethics approval was therefore not required.

Data Availability Statement

The data extraction tables supporting this review are available from the corresponding author upon reasonable request. All included studies are publicly available in the cited peer-reviewed sources.

REFERENCES

[1] Cheng, P., Gilchrist, A., Robinson, K.M., Paul, L., 2009. The risk and consequences of clinical miscoding due to inadequate medical documentation: a case study of the impact on health services funding. Health Information Management Journal 38(1), 35-46. https://doi.org/10.1177/183335830903800105

[2] Pongpirul, K., Walker, D.G., Rahman, H., Robinson, C., 2011. DRG coding practice: a nationwide hospital survey in Thailand. BMC Health Services Research 11, 290. https://doi.org/10.1186/1472-6963-11-290

[3] Nouraei, S.A.R., Hudovsky, A., Virk, J.S., Chatrath, P., Sandhu, G.S., 2013. An audit of the nature and impact of clinical coding subjectivity variability and error in otolaryngology. Clinical Otolaryngology 38(6), 512-524. https://doi.org/10.1111/coa.12153

[4] Rosenbaum, B.P., Lorenz, R.R., Luther, R.B., Knowles-Ward, L., Kelly, D.L., Weil, R.J., 2014. Improving and measuring inpatient documentation of medical care within the MS-DRG system: education, monitoring, and normalized case mix index. Perspectives in Health Information Management 11(Summer), 1c. PMID: 25214820.

[5] Frazee, R.C., Matejicka, A.V., Abernathy, S.W., Davis, M., Isbell, T.S., Regner, J.L., Smith, R.W., Jupiter, D.C., Papaconstantinou, H.T., 2015. Concurrent chart review provides more accurate documentation and increased calculated case mix index, severity of illness, and risk of mortality. Journal of the American College of Surgeons 220(4), 652-656. https://doi.org/10.1016/j.jamcollsurg.2014.12.036

[6] Fox, N., Swierczynski, P., Willcutt, R., Elberfeld, A., Mazzarelli, A.J., 2016. Lost in translation: focused documentation improvement benefits trauma surgeons. Injury 47(9), 1919-1923. https://doi.org/10.1016/j.injury.2016.04.035

[7] Nouraei, S.A.R., Virk, J.S., Hudovsky, A., Wathen, C., Darzi, A., Parsons, D., 2016. Accuracy of clinician-clinical coder information handover following acute medical admissions: implication for using administrative datasets in clinical outcomes management. Journal of Public Health 38(2), 352-362. https://doi.org/10.1093/pubmed/fdv041

[8] Tsopra, R., Peckham, D., Beirne, P., et al., 2018. The impact of three discharge coding methods on the accuracy of diagnostic coding and hospital reimbursement for inpatient medical care. International Journal of Medical Informatics 115, 35-42. https://doi.org/10.1016/j.ijmedinf.2018.03.015

[9] Liu, W., Walsh, T., 2018. The impact of implementation of a clinically integrated problem-based neonatal electronic health record on documentation metrics, provider satisfaction, and hospital reimbursement: a quality improvement project. JMIR Medical Informatics 6(2), e40. https://doi.org/10.2196/medinform.9776

[10] Elsamadicy, A.A., Sergesketter, A., Sampson, J.H., Gottfried, O.N., 2018. Institutional review of mortality in 5434 consecutive neurosurgery patients: are we improving? Neurosurgery 83(6), 1269-1276. https://doi.org/10.1093/neuros/nyx603

[11] Zafirah, S.A., Nur, A.M., Puteh, S.E.W., Aljunid, S.M., 2018. Potential loss of revenue due to errors in clinical coding during the implementation of the Malaysia diagnosis related group (MY-DRG) Casemix system in a teaching hospital in Malaysia. BMC Health Services Research 18(1), 1-11. https://doi.org/10.1186/s12913-018-2843-1

[12] Tsopra, R., Wyatt, J.C., Beirne, P., Rodger, K., Callister, M., Ghosh, D., Clifton, I.J., Whitaker, P., Peckham, D., 2019. Level of accuracy of diagnoses recorded in discharge summaries: a cohort study in three respiratory wards. Journal of Evaluation in Clinical Practice 25(1), 36-43. https://doi.org/10.1111/jep.13020

[13] DeCicco, D., Krupica, T.M., Pellegrino, R., Dimachkie, Z.O., 2022. Hospital-wide intervention in billing and coding to capture complexity of care at an academic referral center. Journal of Healthcare Management 67(6), 416-424. https://doi.org/10.1097/JHM-D-21-00213

[14] Nouraei, S.A.R., Hudovsky, A., Frampton, A.E., Mufti, U., White, N.B., Wathen, C.G., Sandhu, G.S., Darzi, A., 2015. A study of clinical coding accuracy in surgery: implications for the use of administrative big data for outcomes management. Annals of Surgery 261(6), 1096-1107. https://doi.org/10.1097/SLA.0000000000000851

[15] Reyes, C., Greenbaum, A., Porto, C., Russell, J.C., 2017. Implementation of a clinical documentation improvement curriculum improves quality metrics and hospital charges in an academic surgery department. Journal of the American College of Surgeons 224(3), 301-309. https://doi.org/10.1016/j.jamcollsurg.2016.11.010

[16] Aiello, F.A., Judelson, D.R., Durgin, J.M., et al., 2018. A physician-led initiative to improve clinical documentation results in improved health care documentation, case mix index, and increased contribution margin. Journal of Vascular Surgery 68(5), 1524-1532. https://doi.org/10.1016/j.jvs.2018.02.038

[17] Elkbuli, A., Godelman, S., Miller, A., et al., 2018. Improved clinical documentation leads to superior reportable outcomes: an accurate representation of patient's clinical status. International Journal of Surgery 53, 288-291. https://doi.org/10.1016/j.ijsu.2018.03.081

[18] Radhakrishnan, S.T., Perry, R., Misra, S., Ray, S., Ruban, A., Quayson, B.I., Fofaria, R., Hudovsky, A., Williams, H.R.T., 2024. Targeted education for clinicians and clinical coding staff improves the accuracy of clinical coding: a quality improvement project. Future Healthcare Journal 11(1), 100127. https://doi.org/10.7861/fhj.2023-0021

[19] Burns, E.M., Rigby, E., Mamidanna, R., Bottle, A., Aylin, P., Ziprin, P., Faiz, O.D., 2012. Systematic review of discharge coding accuracy. Journal of Public Health 34(1), 138-148. https://doi.org/10.1093/pubmed/fdr054

[20] Campbell, S.E., Campbell, M.K., Grimshaw, J.M., Walker, A.E., 2001. A systematic review of discharge coding accuracy. Journal of Public Health Medicine 23(3), 205-211. https://doi.org/10.1093/pubmed/23.3.205

[21] Spurgeon, A., Hiser, B., Hafley, C., Litofsky, N.S., 2011. Does improving medical record documentation better reflect severity of illness in neurosurgical patients? Clinical Neurosurgery 58, 155.

[22] Pongpirul, K., Walker, D.G., Winch, P.J., Robinson, C., 2011. A qualitative study of DRG coding practice in hospitals under the Thai Universal Coverage Scheme. BMC Health Services Research 11, 71. https://doi.org/10.1186/1472-6963-11-71

[23] Rangachari, P., 2007. Coding for quality measurement: the relationship between hospital structural characteristics and coding accuracy from the perspective of quality measurement. Perspectives in Health Information Management 4, 3.

[24] Page, M.J., McKenzie, J.E., Bossuyt, P.M., et al., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, n71. https://doi.org/10.1136/bmj.n71