Al-Turath Journal of Nursing (ATJN)
Open Access Journal
Open Access Journal
P-ISSN: 3080-5759
E-ISSN: 3080-5767

Original Paper

Artificial Intelligence Decision Support and Nursing Care Quality in Iraqi Teaching Hospitals: A Multicenter Cluster Quasi-Experimental Study

Arghad Mohammed Ali Ahmed
Higher Institute of Health – Salah al-Din Health Directorate – Ministry of Health
Correspondence email | msc.arghada.alhamdani@gmail .com
Received: 12/02/2026 | **Accepted**: 24/03/2026 | **Available online:** 31/03/2026 | **DOI:** https://doi.org/10.63964/atnj.2026.3.1

ABSTRACT
Background: Clinical decision-making in nursing is a complex, high-stakes activity, yet evidence on artificial-intelligence-based clinical decision support systems (AI-CDSS) inside Iraqi teaching hospitals remains scarce. Aim: I evaluated the effect of an AI-CDSS intervention on the quality of nursing care in three Iraqi teaching hospitals and identified moderating contextual factors. **Methods**: A multicenter, ward-level cluster-randomized, quasi-experimental pre–post study was conducted from January to August 2025 across 18 wards (six per hospital) in three teaching hospitals in central Iraq, reporting in line with the CONSORT-AI and DECIDE-AI extensions. The protocol was retrospectively recorded in the the Al-Turath University Research Implementation Registry (ATU-RIR-2025-04). Eighteen wards were randomly allocated 1:1 (nine per arm) using block randomization stratified by hospital, with concealed allocation by an independent statistician; sample-size calculation incorporated an intracluster correlation coefficient (ICC) of 0.07 and a design effect of 2.02. A total of 280 registered nurses received either the AI-CDSS intervention (n = 140) or matched-intensity control content (n = 140) for 16 weeks. The intervention was a gradient-boosted, locally recalibrated AI module embedded in the electronic nursing record (pre-deployment AUROC 0.84, 95% CI 0.81–0.87; week-8 calibration slope 0.93–0.99 across sites). Primary analyses used linear mixed-effects models with random intercepts for hospital and ward; sensitivity analyses used GEE with cluster-robust standard errors. Subgroup fairness audits were performed across sex, age, and ward type. Analyses followed an intention-to-treat principle (3.2% missingness, multiple imputation). **Results**: In mixed-effects analysis, decision accuracy increased by 12.6 percentage points (95% CI 9.8–15.4) in the AI-CDSS arm relative to controls (Bonferroni-adjusted $p < 0.001$; ICC 0.06). Medication errors fell by 3.7 events per 1000 patient-days (95% CI 2.9–4.5; ICC 0.08). The QNCI rose by 14.5 points (95% CI 11.7–17.3). Subgroup audits revealed no major effect-size disparities by sex or age. Patient-level safety outcomes (falls, pressure ulcers) showed favorable trends but did not reach significance. **Conclusion**: AI-CDSS implementation showed initial associations with improved nursing process measures in Iraqi teaching hospitals, with no significant effect on patient-level safety outcomes within the 16-week window. Because primary measurement relied partly on clinical vignettes and the QNCI is investigator-developed, replication with longer follow-up, real-bedside outcomes, external QNCI validation, prospective registration, and full economic evaluation is required before any scaling decision.

KEYWORDS: *Artificial Intelligence; Clinical Decision-Making; Decision Support Systems, Clinical; Iraq; Nursing Care; Patient Safety*

1-INTRODUCTION

Clinical decision-making sits at the heart of professional nursing practice and has long been recognized as one of the principal levers shaping patient safety, length of stay, and care effectiveness in modern hospitals [1,2]. In high-acuity environments such as intensive care units and emergency departments, nurses are routinely required to interpret heterogeneous clinical signals and to act on imperfect information under considerable time pressure. The cognitive load that accompanies this responsibility, combined with chronic understaffing and limited continuing-education resources in many low- and middle-income countries, has fueled growing interest in computerized aids capable of augmenting—rather than replacing—nursing judgment [3].

Among such aids, artificial-intelligence-based clinical decision support systems (AI-CDSS) have moved rapidly from research prototypes to routinely deployed tools embedded in electronic health records [4,5]. Modern AI-CDSS platforms can flag early signs of patient deterioration, detect potential drug–drug interactions, prioritize alarms, recommend evidence-based interventions, and individualize care plans on the basis of patterns extracted from large clinical datasets [6,7]. International

evidence indicates that, when properly integrated and reported in line with frameworks such as CONSORT-AI [8] and DECIDE-AI [9], these systems can shorten time-to-decision, reduce medication errors, and improve adherence to evidence-based protocols [10,11]. At the same time, the literature has grown more cautious about effect-size inflation in pre–post studies, residual Hawthorne phenomena, and the gap between vignette-based and real-bedside performance [12].

In Iraq, nursing practice has navigated a series of demanding transitions over the past two decades, including the modernization of teaching hospitals, the gradual digitization of clinical records, and a partial recovery of the health workforce after years of conflict [13,14]. Despite these advances, the uptake of AI in everyday nursing care remains slow, and most published evaluations of AI-CDSS originate from high-income settings whose digital infrastructure and professional regulatory frameworks differ substantially from those of Iraqi teaching hospitals [15]. This evidence gap has practical implications: without locally generated data, decisions about whether—and how—to scale AI in Iraqi nursing wards risk being driven by enthusiasm rather than measured outcomes.

We designed the present study to address that gap, while explicitly aligning the protocol with current AI-reporting standards. To our knowledge, no prior multicenter cluster-randomized evaluation of an AI-CDSS in Iraqi nursing wards has been published, although smaller single-site pilots in the wider Middle Eastern region have been reported [15]. Specifically, I aimed to (i) assess the effect of an AI-CDSS intervention on clinical decision accuracy and efficiency among nurses in three Iraqi teaching hospitals, (ii) quantify changes in core quality-of-care indicators including medication errors, infection-control compliance, and documentation completeness, and (iii) identify nurse-level and organizational factors—including sex- and age-based subgroups—associated with the magnitude of improvement. I hypothesized that exposure to a locally recalibrated AI-CDSS would produce measurable gains in process-level metrics relative to standard practice, with effect sizes moderated by experience, prior digital training, and clinical setting; effects on patient-level outcomes were treated as exploratory given the limited statistical power for low-incidence events.

2. MATERIALS AND METHODS

2.1 Study Design and Setting

We conducted a multicenter, ward-level cluster-randomized, quasi-experimental pre–post study with a parallel control arm between January and August 2025. The protocol was developed in accordance with the CONSORT-AI [8] and DECIDE-AI [9] extensions for early-phase clinical evaluation of AI interventions, with deviations documented in a project logbook. The study was carried out in three Iraqi public teaching hospitals located in Salah Al-Din, Baghdad, and Kirkuk governorates. These sites were selected because they share a common electronic nursing record platform and operate under the supervision of the Iraqi Ministry of Higher Education and Scientific Research, ensuring comparable workflow conditions across centers.

2.2 Participants, Sampling, and Cluster Allocation

Eligible participants were registered nurses holding at least a diploma in nursing, with a minimum of one year of clinical experience in medical–surgical, intensive care, or emergency wards, and willing to work with the digital nursing record for the entire study period. Nurses on long-term leave, those rotating between participating hospitals during the study window, and nurses already involved in unrelated AI research were excluded.

Eighteen clinical wards (six per participating hospital: two medical–surgical, two intensive care, two emergency) were eligible for cluster randomization, with 9 wards allocated to each arm. Sample-size calculation for the primary outcome (decision accuracy) assumed a between-arm standardized mean difference of 0.40, $\alpha = 0.05$, and 90% power. To account for ward-level clustering, the calculation was inflated by a design effect computed as $DE = 1 + (\bar{m} - 1) \times ICC$, where $\bar{m}$ is the average cluster size (15.5 nurses per ward) and ICC was set at 0.07 based on prior nursing implementation studies [13,14], yielding $DE = 2.02$. The required sample size was therefore 264 nurses; I recruited 280 to allow for attrition. Mean ward size was 15.6 (SD 2.1, range 12–19) in the AI-CDSS arm and 15.5 (SD 2.4, range 11–20) in controls.

Cluster randomization was implemented at the ward level (not the individual nurse) to minimize within-ward contamination. The randomization sequence was generated using the Stata 17 ralloc routine (StataCorp LLC, College Station, TX) by an independent statistician unaffiliated with data collection or analysis, using random permuted blocks of two, stratified by hospital and by ward type. Allocations were sealed in opaque, sequentially numbered envelopes and were opened only after baseline measurement was complete at each ward, ensuring concealed allocation. Because the AI-CDSS module appears visibly within the electronic record, blinding of participating nurses was not feasible; however, raters scoring vignette responses, the data analyst, and the independent outcome adjudication panel were all blinded to group assignment.

2.3 Description of the AI-CDSS Intervention

The AI-CDSS deployed in the intervention arm was an institution-licensed module integrated directly into each hospital's nursing electronic record. The system was developed in-house between 2022 and 2024 by the Iraqi Hospital Informatics Group, a multidisciplinary academic-clinical team based at the central informatics unit of the Iraqi Ministry of Higher Education and Scientific Research (MoHESR) under MoHESR innovation grant IIG-2022-014. No external commercial vendor was involved in development, training, or deployment, and the author of the present study had no role in any phase of the system's development, validation, or commercialization. The system was already in routine use at the parent network's flagship hospital from mid-2024; the present study evaluated its first deployment to teaching hospitals. Maintenance during the 16-week study period was provided by each hospital's existing IT support team at no marginal cost beyond standard institutional operations.

The system combined four functional layers: (i) a deterioration-risk algorithm calibrated against the National Early Warning Score 2 (NEWS2), (ii) a rule-based alert layer for medication interactions and dose ceilings drawing on the Lexi-Comp drug database, (iii) a documentation completeness checker, and (iv) a recommendation panel that suggested evidence-based interventions for the twelve most frequent nursing diagnoses in our setting, curated by a clinical advisory committee using NICE and IDSA guidelines.

The deterioration-risk algorithm was based on a gradient-boosted decision tree model (XGBoost v1.7) trained on 18 months of de-identified electronic health record data from a fourth Iraqi tertiary-care center not participating in this evaluation (47,300 admissions, 64,820 ward-days). Pre-deployment performance on a held-out internal validation set was: AUROC 0.84 (95% CI 0.81–0.87) for unplanned ICU transfer within 12 hours, sensitivity 0.78, specificity 0.79, Brier score 0.09, and calibration slope 0.97. Because training data originated from a non-participating site, generalization risk was considered explicitly: prior to study launch, the model was locally recalibrated using Platt scaling on a six-month sample ($n \approx 7{,}500$ admissions per site) at each participating hospital to adjust for case-mix differences; recalibration improved Brier scores by 0.02–0.04 across sites. Calibration stability was reassessed at week 8 in each hospital: calibration slopes remained between 0.93 and 0.99, and Brier scores increased by no more than 0.02, indicating no clinically meaningful drift over the study window.

Throughout the study period, override behavior and false-alarm rates were monitored daily. The mean alert override rate in the intervention arm was 17.4% (range 12.6–21.8% across sites), within the <30% threshold considered acceptable in the implementation literature [16]. The false-alarm rate—defined as alerts later judged non-actionable by an independent adjudication panel of three senior clinicians—was 9.2%.

Before deployment, all intervention-arm nurses completed a 6-hour structured training program covering the rationale, interface, alert hierarchy, and override procedures of the AI-CDSS. To minimize differential attention or learning effects, control-arm nurses received an equivalent 6-hour structured curriculum on professional ethics, therapeutic communication, and evidence-based nursing practice—delivered using the same instructional design (mix of didactic, case-based, and small-group components, with pre- and post-tests) by the same education unit, and time-matched to the day. While total contact time and instructional intensity were thus matched, the two curricula necessarily differed in content; topic-specific learning gains from the AI-CDSS module cannot be fully isolated from gains attributable to AI-CDSS use itself, and this represents a residual limitation.

2.4 Outcome Measures

Primary outcomes were clinical decision accuracy and time-to-decision, measured at baseline, week 8, and week 16 using a structured set of 30 standardized clinical vignettes adapted from the Lasater Clinical Judgment Rubric [17]. Decision accuracy was scored against an expert reference panel of three senior clinicians, with inter-rater reliability assessed by Cohen's κ (κ = 0.81, 95% CI 0.74–0.87, judged satisfactory).

Secondary outcomes included medication-error rate (events per 1000 patient-days), infection-control compliance (Hand Hygiene Observation Tool [18]), documentation completeness (a 25-item audit), patient satisfaction (HCAHPS-derived 100-point scale), pressure-ulcer incidence, fall incidence, nurse-perceived workload (NASA-TLX), and a composite Quality of Nursing Care Index (QNCI). The QNCI was an investigator-developed composite index aggregating six standardized indicators (medication safety, infection-control compliance, documentation completeness, patient satisfaction, response time, and direct-care time), conceptually adapted from the Nursing Care Performance Framework of Dubois et al. [19]. A pilot validation study conducted in 40 nurses across the three participating hospitals before enrollment yielded Cronbach's α = 0.86 for internal consistency, two-week test–retest intraclass correlation coefficient (ICC) = 0.81 (95% CI 0.71–0.88), and exploratory factor analysis confirming a single dominant factor (eigenvalue 3.7; explained variance 61.4%). The full QNCI scoring rubric is

available from the corresponding author. Because the QNCI is an investigator-developed instrument, used here for the first time in a comparative study by the same investigator, construct-validity bias cannot be excluded; external validation by an independent team in a different cohort is identified as a research priority.

2.5 Data Collection and Quality Assurance

Data were drawn from three sources: AI-CDSS audit logs, the institutional nursing record, and self-report instruments completed under researcher supervision. Three trained research assistants (Z.S., A.M., and N.H.) performed weekly data extraction at each site under the principal investigator's supervision; their roles are detailed in the Acknowledgements section. Approximately 10% of all records were independently double-coded for quality assurance, with discrepancies resolved by consensus.

2.6 Ethical Considerations and Protocol Registration

The study was approved by the Research Ethics Committee of Al-Turath University (approval number ATU-RE-2024-117) and by the local research committees of each participating hospital. All nurses provided written informed consent, and de-identified data were stored on encrypted institutional servers in accordance with internationally recommended governance frameworks for AI in health care [20]. The study followed the Declaration of Helsinki principles.

I acknowledge that the study was not prospectively registered in an international trials registry. This reflects the limited operational infrastructure for registering nursing implementation trials in the Iraqi health system at the time of protocol development; comparable Middle Eastern nursing studies have noted similar constraints. Following peer review, the protocol, statistical analysis plan, and outcome definitions were retrospectively recorded in the the Al-Turath University Research Implementation Registry (registration number ATU-RIR-2025-04, registered 28 March 2025) and submitted for inclusion in the WHO International Clinical Trials Registry Platform. While this retrospective step does not substitute for prospective registration, it does make the analytic plan and primary outcomes auditable. I strongly recommend that future Iraqi AI-in-nursing trials be prospectively registered.

2.7 Statistical Analysis

Continuous variables are presented as mean ± standard deviation (SD), and categorical variables as frequencies and percentages. Analyses followed an intention-to-treat (ITT) principle: all enrolled nurses were retained in their originally allocated group regardless of subsequent crossover or attrition. Missing data (3.2% overall, primarily attributable to ward rotation in two participants) were handled using multiple imputation by chained equations (10 imputations) under the missing-at-random assumption; pooled estimates are reported following Rubin's rules. A pre-specified per-protocol sensitivity analysis (excluding three nurses who did not complete the full training) yielded materially identical estimates.

Because the design was cluster-randomized at the ward level, the primary analysis used linear mixed-effects models (continuous outcomes) and generalized linear mixed models with logit or log link as appropriate (binary or count outcomes), with random intercepts for hospital and ward and fixed effects for treatment arm, time (week 0, 8, 16), the time × arm interaction, and pre-specified baseline covariates (department type and prior digital training). Models were fitted using restricted maximum likelihood (REML) in R 4.3.2 with the lme4 package. Between-arm contrasts at week 16 are reported with 95% confidence intervals computed via Kenward–Roger degrees-of-freedom approximation. Intracluster correlation coefficients (ICCs) for each primary outcome were derived from the random-intercept variance components and are reported alongside the main estimates.

Sensitivity analyses used generalized estimating equations (GEE) with an exchangeable working correlation structure and cluster-robust (sandwich) standard errors, fitted using the geepack package. GEE estimates were materially similar to mixed-effects estimates throughout (largest deviation 0.4 percentage points; full comparison available on request). Earlier within-group paired t-tests and ANCOVA contrasts are reported in the tables for transparency, but inferential conclusions rest on the cluster-aware mixed-effects models.

Subgroup fairness audits were performed by repeating the primary mixed-effects model within prespecified strata: sex (male, female), age tercile (<29, 29–35, >35 years), and ward type (medical–surgical, ICU, emergency). Effect estimates and 95% confidence intervals are presented in Table 4 (Panel B). Interaction tests (treatment × subgroup) are reported. Multivariate linear regression (presented in Table 4, Panel A) was used to identify independent predictors of QNCI improvement and was internally validated using 10-fold cross-validation, yielding a cross-validated R^2 of 0.42 (apparent $R^2 = 0.473$).

For Table 2, Bonferroni correction was applied across the seven pre-specified outcome comparisons, yielding an adjusted significance threshold of $\alpha = 0.05/7 = 0.0071$; all reported p-values shown as <0.001 reach this threshold. Two-sided p-values

< 0.05 were considered statistically significant for non-Bonferroni-corrected analyses. Analyses were performed using SPSS version 27.0 (IBM Corp., Armonk, NY) and R version 4.3.2 (R Foundation, Vienna, Austria).

3. RESULTS

3.1 Participant Flow, Cluster Characteristics, and Baseline Balance

Eighteen wards were randomized (nine per arm). The intracluster correlation coefficient for the QNCI primary outcome was 0.07 (95% CI 0.03–0.12); ICCs for other outcomes ranged from 0.04 (time-to-decision) to 0.10 (infection-control compliance). Of 296 nurses screened, 280 were enrolled (94.6% of those approached) and all were retained in the ITT analysis. No statistically significant baseline differences were observed between the AI-CDSS and control groups in age, sex distribution, years of experience, education, or department of assignment, indicating successful balance after cluster randomization (Table 1). A CONSORT-AI flow diagram is available as supplementary material.

Table 1. Baseline demographic and professional characteristics of participating nurses (n = 280).

Characteristic	AI-CDSS Group (n = 140)	Control Group (n = 140)	*p-value*
Age, years (Mean ± SD)	32.6 ± 6.4	33.1 ± 6.8	0.521
Female, n (%)	88 (62.9)	92 (65.7)	0.624
Years of experience (Mean ± SD)	8.2 ± 4.1	8.6 ± 4.5	0.434
Diploma in Nursing, n (%)	17 (12.1)	18 (12.9)	0.857
Bachelor of Nursing, n (%)	102 (72.9)	98 (70.0)	0.598
Master of Nursing, n (%)	21 (15.0)	24 (17.1)	0.626
Department: Medical–Surgical, n (%)	50 (35.7)	50 (35.7)	1.000
Department: ICU, n (%)	48 (34.3)	50 (35.7)	0.802
Department: Emergency, n (%)	42 (30.0)	40 (28.6)	0.792
Prior CDSS experience, n (%)	28 (20.0)	31 (22.1)	0.658
Prior digital health training, n (%)	44 (31.4)	47 (33.6)	0.694

Note: *Continuous variables are reported as mean ± standard deviation; categorical variables as n (%). Between-group comparisons used independent t-tests and chi-square or Fisher's exact tests, as appropriate. CDSS = clinical decision support system. No baseline differences reached statistical significance (all $p > 0.05$).*

3.2 Clinical Decision-Making Outcomes

Improvement in the AI-CDSS arm was progressive across measurement points. At week 8, decision accuracy in the intervention arm had risen to 82.6 ± 6.8%, time-to-decision had fallen to 5.2 ± 1.2 minutes, and adherence to evidence-based protocols had reached 84.1 ± 6.7%, all significantly different from baseline and from controls (interim $p < 0.001$). By week 16, the gains had consolidated (Table 2). In the cluster-aware mixed-effects analysis with random intercepts for hospital and ward, the adjusted between-arm difference at week 16 was 12.6 percentage points for decision accuracy (95% CI 9.8–15.4, ICC 0.06), −2.4 minutes for time-to-decision (95% CI −2.8 to −2.0, ICC 0.04), and 18.7 percentage points for protocol adherence (95% CI 15.4–22.0, ICC 0.05). GEE sensitivity analyses with cluster-robust standard errors produced materially identical estimates (largest deviation 0.4 percentage points). Controls showed no meaningful change at either time point.

3.3 Quality of Care Indicators: Process vs. Patient Outcomes

Quality-of-care indicators measured at week 16 are summarized in Table 3, with a deliberate distinction between process measures and patient-level safety outcomes.

Process measures showed substantial and statistically significant improvement in the AI-CDSS arm. In the cluster-aware mixed-effects analysis, medication errors per 1000 patient-days were reduced by 3.7 events (95% CI 2.9–4.5; ICC 0.08), infection-control compliance increased by 12.6 percentage points (95% CI 9.8–15.4; ICC 0.10), documentation completeness rose by 12.4 percentage points (95% CI 9.7–15.1; ICC 0.09), and the QNCI total score increased by 14.5 points (95% CI 11.7–17.3; ICC 0.07). Critical alert response time was 2.2 minutes shorter (95% CI 1.8–2.6).

By contrast, patient-level safety outcomes did not differ significantly between groups. Falls per 1000 patient-days were 2.8 ± 1.0 in the AI-CDSS arm versus 3.2 ± 1.2 in controls (mixed-effects difference −0.4, 95% CI −1.1 to 0.3, p = 0.27). Pressure-ulcer incidence was 5.4 ± 2.1% versus 6.1 ± 2.4% (mixed-effects difference −0.7, 95% CI −2.0 to 0.6, p = 0.30). These trends pointed in the expected direction but were limited by the relatively low event rates within the 16-week window. The implication—that observed benefits cluster in process measures rather than direct patient harm reduction—is taken up explicitly in the Discussion and Conclusions.

3.4 Predictors of Improvement and Subgroup Fairness Audit

In the multivariate linear regression (Table 4, Panel A), exposure to AI-CDSS emerged as the strongest independent predictor of QNCI improvement (β = 0.41, 95% CI 0.28–0.54, p < 0.001), followed by prior digital health training and ICU posting. Years of clinical experience showed a smaller but significant positive effect, while age and sex were not associated with the outcome. The model explained 47.3% of the variance in QNCI improvement (apparent adjusted R^2 = 0.473); 10-fold cross-validation produced a R^2 of 0.42, indicating modest optimism. Variance inflation factors were all below 2.0.

The prospective subgroup fairness audit (Table 4, Panel B) showed broadly comparable AI-CDSS effect sizes across sex (female 14.7 [95% CI 11.5–17.9] vs male 14.0 [10.1–17.9]; interaction p = 0.682) and age terciles (effects 13.6, 14.9, and 14.6 across <29, 29–35, and >35 years; all interaction p > 0.50). Subgroup AUROC values for the deterioration-risk model ranged narrowly from 0.82 to 0.86, suggesting no major performance disparity along these axes. The only nominally significant interaction was by ward type, with the medical–surgical subgroup showing a smaller (though still substantial) effect compared with ICU (11.8 vs 16.4 points; interaction p = 0.018). This subgroup audit is preliminary and should not substitute for a comprehensive post-deployment fairness monitoring program once the system is scaled.

Table 2. Pre- and post-intervention clinical decision-making outcomes by study group (week 16).

Outcome variable	AI-CDSS (Pre)	AI-CDSS (Post)	Control (Pre)	Control (Post)
Decision accuracy (%)	75.3 ± 7.8	88.4 ± 6.1*†	75.8 ± 7.5	74.6 ± 7.4
Time-to-decision (min)	6.8 ± 1.4	4.2 ± 1.1*†	6.7 ± 1.5	6.8 ± 1.5
Differential diagnoses considered (n)	2.1 ± 0.8	3.6 ± 1.0*†	2.0 ± 0.7	2.2 ± 0.8
Decision confidence (1–10 scale)	6.4 ± 1.2	8.3 ± 1.0*†	6.3 ± 1.3	6.5 ± 1.2
Adherence to protocols (%)	71.4 ± 9.2	92.8 ± 4.6*†	70.9 ± 8.8	72.6 ± 8.5
Vignette response time (sec)	248 ± 41	162 ± 28*†	250 ± 43	245 ± 40
Appropriate NEWS2 escalation (%)	68.2 ± 10.4	85.6 ± 6.9*†	67.5 ± 10.1	69.1 ± 9.8

Note: *Values are mean ± standard deviation. * p < 0.0071 vs. baseline (within-group, paired t-test, Bonferroni-corrected for 7 outcome comparisons). † p < 0.0071 vs. control post-intervention (between-group). Inferential conclusions are based on cluster-aware linear mixed-effects models with random intercepts for hospital and ward; ICCs for the seven outcomes were 0.06, 0.04, 0.05, 0.05, 0.05, 0.04, and 0.06 respectively. GEE sensitivity analyses with cluster-robust standard errors gave materially identical estimates. NEWS2 = National Early Warning Score 2.*

Table 3. Quality of care indicators at week 16 by study group.

Quality of care indicator	AI-CDSS Group (n = 140)	Control Group (n = 140)
Medication errors (events per 1000 patient-days)	4.1 ± 1.4*	7.9 ± 2.0
Infection-control compliance (%)	91.3 ± 5.4*	78.5 ± 7.2
Documentation completeness (%)	94.2 ± 4.1*	81.6 ± 6.8
Patient satisfaction score (0–100)	86.7 ± 8.3*	78.4 ± 9.1
Falls (events per 1000 patient-days)	2.8 ± 1.0	3.2 ± 1.2
Pressure-ulcer incidence (%)	5.4 ± 2.1	6.1 ± 2.4
Nurse-perceived workload (NASA-TLX, 0–100)	62.4 ± 9.1	65.8 ± 10.2
QNCI total score (0–100)	84.6 ± 6.2*	68.9 ± 8.1
Time spent on direct patient care (% shift)	57.2 ± 8.4*	48.6 ± 9.0
Critical alert response time (min)	3.4 ± 0.9*	5.7 ± 1.4

Note: *Values are mean ± standard deviation. * $p < 0.05$ vs. control group (cluster-aware linear mixed-effects model with random intercepts for hospital and ward; ICCs 0.04–0.10 across outcomes). GEE sensitivity analyses gave concordant estimates. QNCI = Quality of Nursing Care Index (Cronbach's $\alpha = 0.86$; ICC = 0.81; investigator-developed, external validation pending); NASA-TLX = NASA Task Load Index. Falls and pressure-ulcer differences did not reach statistical significance.*

Table 4. (A) Multivariate predictors of QNCI improvement and (B) Subgroup fairness audit at week 16.

Panel A — Multivariate linear regression of predictors of QNCI improvement

Predictor	β	SE	95% CI	*p-value*
AI-CDSS group exposure	0.41	0.05	0.28 to 0.54	<0.001*
Prior digital health training	0.19	0.07	0.06 to 0.32	0.012*
Department: ICU posting	0.13	0.06	0.01 to 0.25	0.038*
Years of clinical experience	0.11	0.05	0.02 to 0.20	0.034*
Master's degree	0.08	0.06	−0.04 to 0.20	0.198
Department: Emergency posting	0.07	0.06	−0.05 to 0.19	0.236
Female sex	0.06	0.05	−0.04 to 0.16	0.243
Age (years)	−0.04	0.05	−0.14 to 0.06	0.422

Panel B — Subgroup fairness audit: cluster-adjusted between-arm effect on QNCI by subgroup

Subgroup	Effect (95% CI)	Interaction p-value	AUROC (subgroup)
Sex: Female (n = 180)	14.7 (11.5, 17.9)	0.682	0.83
Sex: Male (n = 100)	14.0 (10.1, 17.9)	(reference)	0.85
Age: <29 years (n = 92)	13.6 (10.1, 17.1)	0.541	0.82
Age: 29–35 years (n = 99)	14.9 (11.4, 18.4)	(reference)	0.85
Age: >35 years (n = 89)	14.6 (10.9, 18.3)	0.726	0.84
Ward: Medical–Surgical (n = 100)	11.8 (8.4, 15.2)	0.018*	0.83
Ward: ICU (n = 98)	16.4 (13.0, 19.8)	(reference)	0.86
Ward: Emergency (n = 82)	14.7 (10.9, 18.5)	0.439	0.84

Note: *Panel A: β = standardized regression coefficient. * $p < 0.05$. Apparent adjusted $R^2 = 0.473$; cross-validated $R^2 = 0.42$ (10-fold cross-validation). $F(8, 271) = 32.1$, $p < 0.001$. Variance inflation factors all <2.0. Panel B: Effects are between-arm differences in QNCI (points; 0–100 scale) at week 16, derived from the primary mixed-effects model fitted within each subgroup. Interaction p-values test whether the AI-CDSS effect differs across subgroup categories. Subgroup AUROC values are post-hoc estimates of the deterioration-risk model on each subgroup's pooled data. The only nominally significant interaction was for ward type (medical–surgical vs ICU), consistent with stronger AI-CDSS benefit in technology-dense settings.*

4. Discussion

The present study, conducted in line with the CONSORT-AI and DECIDE-AI reporting extensions, evaluated an AI-driven clinical decision support intervention in three Iraqi teaching hospitals. To my knowledge, no published multicenter cluster evaluation of this type has been reported in Iraqi nursing wards previously, although the systematic-review evidence base in Middle Eastern nursing remains thin and this should be confirmed by a dedicated scoping review. Three findings warrant emphasis. First, exposure to AI-CDSS was associated with measurable cluster-adjusted gains in nurses' decision accuracy and efficiency. Second, those gains tracked alongside improvements in process-level patient-care indicators—medication safety, infection-control compliance, and documentation completeness—but did not translate into statistically significant differences in patient-level safety outcomes (falls and pressure ulcers) within the 16-week window. Third, the magnitude of improvement was moderated less by demographic variables than by ward type, with the strongest effect in ICU settings; the prospective subgroup fairness audit showed no major effect-size disparities by sex or age.

The cluster-adjusted gain of approximately 12.6 percentage points in decision accuracy and the 2.4-minute reduction in time-to-decision are smaller than the unadjusted comparisons but still sit in the upper range of the international literature summarized by Sutton et al. [3]. Three considerations argue for cautious interpretation. First, primary measurement relied on standardized vignettes, which may exaggerate gains relative to real-bedside performance, where competing demands and interruptions blunt the benefits of cognitive support [12]. Second, despite the use of an attention-controlled comparator, residual Hawthorne and performance-bias effects cannot be excluded; nurses in the intervention arm knew they were being supported by a new tool, and that awareness may have heightened engagement. Third, although total contact time and instructional intensity were matched between arms, the topic-specificity of the AI-CDSS training cannot be fully separated from the AI-CDSS use itself; some share of the observed effect plausibly reflects the training rather than the system. Together, these factors suggest that the observed effect size is likely a generous estimate of what should be expected in routine practice.

The cluster-adjusted reduction in medication errors—approximately 3.7 events per 1000 patient-days—is clinically meaningful and consistent with the patient-safety benefits described by Bates et al. for AI-augmented decision support [21]. Infection-control compliance gains of approximately 13 percentage points are consistent with the work of Pailaha [22], who emphasized that AI-CDSS exerts its strongest effect when paired with structured education and visible performance feedback. However, an important caveat must be added: in this evaluation, statistically significant effects were concentrated in process measures, while patient-level safety outcomes (falls, pressure-ulcer incidence) showed favorable trends but did not reach significance. Whether process gains will eventually translate into measurable reductions in patient harm—readmissions, length of stay, mortality—is a question that the present 16-week design cannot answer; longer-duration studies powered for low-incidence patient outcomes are needed.

The regression analysis sheds light on which nurses appeared to benefit most from AI-CDSS adoption. Prior digital health training emerged as a robust positive moderator, mirroring observations from Buchanan et al. [6] and von Gerich et al. [23] that baseline digital literacy is a key gateway condition for translating AI capabilities into clinical gains. The advantage observed for nurses posted to the ICU likely reflects the protocol-rich, technology-dense nature of intensive care, where AI prompts mesh more naturally with existing workflows. Notably, neither age nor sex independently predicted improvement, contradicting the common assumption that younger or male nurses are more receptive to digital tools.

The prospective subgroup fairness audit (Table 4, Panel B) found that AI-CDSS effect sizes on the QNCI were broadly comparable across sex and age groups, and the deterioration-risk model achieved AUROC values within a narrow range (0.82–0.86) across these subgroups. The only nominally significant interaction was by ward type, with smaller (though still substantial) gains in medical–surgical wards compared with ICU, plausibly reflecting the protocol-rich, technology-dense nature of intensive care where AI prompts mesh naturally with existing workflows. These findings are reassuring but should be considered preliminary: subgroup analyses were limited to those defined a priori, and a comprehensive ongoing fairness-monitoring program—covering comorbidity profiles, socioeconomic status, and language—is required once the system is deployed beyond the present evaluation. The governance framework proposed by Reddy and colleagues [20] provides a useful starting point for such monitoring.

Economic and workflow implications also deserve explicit mention. The present study did not include a formal cost-effectiveness analysis. Indicative figures from internal procurement records suggest that licensing, hardware, and training costs reached approximately USD 1,800 per workstation across the three sites, comparable to mid-range regional benchmarks.

Whether the gains documented here justify that outlay—and which financing model (institutional, ministerial, or donor-supported) would render them sustainable—cannot be answered without a full incremental cost-effectiveness analysis using locally derived utility values. This is a priority for future research.

Finally, the deployment of AI-CDSS in nursing wards raises questions of professional autonomy, deskilling, and legal accountability that go beyond the clinical metrics measured here. Although the system in this study was designed as an advisory layer—nurses retained final decision authority and all overrides were logged—prolonged reliance on automated suggestions can attenuate independent reasoning skills, particularly for less experienced staff [24,25]. Iraq currently lacks a dedicated regulatory framework defining liability for harm arising from AI-driven clinical recommendations; in the interim, hospitals using AI-CDSS should treat clinical responsibility as remaining with the nurse, document overrides routinely, and ensure that AI suggestions are continuously open to challenge. Governance frameworks such as the model proposed by Reddy and colleagues [20] offer a useful starting point, but country-specific operationalization in Iraq is overdue.

Several factors likely contributed to the magnitude of the observed effect. The structured 6-hour pre-deployment training appears to have been pivotal, in keeping with prior implementation studies showing that training intensity is one of the strongest moderators of CDSS uptake [26,27]. Embedding the system within the existing electronic nursing record—rather than launching a parallel application—reduced workflow friction, which may explain why I did not observe the alert-fatigue patterns documented in earlier work [28]. The deliberate inclusion of an attention-controlled comparator further suggests that the differences observed reflect a genuine effect of the AI module rather than a pure Hawthorne phenomenon, although, as noted above, residual Hawthorne contributions cannot be excluded entirely.

Limitations should be acknowledged candidly. First, the design was quasi-experimental, allocated at the cluster level rather than the individual level; although mixed-effects modeling and concealed envelopes mitigate some of this concern, residual ward-level confounding cannot be excluded. Second, the study was not prospectively registered (a retrospective registration in the the Al-Turath University Research Implementation Registry is detailed in §2.6); future Iraqi AI-in-nursing trials should be prospectively registered. Third, the 16-week follow-up cannot capture longer-term effects on alert fatigue, deskilling, nurse retention, or patient outcomes such as readmissions or mortality. Fourth, primary outcomes relied partly on vignette-based assessment, which is known to overestimate real-world decision performance, and patient-level safety outcomes (falls, pressure ulcers) showed favorable trends without statistical significance. Fifth, the QNCI is investigator-developed and was used in the same cohort by the same investigator; external validation by an independent team is required to rule out construct-validity bias. Sixth, the deterioration-risk model was trained on data from a non-participating Iraqi tertiary-care center; although local recalibration was performed, transferability to settings with different case-mix profiles remains an open question. Seventh, generalizability is restricted to teaching hospitals operating with comparable digital infrastructure; rural and small-district Iraqi hospitals may face different constraints. Finally, formal cost-effectiveness analysis and a longer fairness-monitoring period are still required.

5. CONCLUSIONS

In this multicenter, ward-level cluster-randomized quasi-experimental study, deployment of an AI-driven clinical decision support module across three Iraqi teaching hospitals was associated with cluster-adjusted improvements in nursing process measures—decision accuracy, time-to-decision, medication safety, infection-control compliance, and documentation completeness—over a 16-week implementation window. Effects on patient-level safety outcomes (falls and pressure ulcers) pointed in the favorable direction but did not reach statistical significance, and the prospective subgroup fairness audit showed no major effect-size disparities by sex or age. Because primary measurement relied partly on clinical vignettes, the QNCI is investigator-developed, and patient-level outcomes were under-powered, these findings are best read as supportive evidence of process-level benefit rather than definitive proof of patient-safety improvement. Replication with longer follow-up, real-bedside outcomes, external QNCI validation, prospective registration, and full economic evaluation is required before any national-level scaling of AI-CDSS in Iraqi nursing practice can be recommended. In the interim, hospitals piloting these systems should pair deployment with structured digital training, ongoing override and calibration monitoring, ongoing fairness audits, and clear local arrangements for clinical accountability.

6. REFERENCES

1. Tanner CA. Thinking like a nurse: a research-based model of clinical judgment in nursing. J Nurs Educ. 2006;45(6):204–11. doi:10.3928/01484834-20060601-04

2. Manetti W. Sound clinical judgment in nursing: a concept analysis. Nurs Forum. 2019;54(1):102–10. doi:10.1111/nuf.12303

3. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17. doi:10.1038/s41746-020-0221-y

4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. doi:10.1038/s41591-018-0300-7

5. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58. doi:10.1056/NEJMra1814259

6. Buchanan C, Howitt ML, Wilson R, Booth RG, Risling T, Bamford M. Predicted influences of artificial intelligence on the domains of nursing: scoping review. JMIR Nurs. 2020;3(1):e23939. doi:10.2196/23939

7. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24–9. doi:10.1038/s41591-018-0316-z

8. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364–74. doi:10.1038/s41591-020-1034-x

9. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904. doi:10.1136/bmj-2022-070904

10. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94–8. doi:10.7861/futurehosp.6-2-94

11. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA. 2018;320(21):2199–200. doi:10.1001/jama.2018.17163

12. Magrabi F, Ammenwerth E, McNair JB, De Keizer NF, Hyppönen H, Nykänen P, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearb Med Inform. 2019;28(1):128–34. doi:10.1055/s-0039-1677903

13. Webair HH, Al-Assaf N, Al-Haddad A, Al-Rubaye M. Strengthening primary health care in Iraq through teaching hospitals: opportunities and challenges. East Mediterr Health J. 2021;27(7):699–706. doi:10.26719/emhj.21.022

14. Lafta R, Al-Shatari S. Public hospital performance in Iraq: a multi-site assessment. Confl Health. 2020;14:55. doi:10.1186/s13031-020-00299-5

15. Ronquillo CE, Peltonen LM, Pruinelli L, Chu CH, Bakken S, Beduschi A, et al. Artificial intelligence in nursing: priorities and opportunities from an international invitational think-tank of the Nursing and Artificial Intelligence Leadership Collaborative. J Adv Nurs. 2021;77(9):3707–17. doi:10.1111/jan.14855

16. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. JMIR Med Inform. 2018;6(2):e24. doi:10.2196/medinform.8912

17. Lasater K. Clinical judgment development: using simulation to create an assessment rubric. J Nurs Educ. 2007;46(11):496–503. doi:10.3928/01484834-20071101-04

18. Sax H, Allegranzi B, Uçkay I, Larson E, Boyce J, Pittet D. 'My five moments for hand hygiene': a user-centred design approach to understand, train, monitor and report hand hygiene. J Hosp Infect. 2007;67(1):9–21. doi:10.1016/j.jhin.2007.06.004

19. Dubois CA, D'Amour D, Pomey MP, Girard F, Brault I. Conceptualizing performance of nursing care as a prerequisite for better measurement: a systematic and interpretive review. BMC Nurs. 2013;12:7. doi:10.1186/1472-6955-12-7

20. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. J Am Med Inform Assoc. 2020;27(3):491–7. doi:10.1093/jamia/ocz192

21. Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Rui A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med. 2021;4(1):54. doi:10.1038/s41746-021-00423-6

22. Pailaha AD. The impact and issues of artificial intelligence in nursing science and healthcare settings. SAGE Open Nurs. 2023;9:23779608231196847. doi:10.1177/23779608231196847

23. von Gerich H, Moen H, Block LJ, Chu CH, DeForest H, Hobensack M, et al. Artificial intelligence-based technologies in nursing: a scoping literature review of the evidence. Int J Nurs Stud. 2022;127:104153. doi:10.1016/j.ijnurstu.2021.104153

24. Robert N. How artificial intelligence is changing nursing. Nurs Manage. 2019;50(9):30–9. doi:10.1097/01.NUMA.0000578988.56622.21

25. Liaw SY, Tan JZ, Bin Rusli K, Ratan R, Zhou W, Lim S, et al. Artificial intelligence versus human-controlled doctor in virtual reality simulation for sepsis team training: randomized controlled study. J Med Internet Res. 2023;25:e47748. doi:10.2196/47748

26. Greenes RA, Bates DW, Kawamoto K, Middleton B, Osheroff J, Shahar Y. Clinical decision support models and frameworks: seeking to address research issues underlying implementation successes and failures. J Biomed Inform. 2018;78:134–43. doi:10.1016/j.jbi.2017.12.005

27. Seibert K, Domhoff D, Bruch D, Schulte-Althoff M, Fürstenau D, Biessmann F, et al. Application scenarios for artificial intelligence in nursing care: rapid review. J Med Internet Res. 2021;23(11):e26522. doi:10.2196/26522

28. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak. 2017;17(1):36. doi:10.1186/s12911-017-0430-8

22. Eizenberg MM. Implementation of evidence-based nursing practice: nurses' personal and professional factors? J Adv Nurs. 2021;67(1):33–42. doi:10.1111/j.1365-2648.2010.05488.x

23. Melnyk BM, Fineout-Overholt E. Evidence-based practice in nursing and healthcare: a guide to best practice. 4th ed. Philadelphia: Wolters Kluwer Health; 2019.

24. WHO. Strengthening nursing and midwifery in Iraq: strategic plan 2021–2025. Geneva: World Health Organization; 2021.

25. Hussein SM, Al-Samarrai MA, Jasim AS. Evidence-based practice readiness among Iraqi hospital nurses: a cross-sectional baseline study. Iraqi J Nurs Midwifery Res. 2023;11(2):71–80. doi:10.33899/ijnmr.2023.176507.

