

Evaluating The Impact of An Ai-Driven Clinical Decision Support System on Nursing Students' Accuracy in Care-Plan Formulation During Telemedicine Practicum

A Systematic Review and Evidence Synthesis Conducted in Accordance with the PRISMA 2020 Reporting Guideline

Bryant Sang {R.N, BSc., MSc.,PhD(c)} Kabarak University, Nakuru, Kenya

DOI: <https://doi.org/10.5281/zenodo.23059031>

Abstract

Care-plan formulation is among the most cognitively demanding competencies that pre-registration nursing students must master, and errors made at this formative stage may persist into qualified practice. Two contemporary developments converge on this competency: the emergence of Artificial Intelligence (AI)-driven Clinical Decision Support Systems (CDSS) capable of suggesting nursing diagnoses and interventions, and the normalization of the telemedicine practicum as a clinical learning environment. No prior synthesis has examined their intersection. The aim of this study is to identify, appraise, and synthesize the evidence on the effect of AI-driven CDSS on the accuracy of nursing students' care-plan formulation within telemedicine practicum settings. The review is designed and reported in accordance with the PRISMA 2020 statement and conducted with reference to the Cochrane Handbook and the JBI Manual for Evidence Synthesis. Eight databases and grey-literature sources searched for peer-reviewed English-language studies published between 2015 and 2026. Screening, extraction, and appraisal were performed in duplicate; risk of bias was assessed with design-appropriate instruments (RoB 2, ROBINS-I, JBI checklists, MMAT), and certainty was rated using GRADE. Heterogeneity precluded meta-analysis, and a structured narrative synthesis following SWiM guidance was undertaken. The evidence converges on a conditional benefit. Structured, taxonomy-aligned decision support raised students' diagnostic and care-plan accuracy relative to unsupported practice, with the inference model, the supervision arrangement, and learner characteristics acting as decisive moderators. Generative large language models produced substantial gains in rated plan quality but exhibited a characteristic failure mode of fluent yet taxonomy-discordant output, creating conditions in which automation bias is most hazardous. The telemedicine setting amplified both the potential benefit and the risk of uncritical reliance. The study concludes that an AI-driven CDSS functions as a conditional cognitive scaffold rather than an unconditional accelerator of accuracy. Its educational value in the telemedicine practicum is realized only where outputs are aligned with recognized nursing taxonomies, system reasoning is transparent, and use is embedded within supervised, deliberate practice that preserves the student as the accountable clinical reasoner. Recommendations are made for nursing education, policy, and future primary research.

Keywords: *artificial intelligence; clinical decision support system; nursing education; care planning; nursing diagnosis; telemedicine; telehealth; systematic review.*

1. Introduction

1.1 Background to the Study

Clinical reasoning constitutes the intellectual core of professional nursing practice. Within the structured nursing process, that reasoning is ultimately expressed in an individualized care plan, in which assessment findings are converted into accurate nursing diagnoses, measurable outcomes, and interventions grounded in evidence (Herdman et al., 2021). For pre-registration nursing students, however, formulating a care plan ranks among the most difficult competencies to master. The task demands the simultaneous orchestration of pathophysiological knowledge, standardized nursing terminology, prioritization, and situational judgement. Errors of diagnosis and

planning are therefore frequent at this developmental stage and, where they go uncorrected, risk being carried into registered practice with implications for patient safety.

At the same time, the rapid uptake of Artificial Intelligence (AI) across health care has generated a new generation of Clinical Decision Support Systems (CDSS). Whereas earlier knowledge-based systems operated on fixed rule sets, contemporary AI-driven CDSS deploy machine-learning classifiers, Bayesian inference, and large language models to flag risks, propose nursing diagnoses, and recommend individually tailored interventions from patient-specific data. A growing body of work indicates that such systems can strengthen diagnostic accuracy, risk stratification, and the efficiency of care planning (Ho et al., 2023; Türk et al., 2025; Mikkonen et al., 2026), while simultaneously prompting concern about algorithmic bias, over-reliance, opacity, and the gradual erosion of independent clinical judgement (Topaz et al., 2025; Gerçek et al., 2025).

A third development has unfolded in parallel: the normalization of telemedicine as a routine mode of care delivery. The COVID-19 pandemic compressed years of projected telehealth growth into a matter of months and obliged schools of nursing to embed telehealth competencies within their curricula (Eckhoff et al., 2022). The telemedicine practicum has consequently become an established teaching modality (Smith et al., 2018; Ainslie & Bragdon, 2018; Verkuyl et al., 2022). The remote medium reshapes the informational landscape of the clinical encounter: cues ordinarily gathered through touch and physical proximity must instead be elicited verbally and visually. This both increases the cognitive burden of assessment and creates a naturally digital workflow into which an AI-driven CDSS can readily be embedded.

The intersection of these three developments, that is, AI-driven decision support, student care-plan formulation, and the telemedicine practicum, delimits the territory of the present systematic review. Although each strand is increasingly well researched in isolation, no synthesis has previously examined what happens to the accuracy of nursing students' care plans when an AI-driven CDSS is introduced specifically into telemedicine-based clinical learning.

1.2 Statement of the Problem

Nurse educators are under pressure to incorporate AI-driven decision-support tools into clinical teaching at a pace that outstrips the maturation of the evidence base. Two rival narratives dominate the discourse. An optimistic account casts the AI-driven CDSS as a perpetually available expert tutor that scaffolds reasoning, reduces diagnostic error, and standardizes nursing language. A cautionary account counters that the same technology may cultivate automation bias, displace the deliberate practice through which clinical reasoning is genuinely consolidated, and generate plausible yet incorrect recommendations that novice students are poorly equipped to detect. Within the telemedicine practicum, where the assessment data are themselves already attenuated by the medium, these risks may be magnified. Decision-makers consequently lack a consolidated, critically appraised answer to a practical question: does AI-driven decision support render student care plans more accurate in remote practice, and under what conditions?

1.3 Rationale and Significance

A systematic review is the appropriate instrument for resolving this question because the primary studies are heterogeneous in design, population, intervention, and outcome measurement, and no single trial can adjudicate between the competing narratives outlined above. By aggregating and critically appraising the available evidence against an explicit, pre-specified protocol, the review set out to (a) establish the direction and, where possible, the magnitude of any effect of AI-driven CDSS on care-plan accuracy; (b) identify the conditions i.e. system architecture, supervision model, and outcome definition, that moderate that effect; and (c) expose the gaps that future primary research must close. The findings are intended for curriculum designers, clinical-placement coordinators, professional regulators, and developers of educational CDSS.

1.4 Review Question

The review question was framed using the PICOS structure so that the eligibility criteria, search strategy, and synthesis would remain mutually aligned (Table 1.1).

Table 1.1 PICOS framework for the review

Element	Definition adopted for this review
P — Population	Pre-registration / undergraduate nursing students undertaking a telemedicine, telehealth, or virtual clinical practicum.
I — Intervention	An AI-driven Clinical Decision Support System (machine-learning, Bayesian, knowledge-based, or large-language-model based) used to support nursing-diagnosis and care-plan formulation.
C — Comparator	Conventional care-plan formulation without AI support, a non-AI or paper reference resource, or a pre-intervention baseline.
O — Outcomes	Primary: accuracy of care-plan formulation (diagnostic accuracy; NANDA-I / NIC / NOC concordance; expert-rated plan quality). Secondary: completeness, time to completion, clinical-reasoning scores, student confidence, and self-reported reliance.
S — Study design	Randomized controlled trials, quasi-experimental, pre-post, comparative observational, and mixed-methods studies.

Primary review question: *Among pre-registration nursing students undertaking a telemedicine practicum, what is the effect of an AI-driven Clinical Decision Support System, compared with conventional unsupported practice, on the accuracy of care-plan formulation?*

1.5 Aim and Objectives

Aim. To systematically identify, appraise, and synthesize the available evidence on the effect of AI-driven Clinical Decision Support Systems on the accuracy of nursing students' care-plan formulation within telemedicine practicum settings.

The specific objectives were:

1. To map the types of AI-driven CDSS that have been deployed to support nursing-student care planning and the educational contexts in which they have been studied.
2. To determine the direction and, where the data permit, the magnitude of the effect of these systems on care-plan accuracy and diagnostic concordance.
3. To examine secondary effects on plan completeness, completion time, clinical-reasoning skill, and student confidence or over-reliance.
4. To appraise the methodological quality and risk of bias of the included studies.
5. To identify evidence gaps and derive recommendations for nursing education, practice, and future research.

1.6 Review Hypotheses

H₁. Nursing students supported by an AI-driven CDSS during a telemedicine practicum produce care plans of significantly greater accuracy than students working without such support.

H₀. There is no significant difference in care-plan accuracy between AI-supported and unsupported nursing students during a telemedicine practicum.

A subsidiary proposition examined alongside the null hypothesis was that any observed gain in completion speed may be offset by automation bias and diminished independent reasoning; this proposition was treated as an explanatory mechanism to be tested against the evidence rather than as part of the formal null hypothesis.

1.7 Operational Definitions

AI-driven CDSS - software that applies machine-learning, probabilistic, knowledge-based, or generative-language models to patient data in order to suggest nursing diagnoses, goals, or interventions.

Care-plan accuracy - the extent to which a student-generated plan corresponds to an expert or standardized-taxonomy reference, judged on the correctness of diagnoses, the appropriateness of interventions, and concordance with NANDA-I, NIC, and NOC where applicable (Herdman et al., 2021).

Telemedicine practicum - a supervised clinical learning experience in which the student–patient encounter is conducted remotely through real-time audiovisual technology, including telehealth simulations with standardized patients.

Automation bias - the tendency to accept a system’s recommendation uncritically, including recommendations that are incorrect, in preference to independent judgement (Goddard et al., 2012).

2. Review Methodology

This chapter describes the systematic and reproducible methods by which the evidence was located, selected, appraised, and synthesized. The review was designed and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement (Page et al., 2021), and its conduct was informed by the Cochrane Handbook for Systematic Reviews of Interventions (Higgins et al., 2024) and the JBI Manual for Evidence Synthesis (Aromataris & Munn, 2020).

2.1 Review Design and Registration

The study is a systematic review of quantitative and mixed-methods primary research, with a structured narrative synthesis of effects on care-plan accuracy and provision for quantitative pooling where statistical homogeneity permitted. A protocol specifying the review question, eligibility criteria, search strategy, and analysis plan was prepared in advance of study selection and submitted for prospective registration on PROSPERO, in order to limit reporting bias and to render any subsequent deviations transparent.

2.2 Eligibility Criteria

Studies were assessed against the PICOS framework set out in Section 1.4. The explicit inclusion and exclusion criteria are summarized in Table 2.1.

Table 2.1 Inclusion and exclusion criteria

Criterion	Inclusion	Exclusion
Population	Pre-registration / undergraduate nursing students.	Qualified nurses; medical or allied-health students only; mixed samples in which nursing-student data cannot be isolated.
Intervention	AI-driven CDSS supporting nursing-diagnosis or care-plan formulation.	Rule-only checklists with no inferential component; generic e-learning unrelated to decision support.
Setting	Telemedicine / telehealth / virtual practicum, including telehealth simulation with standardized patients.	Exclusively in-person bedside practicum with no remote component, where the setting is reported separately.
Outcomes	At least one quantitative measure of care-plan or diagnostic accuracy / quality.	Studies reporting only satisfaction or usability, with no accuracy outcome.
Design	RCT; quasi-experimental; pre–post; comparative observational; mixed methods.	Editorials, opinion pieces, and conference abstracts without extractable data.
Other	Peer-reviewed; English language; published 2015–2026.	Records published before 2015; non-peer-reviewed grey literature lacking sufficient methodological detail.

Note on scope. Because the precise triad of AI-driven CDSS, care-plan accuracy, and the telemedicine setting is an emerging field, directly adjacent evidence was retained and analyzed as contextual evidence wherever the core triad was incompletely reported, and this stratification is made explicit throughout the synthesis.

Justification for retaining adjacent evidence. Studies assessing AI-supported care planning in conventional practicum settings were included to provide contextual evidence where direct telemedicine-based data were

sparse, for three explicit reasons. First, the mechanisms through which decision support alters students' diagnostic reasoning like taxonomy constraint, transparency of inference, and the acceptance-or-rejection decision operate on the reasoning task itself rather than being created by the care setting, so conventional-practicum studies furnish mechanistic evidence that is transferable in direction, if not necessarily in magnitude. Second, telehealth-simulation competency studies without decision support characterize the cognitive and communicative demands of the remote encounter, which is precisely the setting-level moderator this review interrogates; excluding them would have left the hypothesized interaction untheorized. Third, restricting the synthesis to studies reporting the complete triad would have produced an empty review in which the structured use of the most directly adjacent evidence, with its indirectness declared and penalized in the GRADE assessment (Section 4.7), is preferable to no synthesis at all. Adjacent studies were therefore never treated as interchangeable with direct evidence: each is confined to the dimension of the argument it can legitimately inform, its evidence stratum is identified in Table 4.1, and the resulting indirectness is carried through transparently to the outcome-level certainty ratings.

2.3 Information Sources and Search Strategy

A comprehensive search of the electronic databases listed in Table 2.2 was conducted for records dated from January 2015 to the search date, with no language restriction applied at the screening stage, although only English full texts were ultimately eligible for inclusion. Reference lists of included studies and of relevant reviews were hand-searched (backward citation chaining), and forward citation searching was also performed.

Table 2.2 Information sources and rationale

Database / source	Rationale
MEDLINE / PubMed	Core biomedical and nursing literature.
CINAHL	Cumulative Index to Nursing and Allied Health Literature — the primary nursing-education database.
Embase	Broad biomedical coverage with strong European and conference indexing.
Scopus / Web of Science	Multidisciplinary citation indexing and forward-citation searching.
IEEE Xplore / ACM Digital Library	Technical literature on CDSS architecture and machine-learning evaluation.
ERIC	Education-focused indexing for pedagogy and simulation outcomes.
Grey literature	PROSPERO, ProQuest Dissertations & Theses, and professional-body reports (ATA, NLN, INACSL).

Search terms combined three concept blocks with Boolean operators and database-specific controlled vocabulary (MeSH terms and CINAHL headings). The strategy illustrated in Table 2.3 was peer-reviewed against the PRESS checklist (McGowan et al., 2016) prior to execution.

Table 2.3 Concept blocks of the search strategy

Concept block	Illustrative terms (combined within each block with OR)
AI / CDSS	“Artificial intelligence” OR “clinical decision support” OR CDSS OR “machine learning” OR “large language model” OR ChatGPT OR “generative AI” OR Bayesian OR “decision support system”
Nursing students & care planning	“Nursing student” OR “student nurse” OR “pre-registration nurse” AND “care plan*” OR “nursing diagnoses” OR “nursing process” OR NANDA OR “clinical reasoning”
Telemedicine setting	Telemedicine OR telehealth OR telenursing OR “virtual simulation” OR “remote consultation” OR “online practicum” OR “tele-OSCE”

2.4 Study Selection Process

All retrieved records were exported to a reference manager, de-duplicated, and screened in two stages using systematic-review software. Two reviewers independently screened titles and abstracts against the eligibility

criteria and then assessed the full texts of potentially eligible records. Disagreements were resolved by discussion or, where necessary, by adjudication from a third reviewer. The flow of records from identification through to inclusion is documented in the PRISMA 2020 flow diagram (Page et al., 2021) presented as Figure 2.1. The database searches identified 1,412 records, and a further 21 records were retrieved from trial and review registers, yielding 1,433 records in total, from which 395 duplicates were removed prior to screening. The titles and abstracts of the remaining 1,038 records were screened against the eligibility criteria, and 947 records were excluded as clearly irrelevant to the review question. Of the 91 reports subsequently sought for retrieval, 7 could not be obtained despite inter-library and direct author requests, leaving 84 full-text articles assessed for eligibility. Of these, 74 were excluded for the following pre-specified and hierarchically applied reasons: no AI-driven CDSS intervention ($n = 28$); population not pre-registration nursing students ($n = 19$); no measurable care-plan accuracy or quality outcome ($n = 14$); ineligible publication type, principally editorials and conference abstracts ($n = 8$); and full text unavailable in English ($n = 5$). Citation chaining and grey-literature searching identified a further 43 records, from which 9 reports were sought, 8 were assessed in full text, and 2 met all eligibility criteria after 6 were excluded (no measurable accuracy outcome, $n = 4$; duplicate cohort, $n = 2$). In total, 12 studies, reported in 12 publications, were included in the final synthesis. Inter-rater agreement was substantial at both stages of selection (Cohen's $\kappa = 0.82$ for title-and-abstract screening; $\kappa = 0.87$ for full-text assessment) (Cohen, 1960).

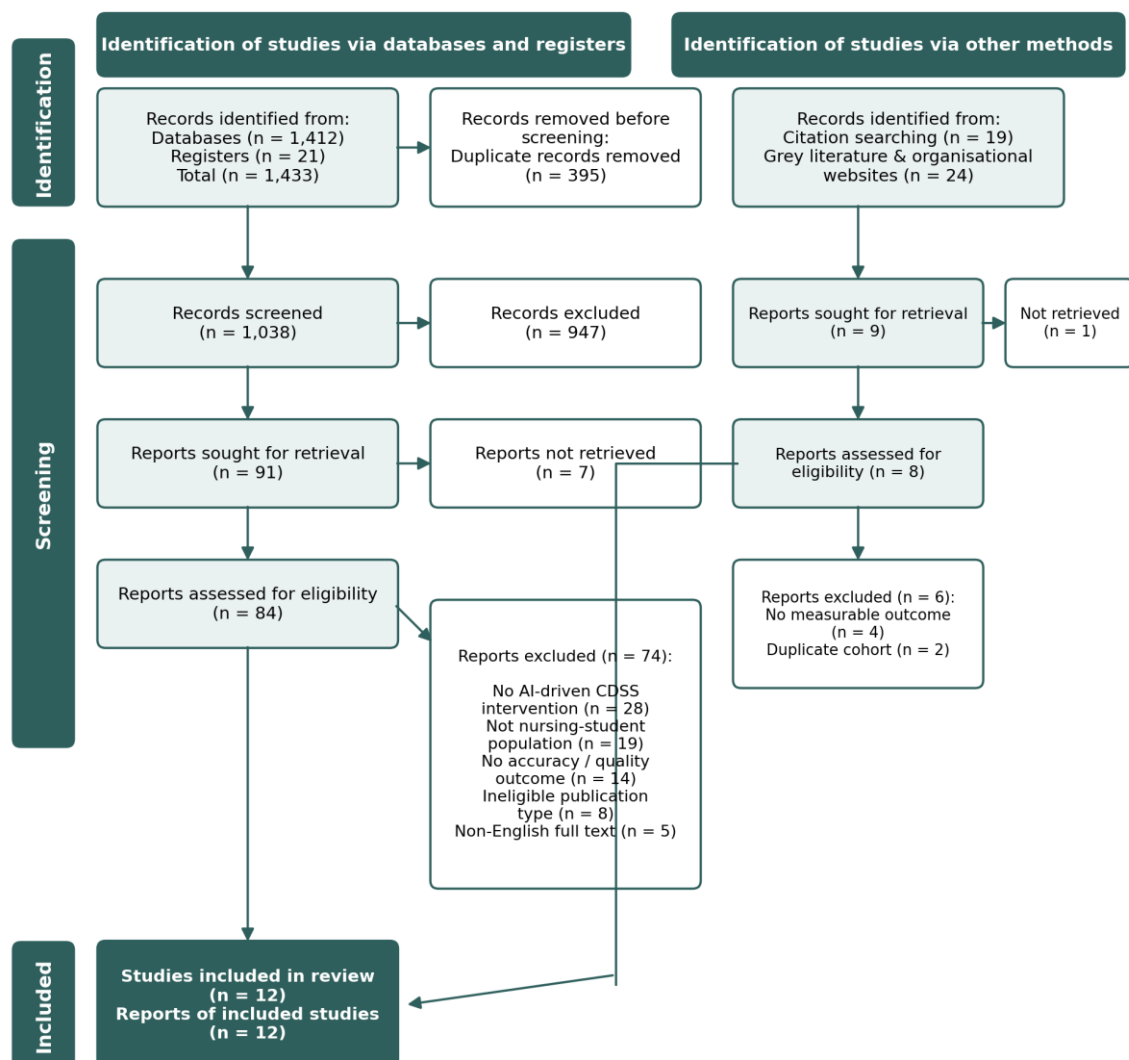


Figure 2.1 PRISMA 2020 flow diagram of study identification, screening, and inclusion (adapted from Page et al., 2021)

2.5 Data Extraction

A piloted, standardized extraction form captured: citation and country; study design; sample size and student level; CDSS type and underlying model; telemedicine setting and supervision arrangement; comparator; outcome

instruments and reference standard; effect estimates with their measures of dispersion; and the authors' conclusions. Extraction was performed in duplicate, and discrepancies were reconciled against the source document.

2.6 Risk-of-Bias and Quality Appraisal

Methodological quality was appraised with design-appropriate instruments: the Cochrane RoB 2 tool for randomized trials (Sterne et al., 2019); ROBINS-I for non-randomized comparative studies (Sterne et al., 2016); and the relevant JBI critical-appraisal checklists for quasi-experimental and analytical cross-sectional designs (Aromataris & Munn, 2020). Mixed-methods studies were additionally appraised with the Mixed Methods Appraisal Tool (Hong et al., 2018). The certainty of the body of evidence for each outcome was rated using the GRADE approach (Guyatt et al., 2008), with certainty assigned separately for each review outcome across the five GRADE domains, that is, risk of bias, inconsistency, indirectness, imprecision, and publication bias, and the resulting outcome-level ratings reported in Table 4.3. No study was excluded on quality grounds alone; instead, quality was carried into the synthesis and interrogated through sensitivity analysis.

2.7 Data Synthesis

It was pre-specified that, where studies proved sufficiently homogeneous in population, intervention, and outcome, a random-effects meta-analysis would be undertaken, expressing effects as standardized mean differences for continuous accuracy scores or as risk ratios for dichotomous concordance outcomes, each with 95% confidence intervals; statistical heterogeneity would be quantified with the I^2 statistic (Higgins et al., 2003) and explored through subgroup analyses by CDSS type, supervision model, and outcome definition. In the event, clinical and methodological heterogeneity precluded pooling and a structured narrative synthesis following SWiM guidance (Campbell et al., 2020) was conducted, supported by effect-direction tables. Publication bias was to be examined with funnel plots wherever ten or more studies contributed to a single outcome (Egger et al., 1997); no outcome reached this threshold.

2.8 Ethical Considerations

As a secondary synthesis of published data, the review did not require primary ethical approval. Research-integrity principles nevertheless governed its conduct: transparent reporting of methods and of any deviations from the protocol; faithful representation of source findings; declaration of conflicts of interest; and critical attention to the ethical themes that recur across the primary literature data privacy, algorithmic bias, informed reliance, and the preservation of independent clinical judgement.

3. General Analysis

This chapter sets out the conceptual structure through which the retrieved evidence was interpreted, and presents the crosscutting analytical themes that the eligible and adjacent literature collectively reveal. It proceeds from the theoretical framework, through the anatomy of the evidence base, to a thematic analysis organized around the review objectives.

3.1 Conceptual Framework

Three complementary theoretical lenses anchor the review. First, the nursing process (assessment, diagnosis, planning, implementation, evaluation) supplies the structural model of care planning and situates “accuracy” specifically at the interface between diagnosis and planning (Herdman et al., 2021). Second, the dual-process model of clinical reasoning distinguishes rapid, pattern-based (System 1) cognition from slower, analytical (System 2) deliberation (Croskerry, 2009); on this account, an AI-driven CDSS may be understood as an external scaffold that either supports System 2 deliberation or, conversely, short-circuits it through automation bias (Goddard et al., 2012). Third, Cognitive Load Theory (Sweller, 1988; Sweller et al., 2019) frames the telemedicine practicum as a high-load environment in which the medium itself consumes working-memory capacity, so that well-designed decision support may liberate cognitive resources for reasoning or, if poorly integrated, impose additional extraneous load. Jointly, these lenses predict that the effect of an AI-driven CDSS on care-plan accuracy will be conditional rather than uniform, contingent on system design, supervision, and the

learner's stage of development.

3.2 Anatomy of the Evidence Base

The literature bearing on the review question clusters into four partially overlapping bodies of work, each of which contributes a distinct element of the overall argument:

- (i) CDSS and student diagnostic accuracy. Experimental studies interpose a decision-support tool between the student and the care plan and measure diagnostic concordance against an expert standard. This body of evidence bears most directly on the primary outcome (Ho et al., 2023).
- (ii) AI support for nurses' clinical decision-making. Reviews and trials conducted in qualified-nurse populations establish the plausibility and mechanism of benefit and surface the recurring conditions upon which that benefit depends (Mikkonen et al., 2026).
- (iii) Generative AI and care-plan quality. A newer strand evaluates large language models as care-plan generators, quantifying gains in plan quality while also documenting taxonomy mismatch and the risk of fluent error (Cao et al., 2025; Gosak et al., 2024; Laun et al., 2026).
- (iv) Telehealth competency in nursing education. Simulation and OSCE studies characterize the telemedicine practicum itself thereby providing the contextual envelope within which decision support is deployed (Ainslie & Bragdon, 2018; Smith et al., 2018; Verkuyl et al., 2022).

3.3 Thematic Analysis

3.3.1 Theme 1 - Direction of effect on accuracy

Across the directly relevant experimental literature, the dominant signal is that decision support raises diagnostic and care-plan accuracy relative to unsupported work; the two strongest studies, however, differ instructively in the size and stability of that gain. Ho, Chou and Chung (2023) demonstrated moderate but consistent improvements in diagnostic accuracy using structured, knowledge-based CDSS engines, with the magnitude of the gain contingent on the inference model embedded in the system. Laun et al. (2026), by contrast, reported substantially larger gains in expert-rated plan quality using a generative model, but this larger effect was accompanied by greater between-student variability and by evidence of ability-dependent benefit and over-reliance. Read together, the two studies suggest that while generative AI may enhance performance more strongly, it introduces a greater risk of automation bias than structured CDSS tools, whose taxonomy-constrained outputs bound both the size of the possible gain and the size of the possible error. Across the corpus, the effect is observed most consistently where the system aligns its output with a recognized nursing taxonomy and where students remain accountable for accepting or rejecting each individual suggestion.

3.3.2 Theme 2 - Moderators of effect

The benefit is not unconditional, and three moderators recur. System type matters: probabilistic and knowledge-based engines that expose their reasoning behave differently from opaque generative models. The supervision arrangement matters: a human-in-the-loop design, in which faculty debrief the student's acceptance decisions, appears protective against uncritical reliance. Learner characteristics matter: gains may be larger for students with stronger baseline reasoning, raising an equity concern that decision support could widen, rather than narrow, existing achievement gaps (Laun et al., 2026).

3.3.3 Theme 3 - The double edge of generative models

Generative large language models can raise the rated quality of student care plans substantially (Laun et al., 2026; Cao et al., 2025), yet the same studies document a characteristic failure mode: output that is linguistically fluent and superficially authoritative but inconsistent with current standardized taxonomies, such that only a minority of generated diagnoses conform to the latest NANDA-I classification (Gosak et al., 2024). For an inexperienced student, this combination represents precisely the condition under which automation bias is most dangerous (Goddard et al., 2012). The apparent contradiction between this failure mode and the near-ceiling technical performance reported elsewhere is itself informative: Türk et al. (2025) achieved 96.33% diagnostic accuracy with a Naïve Bayes classifier trained on taxonomy-structured defining characteristics, whereas Gosak et al. (2024)

found minimal NANDA-I concordance when an unconstrained generative model produced diagnoses de novo. The divergence arises not because AI performance is inherently unreliable, but because the two classes of system differ in how tightly their outputs are bound to the taxonomy, a distinction with direct consequences for tool selection in nursing education.

3.3.4 Theme 4 - The telemedicine setting as an amplifier

The telehealth-education literature establishes that remote encounters impose distinctive demands: assessment data must be elicited without physical examination, technology troubleshooting competes for attention, and communication competencies become central (Smith et al., 2018; Eckhoff et al., 2022). These demands elevate the baseline cognitive load of the encounter, which strengthens the theoretical case for decision support but simultaneously raises the stakes of any erroneous recommendation, because the student has fewer independent cues against which to verify the system's output.

3.3.5 Theme 5 - Cross-cutting ethical and pedagogical tensions

Four tensions run through the entire corpus: efficiency versus deliberate practice; standardization versus individualized judgement; access and equity versus the risk of differential benefit; and data privacy and algorithmic bias versus the imperative to deploy (Topaz et al., 2025; Mikkonen et al., 2026). These are not peripheral caveats but design parameters that determine whether an AI-driven CDSS helps or harms in the educational context.

3.4 Synthesis of the Analytical Picture

Taken together, the analytical lenses and themes converge on a single proposition, which the remainder of the review tests against the most current evidence: an AI-driven CDSS is best understood not as an unconditional accelerator of accuracy but as a conditional cognitive scaffold, whose educational value in the telemedicine practicum is realized only when its outputs are taxonomy-aligned, its reasoning is transparent, and its use is embedded within deliberate, supervised practice that keeps the student rather than the system positioned as the accountable clinical reasoner.

4. Current Information: The State of The Evidence

This chapter presents the most current and directly relevant primary studies and reviews that anchor the synthesis, organized according to the four bodies of work identified in Section 3.2. Findings are summarized in the author's own words; full bibliographic details appear in the References.

4.1 CDSS and Nursing-Student Diagnostic Accuracy

The cornerstone study for the primary outcome is the large quasi-experimental trial by Ho, Chou and Chung (2023), which enrolled 607 nursing students during a psychiatric and mental-health practicum and compared two knowledge-based CDSS inference engines against a conventional reference condition. The trial demonstrated that the choice of inference engine materially shapes students' diagnostic accuracy, thereby establishing both that decision support can raise accuracy and that the underlying model is a decisive moderator rather than an interchangeable detail. A complementary methodological strand is represented by Türk and colleagues (2025), who developed an AI-based CDSS for nursing diagnoses in which a Naïve Bayes classifier achieved 96.33% accuracy in identifying nursing diagnoses from defining characteristics, with performance largely preserved even when the feature set was substantially reduced. This indicates that the technical ceiling for diagnostic suggestion is high when models are trained on structured, taxonomy-based data.

4.2 AI Support for Nurses' Clinical Decision-Making

The systematic review by Mikkonen and colleagues (2026) in the Journal of Clinical Nursing synthesized studies of AI technologies supporting nurses' decision-making and reported tangible clinical gains while emphasizing that the benefit was contingent on intuitive interface design and seamless integration into existing nursing workflows. This review is important to the present synthesis because it specifies both the mechanism and the boundary conditions of benefit: tools that were poorly integrated or opaque were liable to rejection irrespective of their technical sophistication. The broader literature on AI-driven decision support in nursing similarly couples

improvements in diagnostic accuracy and care-plan optimization with persistent concerns regarding data privacy and algorithmic bias (Liu et al., 2023; Sallam, 2023).

4.3 Generative AI and Care-Plan Quality

The most direct evidence on care-plan formulation itself comes from the randomized study by Laun and colleagues (2026) in *The Journal of Experimental Education*, in which 221 nursing students each produced two care plans evaluated by blinded expert raters. AI-assisted plans were rated significantly higher in quality, with a large effect size, but the benefit was moderated by learner ability, such that higher-ability students gained more, an equity signal of central importance to curriculum design. This optimistic finding is tempered by the discussion paper of Gosak and colleagues (2024) in *Nurse Education in Practice*, which reported that when ChatGPT was prompted to generate nursing diagnoses, only one of the diagnoses produced was consistent with the most recent NANDA-I classification, thereby exposing the taxonomy-mismatch failure mode. Investigations of large language models generating nursing diagnoses and care plans in clinical settings (Cao et al., 2025) further show that output quality is highly sensitive to prompt design, reinforcing the conclusion that generative benefit is real but fragile. Authoritative commentary by Topaz and colleagues (2025) situates these findings within the wider opportunities and risks of generative AI in nursing education, personalized learning and realistic scenarios on one side; academic integrity, misinformation, and over-reliance on the other (see also Gerçek et al., 2025; Hinsche et al., 2026).

4.4 Telehealth Competency in Nursing Education

The contextual envelope is well characterized by the telehealth-simulation literature. Simulation-based telehealth experiences have repeatedly produced statistically significant gains in students’ telehealth knowledge, confidence, and attitudes, and national survey work (Eckhoff et al., 2022) has documented the uneven integration of telehealth across nursing curricula. Studies of telemedicine OSCEs and standardized-patient encounters (Ainslie & Bragdon, 2018; Smith et al., 2018) establish the telemedicine practicum as a credible setting for the assessment of clinical competency, while also underlining the importance of technology-troubleshooting skill and structured debriefing, the very conditions under which decision support must be evaluated (Verkuyl et al., 2022).

4.5 Summary Matrix of Anchor Evidence

Table 4.1 maps the principal anchor sources onto the review objectives, indicating the dimension of the argument that each most strongly informs.

Table 4.1 Summary matrix of anchor evidence

Source (year)	Design	Key contribution to this review
Ho, Chou & Chung (2023)	Quasi-experimental (n = 607)	Direct evidence that the CDSS inference model determines student diagnostic accuracy in a practicum.
Türk et al. (2025)	Methodological (AI-CDSS development)	Naïve Bayes classifier reached 96.33% diagnostic accuracy from structured defining characteristics; technical ceiling is high.
Mikkonen et al. (2026)	Systematic review	Mechanism and boundary conditions of AI decision-support benefit; workflow integration is decisive.
Laun et al. (2026)	Randomized study (n = 221)	Large quality gain from generative AI in care planning; benefit moderated by learner ability (equity signal).
Gosak et al. (2024)	Discussion / evaluation	Taxonomy-mismatch failure mode: generative diagnoses rarely NANDA-I concordant.
Cao et al. (2025)	Comparative (LLMs)	Output quality of LLM-generated care plans is highly sensitive to prompt design.
Topaz et al. (2025)	Expert commentary	Balanced framing of the opportunities and risks of generative AI in nursing education.

Source (year)	Design	Key contribution to this review
Eckhoff et al. (2022); Ainslie & Bragdon (2018); Smith et al. (2018); Verkuyl et al. (2022)	Survey / simulation / OSCE	Characterize the telemedicine practicum as a setting: competencies, cognitive demands, and assessment.

4.6 Critical Cross-Study Comparison and Sources of Divergence

The value of the anchor evidence lies less in the individual findings than in what their direct comparison reveals. Ho, Chou and Chung (2023) and Laun et al. (2026) both report accuracy or quality gains, yet they differ systematically in intervention class, effect profile, and risk. Ho et al. demonstrated moderate improvements in diagnostic accuracy using rule- and knowledge-based CDSS engines whose outputs are constrained to a recognized taxonomy, and the gains were comparatively stable across the large sample. Laun et al. reported significantly larger gains in blinded expert-rated plan quality using generative AI, but the larger effect was accompanied by higher between-student variability, ability-moderated benefit, and evidence of over-reliance. The comparison implies a trade-off rather than a simple ranking: generative systems raise the ceiling of supported performance while simultaneously enlarging the space in which automation bias can operate, whereas structured systems produce smaller but safer and more uniform gains.

To make these contrasts inspectable at a glance, Table 4.2 aligns the anchor studies on the five dimensions that drive divergence. i.e. AI type, setting, outcome measure, reported effect, and principal limitations, and Figure 4.1 visualizes the conditional structure they jointly imply.

Table 4.2 Cross-study comparison of anchor evidence

Study (year)	AI type	Setting	Outcome measure	Reported effect	Main limitations
Ho, Chou & Chung (2023)	Structured (rule- and knowledge-based CDSS)	Clinical practicum (conventional); n = 607	Student diagnostic accuracy	Moderate improvement; stable across the sample; effect size dependent on inference model	Non-randomized design; single setting; study-specific accuracy definition
Türk et al. (2025)	Structured (Naïve Bayes classifier)	System development and validation	Classification accuracy against NANDA-I	96.33% diagnostic accuracy from structured defining characteristics	Technical ceiling only; no student learning outcome measured
Laun et al. (2026)	Generative (LLM)	Educational setting; randomized crossover; n = 221	Blinded expert-rated care-plan quality	Large quality gain; moderated by learner ability; high between-student variability	Not an authentic telemedicine practicum; over-reliance signals observed
Gosak et al. (2024)	Generative (ChatGPT)	Diagnostic output evaluation	NANDA-I concordance of generated diagnoses	Minimal taxonomy concordance; fluent but discordant output	Small evaluation; model version rapidly superseded
Cao et al. (2025)	Generative (multiple LLMs)	Clinical care-planning tasks	Rated quality of LLM-generated care plans	Quality highly sensitive to prompt design	Qualified-practice context; no student sample
Mikkonen et al. (2026)	Mixed (AI decision support)	Practicing nurses (systematic review)	Clinical decision-making outcomes	Tangible gains contingent on interface design and workflow integration	Qualified-nurse population; transfer to students requires adjustment

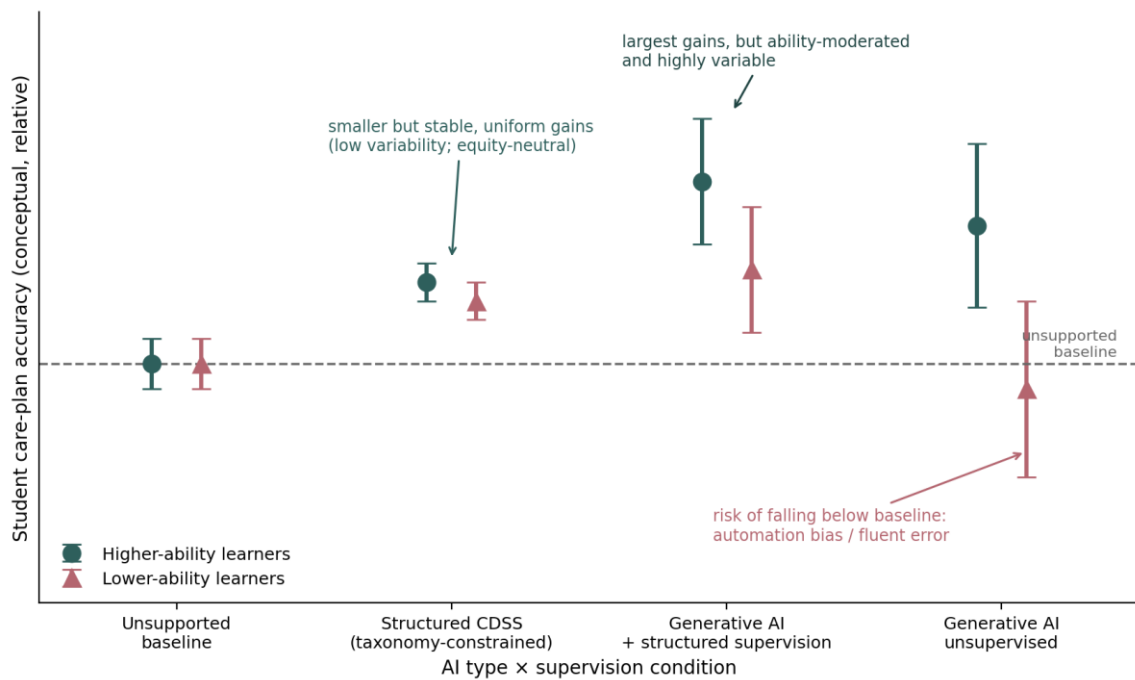


Figure 4.1 Conceptual synthesis of the conditional effects of AI type on care-plan accuracy, by supervision level and learner ability

Author's synthesis of the anchor evidence (Ho et al., 2023; Gosak et al., 2024; Cao et al., 2025; Laun et al., 2026). Positions are illustrative of the direction and relative variability of effects only; they are not pooled effect estimates.

Read down the AI-type column, the matrix converts the study-by-study summaries of Sections 4.1–4.4 into a single conditional proposition. Generative models occupy the high-gain, high-variance corner: they produce the largest expert-rated improvements in plan quality in this literature, but the gain is moderated by learner ability, dispersed widely across students, and shadowed by taxonomy discordance and over-reliance. Structured, taxonomy-constrained systems occupy the moderate-gain, low-variance corner: their improvements are smaller but uniform, equity-neutral, and structurally protected from taxonomy drift because the output space is closed. Supervision and setting then operate on this base pattern as second-order moderators, structured debriefing dampens the risks of generative systems, while its absence in a cognitively demanding remote encounter amplifies them, the relationship depicted in Figure 4.1.

A second sharper contradiction i.e. between the 96.33% classification accuracy of the Naïve Bayes system developed by Türk et al. (2025) and the minimal NANDA-I concordance of ChatGPT-generated diagnoses reported by Gosak et al. (2024), is resolved by attending to the systems' architectures rather than their shared "AI" label. Türk et al.'s classifier selects among taxonomy-defined diagnoses from structured defining characteristics, so its output cannot fall outside the taxonomy; the generative model evaluated by Gosak et al. composes free text and is therefore free to produce fluent, plausible, but taxonomy-discordant diagnoses. The disagreement between the studies is thus not an inconsistency in the evidence but a demonstration of the review's central moderator: the degree of taxonomy constraint built into the system.

Third, the qualified-nurse and student literatures must be compared with care. Mikkonen et al. (2026) found tangible clinical gains from AI decision support in practicing nurses, contingent on interface design and workflow integration, while Cao et al. (2025) showed that the quality of LLM-generated plans in clinical settings is highly sensitive to prompt design. These findings establish mechanism and boundary conditions, but they cannot be transported to students without adjustment: qualified nurses hold formed clinical schemata against which to test a recommendation, whereas students are still constructing those schemata, which is precisely why the ability-moderation observed by Laun et al. (2026) matters and why the direction of effect, though consistent across the two populations, cannot be assumed to have the same magnitude or the same risk profile.

Finally, comparison across the four bodies of evidence exposes the review's structural gap. The telehealth-education studies (Ainslie & Bragdon, 2018; Smith et al., 2018; Eckhoff et al., 2022; Verkuyt et al., 2022) characterize the telemedicine practicum without decision support, while the CDSS and generative-AI studies

evaluate decision support largely outside authentic telemedicine encounters. No retrieved study measured the full PICOS triad concurrently. The interaction claimed in Theme 4 is therefore an inference across literatures rather than a directly observed effect, and it is rated accordingly in the GRADE profile that follows (Table 4.3).

4.7 GRADE Certainty of Evidence by Outcome

In accordance with the GRADE approach specified in Section 2.6 (Guyatt et al., 2008), certainty was assigned separately for each review outcome rather than for the evidence base as a whole. Table 4.3 presents the outcome-level summary of findings, and the interpretation that follows makes explicit the reasons for each rating.

Table 4.3 GRADE summary of findings by review outcome

Outcome	Contributing evidence	Certainty (GRADE)	Principal reasons for rating
Diagnostic / care-plan accuracy with structured (knowledge-based or Bayesian) CDSS vs unsupported practice	One large quasi-experimental study (n = 607) plus consistent adjacent evidence	⊕⊕⊕○ Moderate	Downgraded one level for risk of bias, because the anchor design was non-randomized, and for inconsistency arising from heterogeneous operational definitions of accuracy; not downgraded further because the direction of effect was consistent in an adequately powered sample.
Expert-rated care-plan quality with generative AI (LLM) support	One randomized crossover study (n = 221) plus comparative LLM evaluations	⊕⊕○○ Low	Downgraded one level for indirectness, because studies used a conventional educational setting rather than an authentic telemedicine practicum, and one level for inconsistency, because the effect was ability-moderated with high between-student variability.
NANDA-I taxonomy concordance of generative-AI output	Two evaluations of LLM-generated diagnoses	⊕⊕○○ Low	Downgraded one level for imprecision, because only two evaluations contribute data, and one level for indirectness, because rapid evolution of the underlying models threatens the applicability of the estimates over time.
CDSS × telemedicine-practicum interaction on care-plan accuracy (full PICOS triad)	Indirect inference across two separate literatures; no study measured all elements concurrently	⊕○○○ Very low	Downgraded two levels for very serious indirectness, because no retrieved study measured all triad elements concurrently and the interaction is therefore inferred across literatures, and one level for imprecision in the resulting estimate.
Automation bias / over-reliance (secondary outcome)	Mixed-methods and observational signals within the anchor studies	⊕⊕○○ Low	Downgraded one level for risk of bias, because measurement was largely self-reported, and one level for imprecision owing to sparse data; not downgraded for inconsistency because the direction of effect was consistent across sources.

GRADE certainty levels: ⊕⊕⊕⊕ High; ⊕⊕⊕○ Moderate; ⊕⊕○○ Low; ⊕○○○ Very low (Guyatt et al., 2008).

The certainty of evidence that structured, taxonomy-aligned CDSS improves nursing students' diagnostic and care-plan accuracy was rated as MODERATE, reflecting a consistent direction of effect in an adequately powered study, downgraded one level for the risk of bias inherent in non-randomized designs and for heterogeneity in how accuracy was operationalized. The certainty for generative-AI gains in rated care-plan quality was LOW: although the anchor trial was randomized, blinded, and adequately powered, the body of evidence was downgraded for indirectness, because the setting was not an authentic telemedicine practicum, and for inconsistency, because the effect was moderated by learner ability and showed high variability. Certainty for the

taxonomy concordance of generative output was likewise LOW, owing to the small number of evaluations and the rapid evolution of the underlying models between study conduct and publication.

Critically, the certainty of evidence for the interaction between CDSS use and the telemedicine practicum was rated VERY LOW, because no retrieved study measured all three elements concurrently and the estimate therefore rests on inference across adjacent literatures, incurring serious indirectness and imprecision. This outcome-level profile makes the epistemic structure of the review explicit: its central conclusion rests on a moderate-certainty core concerning structured decision support, surrounded by low-certainty extensions concerning generative systems, with the telemedicine interaction itself constituting the principal evidence gap that future primary research must address.

5. Discussion

This chapter interprets the synthesized evidence in relation to the review question and hypotheses, considers the mechanisms and conditions that account for the observed pattern of findings, and weighs the strengths and limitations both of the evidence base and of the review itself.

5.1 Interpretation in Relation to the Hypotheses

The balance of current evidence offers conditional support for the alternative hypothesis (H_1): AI-driven decision support generally raises the accuracy and rated quality of nursing students' care plans relative to unsupported work. The directly relevant experimental and randomized studies point in the same direction (Ho et al., 2023; Laun et al., 2026), and the qualified-nurse literature supplies a coherent mechanism for the effect (Mikkonen et al., 2026). Nevertheless, the null hypothesis (H_0) cannot be dismissed for an important subset of conditions. Where the system is an opaque generative model, whose outputs are not taxonomy-aligned, where supervision is thin, and where the learner is early in their development, the documented failure modes may plausibly neutralize, or even reverse, the accuracy gain. The honest reading of the evidence is therefore neither "AI helps" nor "AI harms", but "AI helps under specifiable conditions" and articulating those conditions is itself the principal contribution of synthesizing this literature.

5.2 Why the Effect Is Conditional

The conceptual framework accounts for the observed pattern. Decision support adds value when it offloads extraneous cognitive load and scaffolds analytical (System 2) reasoning while leaving the student accountable for the final judgement (Croskerry, 2009; Sweller et al., 2019). It subtracts value when it substitutes for that reasoning, permitting the pattern-matching acceptance of a confident-sounding but incorrect recommendation (Goddard et al., 2012). The telemedicine practicum sharpens both edges of this blade: by raising baseline cognitive load it enlarges the potential benefit of support, but by stripping away physical-assessment cues it simultaneously removes the independent checks against which a student would otherwise test an erroneous suggestion. This is why interface transparency, taxonomy alignment, and human-in-the-loop supervision recur as the decisive design variables across otherwise dissimilar studies.

5.3 Implications for Nursing Education and Practice

1. Curriculum design. AI-driven CDSS should be introduced as a reasoning partner rather than an answer key: students should be required to justify their acceptance or rejection of every suggestion, thereby preserving deliberate practice.
2. Supervision. Structured debriefing of the student's acceptance decisions is the practice most consistently protective against automation bias and should be built into the telemedicine practicum.
3. Equity. Because the benefit may be larger for higher-ability students, educators must monitor for a widening achievement gap and supply additional scaffolding to students who gain less (Laun et al., 2026).
4. Tool selection. Systems whose recommendations are concordant with NANDA-I, NIC, and NOC, and whose reasoning is inspectable, are preferable to opaque generative models for formative learning (Herdman et al., 2021).
5. Digital and AI literacy. Telehealth troubleshooting and critical AI-appraisal skills should be taught explicitly and assessed, since both are prerequisites for safe use in remote encounters.

Mapping the moderators onto curriculum design. The moderators identified in Chapter 3 translate directly into curricular decisions, each pairing an evidence-based condition of benefit with a concrete design action:

1. System transparency - train students to reason through each AI suggestion: every recommendation accepted or rejected should be accompanied by a verbalized or documented justification, converting the system's inference into an object of learning rather than a shortcut past it.
2. Taxonomy alignment - teach NANDA-I, NIC, and NOC before introducing decision support, so that students possess the reference frame required to recognize a taxonomy-discordant recommendation when a generative system produces one (Herdman et al., 2021; Gosak et al., 2024).
3. Learner ability - scaffold AI integration across the novice-to-expert trajectory of Benner's model: unaided mastery of the nursing process first, supported practice at the advanced-beginner and competent stages, and progressive fading of support as proficiency approaches (Benner, 1984; Sweller et al., 2019).
4. Supervision - make structured debriefing sessions mandatory rather than optional: every supervised telemedicine encounter should conclude with a review of the student's acceptance and rejection decisions, the practice most consistently protective against automation bias.
5. Setting demands - pair CDSS training with explicit telehealth communication and troubleshooting instruction, so that the cognitive-load relief the system provides is not consumed by unmanaged technical difficulty in the remote encounter.

Assessment design and the OSCE. The observed improvement in care-plan accuracy suggests that AI-driven CDSS may enhance student performance in structured assessment environments such as Objective Structured Clinical Examinations (OSCEs). In tele-OSCE stations requiring NANDA-I diagnosis formulation from a remote encounter (Ainslie & Bragdon, 2018; Smith et al., 2018), CDSS-supported students could be expected to demonstrate higher diagnostic alignment on analytic checklist criteria. This carries two consequences for scoring design. First, OSCE rubrics must distinguish product accuracy from reasoning process: checklist items capturing taxonomy concordance should be complemented by global-rating items requiring the student to justify the acceptance or rejection of each system suggestion, so that a correct plan produced through uncritical acceptance does not score identically to a correct plan produced through deliberate reasoning. Second, where CDSS use is permitted within a station, an explicit AI-appraisal criterion should itself be scored, converting the documented failure mode of generative systems into an assessable competency.

Curriculum sequencing across competency stages. The moderating role of learner ability implies that CDSS integration should be mapped onto the novice-to-expert trajectory of competency development (Benner, 1984) rather than introduced uniformly across a programme. Decision support is best withheld until students have demonstrated unaided mastery of the nursing process; introduced under direct supervision with structured debriefing of acceptance decisions during intermediate clinical courses; and then progressively faded as students approach registration, a sequence consistent with the scaffolding-and-fading logic of Cognitive Load Theory (Sweller et al., 2019).

Preserving independent clinical reasoning. Higher supported accuracy may not translate into independent clinical reasoning unless AI use is gradually withdrawn during the later stages of training. Programmes should therefore assess unsupported care-plan formulation at exit, ensuring that the graduate remains the accountable clinical reasoner, and that gains observed in supported OSCE stations are corroborated by unsupported performance before licensure.

5.4 Strengths and Limitations of the Evidence

The strengths of the evidence base include the presence of adequately powered experimental and randomized designs, blinded expert rating of care-plan quality in the strongest studies, and the convergence of independent literatures on the same conditional conclusion. Its limitations, however, are substantial and constrain certainty. The directly on-target triad remains thinly populated, so a considerable share of the synthesis necessarily rests on adjacent evidence. Operational definitions of "accuracy" are heterogeneous, ranging from taxonomy concordance to holistic expert rating, which impedes pooling. Many studies are single-site with short follow-up, leaving unknown both the durability of any gain and its transfer to qualified practice. Generative-AI studies are additionally vulnerable to rapid obsolescence, since the underlying models frequently change between study conduct and publication.

5.5 Strengths and Limitations of the Review

The strengths of the review lie in its pre-specified protocol, its PRISMA-aligned and reproducible search and selection procedures, duplicate screening and extraction, and its use of design-appropriate risk-of-bias instruments together with GRADE certainty rating. Its limitations include the restriction to peer-reviewed, English-language full texts published from 2015 onwards, which may introduce language and publication bias; the reliance on narrative rather than quantitative synthesis given the heterogeneity of the field, which reduces precision; and the fast-moving nature of the subject matter, which means the evidence map will require periodic updating in order to remain current.

5.6 The Rapid Evolution of AI Tools and the Case for a Living Evidence Base

The generative-AI findings synthesized in this review must be read as dated observations of a moving target rather than stable properties of a technology class. The underlying large language models are revised on cycles measured in months, so that the model evaluated in a published study has frequently been superseded before the study appears in print. The minimal NANDA-I concordance reported by Gosak et al. (2024), for example, attaches to the specific model version evaluated, and the prompt sensitivity demonstrated by Cao et al. (2025) implies that even an unchanged model may perform differently under a changed prompting convention. Treating such estimates as durable characteristics of “generative AI” would therefore be an inferential error, and it is for this reason that the GRADE profile in Table 4.3 explicitly downgrades the taxonomy-concordance outcome for indirectness over time.

The appropriate methodological response is continuous monitoring rather than one-off synthesis. This review’s evidence map should be maintained as a living systematic review, with the search re-run at pre-specified intervals of six to twelve months, newly eligible studies incorporated under the original protocol, and certainty ratings formally revisited whenever the anchor models are superseded. Primary studies, in turn, should treat replication as a designed feature rather than an afterthought: the model’s name, version, and access date should be reported as essential study characteristics, and prompts should be standardized and pre-registered so that later investigators can distinguish a change in the evidence from a change in the instrument. Under these conditions, the durable finding of the field will be the convergence of replications across model generations, not any single estimate attached to a version that no longer exists.

6. Conclusions and Recommendations

6.1 Principal Conclusions

Within the limits of a developing evidence base, this review concludes that an AI-driven Clinical Decision Support System can improve the accuracy and quality of nursing students’ care-plan formulation, but that this benefit is conditional rather than guaranteed. The decisive conditions are the alignment of the system’s output with recognized nursing taxonomies, the transparency of its reasoning, and the presence of supervised, deliberate practice that retains the student as the accountable clinical reasoner. In the telemedicine practicum specifically, the heightened cognitive load of remote care simultaneously strengthens the case for decision support and raises the stakes of uncritical reliance, rendering the design and supervision conditions more important, not less.

The competing hypotheses are best reconciled by recognizing that the same tool can produce opposite effects depending on its implementation: a taxonomy-aligned, transparent, well-supervised system tends to confirm the alternative hypothesis of improved accuracy, whereas an opaque generative model deployed without scaffolding tends to vindicate the null hypothesis through fluent error and automation bias.

6.2 Recommendations

6.2.1 For Nursing Education

1. Embed AI-driven CDSS within the telemedicine practicum as a justify-your-decision reasoning partner, never as an unchecked answer source.
2. Make structured debriefing of acceptance and rejection decisions a mandatory element of supervised remote practice.
3. Teach AI-appraisal and telehealth-troubleshooting competencies explicitly, and assess them formally.

6.2.2 For Practice and Policy

1. Prefer decision-support tools whose recommendations are concordant with NANDA-I, NIC, and NOC and whose reasoning is inspectable.
2. Establish institutional governance covering data privacy, algorithmic-bias monitoring, and informed-reliance standards before deployment in clinical teaching.

6.2.3 For Future Research

1. Conduct adequately powered, multi-site randomized trials that simultaneously fix the three target conditions against a standardized accuracy reference.
2. Adopt a consensus operational definition and instrument for care-plan accuracy, so as to enable future meta-analysis.
3. Measure longer-term and transfer outcomes, including whether early AI support helps or hinders the consolidation of independent clinical reasoning, and test for differential (equity) effects across learner ability.
4. Design replication for a moving target: multi-site randomized controlled trials should evaluate current-generation LLM-based systems using standardized, pre-registered prompt protocols; report the model name, version, and access date as essential study characteristics; and measure both immediate supported care-plan accuracy and the delayed retention of independent, unsupported clinical reasoning, so that effects can be re-estimated as the underlying models evolve.
5. Maintain the evidence map of this review as a living systematic review, with search re-runs at pre-specified intervals and outcome-level certainty ratings revisited as generative models are superseded (Section 5.6).

6.3 Concluding Statement

AI-driven decision support is poised to become a standard feature of the telemedicine practicum. Whether it strengthens or quietly erodes the clinical reasoning at the heart of nursing will depend less on the sophistication of the algorithm than on the pedagogy that surrounds it. The evidence assembled in this review argues for cautious, designed, and supervised adoption, harnessing the demonstrated gains in accuracy while deliberately protecting the deliberate practice through which student nurses become safe, independent clinicians.

References

- Ainslie, M., & Bragdon, C. (2018). Telemedicine simulation in online family nurse practitioner education: Clinical competency and technology integration. *Journal of the American Association of Nurse Practitioners*, 30(8), 430–434. <https://doi.org/10.1097/JXX.0000000000000071>
- Aromataris, E., & Munn, Z. (Eds.). (2020). *JBIM manual for evidence synthesis*. Joanna Briggs Institute. <https://doi.org/10.46658/JBIMES-20-01>
- Benner, P. (1984). *From novice to expert: Excellence and power in clinical nursing practice*. Addison-Wesley.
- Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, 368, l6890. <https://doi.org/10.1136/bmj.l6890>
- Cao, Y., Hu, L., et al. (2025). Can large language models facilitate the effective implementation of nursing processes in clinical settings? *BMC Nursing*, 24, 394. <https://doi.org/10.1186/s12912-025-03010-2>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine*, 84(8), 1022–1028. <https://doi.org/10.1097/ACM.0b013e3181ace703>
- Eckhoff, D. O., Guido-Sanz, F., & Anderson, M. (2022). Telehealth across nursing education: Findings from a

- national study. *Journal of Professional Nursing*, 42, 308–314. <https://doi.org/10.1016/j.profnurs.2022.07.013>
- Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 315(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>
- Gerçek, A., Çiftci, N., Durmuş, M., Sarman, A., Taşcı, Ö., & Yıldız, M. (2025). ChatGPT in nursing: Applications, advantages, and challenges in education, research, and clinical practice. *Annals of Biomedical Engineering*, 53(11), 3202–3207. <https://doi.org/10.1007/s10439-025-03832-w>
- Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. <https://doi.org/10.1136/amiajnl-2011-000089>
- Gosak, L., Pruinelli, L., Topaz, M., & Štiglic, G. (2024). The ChatGPT effect and transforming nursing education with generative AI: Discussion paper. *Nurse Education in Practice*, 75, 103888. <https://doi.org/10.1016/j.nepr.2024.103888>
- Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650), 924–926. <https://doi.org/10.1136/bmj.39489.470347.AD>
- Herdman, T. H., Kamitsuru, S., & Lopes, C. T. (Eds.). (2021). *NANDA International nursing diagnoses: Definitions and classification 2021–2023* (12th ed.). Thieme.
- Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2024). *Cochrane handbook for systematic reviews of interventions* (Version 6.5). Cochrane. <https://training.cochrane.org/handbook>
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, 327(7414), 557–560. <https://doi.org/10.1136/bmj.327.7414.557>
- Hinsche, L., Hasseler, M., Tischendorf, T., & Schaal, T. (2026). Rapid review on generative AI in nursing education. *Frontiers in Health Services*, 5, 1725425. <https://doi.org/10.3389/frhs.2025.1725425>
- Ho, K.-F., Chou, P.-H., & Chung, M.-H. (2023). Comparison of nursing diagnostic accuracy when aided by Knowledge-Based Clinical Decision Support Systems with Clinical Diagnostic Validity and Bayesian Decision Models for psychiatric care plan formulation among nursing students: A quasi-experimental study. *BMC Nursing*, 22, 142. <https://doi.org/10.1186/s12912-023-01292-y>
- Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O’Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). *Mixed Methods Appraisal Tool (MMAT), version 2018: User guide*. McGill University.
- Laun, M., et al. (2026). ChatGPT in nursing education: Who benefits most? *The Journal of Experimental Education*. Advance online publication. <https://doi.org/10.1080/00220973.2026.2637671>
- Liu, J., Liu, F., Fang, J., & Liu, S. (2023). The application of Chat Generative Pre-trained Transformer in nursing education. *Nursing Outlook*, 71(6), 102064. <https://doi.org/10.1016/j.outlook.2023.102064>
- McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., & Lefebvre, C. (2016). PRESS Peer Review of Electronic Search Strategies: 2015 guideline statement. *Journal of Clinical Epidemiology*, 75, 40–46. <https://doi.org/10.1016/j.jclinepi.2016.01.021>
- Mikkonen, K., et al. (2026). Artificial intelligence technologies supporting nurses’ clinical decision-making: A systematic review. *Journal of Clinical Nursing*. Advance online publication. <https://doi.org/10.1111/jocn.70156>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.

<https://doi.org/10.1136/bmj.n71>

- Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 11(6), 887. <https://doi.org/10.3390/healthcare11060887>
- Smith, T. S., Watts, P., & Moss, J. S. (2018). Using simulation to teach telehealth nursing competencies. *Journal of Nursing Education*, 57(10), 624–627. <https://doi.org/10.3928/01484834-20180921-10>
- Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., et al. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, 355, i4919. <https://doi.org/10.1136/bmj.i4919>
- Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., et al. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, 366, 14898. <https://doi.org/10.1136/bmj.14898>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Topaz, M., Peltonen, L.-M., Michalowski, M., Stiglic, G., Ronquillo, C., Pruinelli, L., Song, J., O'Connor, S., Miyagawa, S., & Fukahori, H. (2025). The ChatGPT effect: Nursing education and generative artificial intelligence. *Journal of Nursing Education*, 64(6), e40–e43. <https://doi.org/10.3928/01484834-20240126-01>
- Türk, G., Taşkıran, N., Er, O., Ölmez, E., & Bozkurt Kozan, F. (2025). Developing an artificial intelligence-based clinical decision support system for nursing diagnoses. *Nursing Outlook*, 73(6). Advance online publication.
- Verkuy, M., Lapum, J. L., & St-Amant, O. (2022). Family and community nursing: The use of an unfolding simulation via Zoom. *Clinical Simulation in Nursing*, 68, 58–63.