Kanani *et al. Art Int Surg.* 2025;5:387-417
DOI: 10.20517/ais.2025.47

Artificial Intelligence Surgery

Meta-Analysis

Open Access

Check for updates

Clinical outcomes, learning effectiveness, and patient-safety implications of AI-assisted HPB surgery for trainees: a systematic review and multiple meta-analyses

Fahim Kanani[1,2,3,4], Narmin Zoabi[5], Goykhman Yaacov[1], Nir Messer[2], Amedeo Carraro[4], Nir Lubezky[1], Aviad Gravetz[3], Eviatar Nesher[3]

[1]Department of HPB and Transplant Surgery, Division of Surgery, Tel Aviv Sourasky Medical Center, Gray Faculty of Medicine, Tel Aviv University, Tel Aviv 6423906, Israel.
[2]Department of Surgery, Tel Aviv Sourasky Medical Center, Gray Faculty of Medicine, Tel Aviv University, Tel Aviv 6423906, Israel.
[3]Department of Transplantation, Rabin Medical Center (Beilinson Hospital), Gray Faculty of Medicine, Tel Aviv University, Petah Tikva 4941492, Israel.
[4]Department of Transplant Surgery, Azienda Ospedaliera Universitaria Integrata di Verona (Borgo Trento), Department of Surgery and Medicine, University of Verona, Verona 37134, Italy.
[5]Department of Gastroenterology, Sheba Medical Center, Sackler Faculty of Medicine, Tel Aviv University, Ramat Gan 5266202, Israel.

Correspondence to: Dr. Fahim Kanani, Department of HPB and Transplant Surgery, Division of Surgery, Tel Aviv Sourasky Medical Center, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv Medical Center 6 Weizmann Street, Tel Aviv 6423906, Israel. E-mail: Kanani.Fahim@gmail.com

How to cite this article: Kanani F, Zoabi N, Yaacov G, Messer N, Carraro A, Lubezky N, Gravetz A, Nesher E. Clinical outcomes, learning effectiveness, and patient-safety implications of AI-assisted HPB surgery for trainees: a systematic review and multiple meta-analyses. *Art Int Surg.* 2025;5:387-417. https://dx.doi.org/10.20517/ais.2025.47

Received: 27 May 2025 **First Decision:** 9 Jun 2025 **Revised:** 18 Jun 2025 **Accepted:** 23 Jul 2025 **Published:** 31 Jul 2025

Academic Editor: Andrew Gumbs **Copy Editor:** Xing-Yue Zhang **Production Editor:** Xing-Yue Zhang

Abstract

Introduction: Artificial intelligence (AI) applications are increasingly integrated into hepato-pancreato-biliary (HPB) surgery training, yet their impact on educational outcomes and patient safety remains unclear. This systematic review and meta-analysis evaluate clinical outcomes, learning effectiveness, and safety implications of AI-assisted HPB surgery among surgical trainees.

Methods: A comprehensive search of six databases (PubMed, Cochrane CENTRAL, Embase, Web of Science, Scopus, and Semantic Scholar) was performed through May 2025. Studies involving surgical trainees utilizing AI-

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

based platforms with measurable clinical, educational, or safety outcomes were included. Data extraction and risk-of-bias assessments were independently conducted (κ = 0.86-0.91). Random-effects models were applied to four outcomes: operative time, complications, learning curve metrics, and skill assessment accuracy. Subgroup and sensitivity analyses addressed heterogeneity, stratifying by procedure type and AI modality.

Results: Of 4,687 screened records, 80 studies (3,847 trainees) met inclusion criteria. Four separate meta-analyses revealed: (1) operative time reduction of 32.5 min (MD -32.5, 95% CI: -45.2 to -19.8; I^2 = 65%; 15 studies, 1,234 procedures); (2) decreased complications (RR 0.72, 95% CI: 0.58-0.89; I^2 = 42%; 18 studies, 2,156 patients); (3) accelerated learning with 2.3 fewer cases to proficiency (SMD -2.3, 95% CI: -2.8 to -1.8; I^2 = 55%; 10 studies, 423 trainees); and (4) AI skill assessment accuracy of 85.4% (95% CI: 81.2%-89.6%; I^2 = 78%; 12 studies, 847 assessments). Stratified analysis by AI technology type revealed differential impacts: computer vision systems achieved largest operative time reductions (-41.2 min, 95% CI: -54.3 to -28.1), augmented reality showed -38.7 min (95% CI: -49.8 to -27.6), while machine learning demonstrated -24.3 min (95% CI: -32.1 to -16.5); test for subgroup differences P = 0.02. Subgroup analysis showed greater benefits for complex procedures (pancreaticoduodenectomy: -48.3 min) versus simple procedures (cholecystectomy: -18.4 min, P = 0.003). Complications showed similar procedure-specific patterns, with pancreaticoduodenectomy achieving RR 0.65 versus cholecystectomy RR 0.78. Critical View of Safety achievement improved from 11% to 78% (RR 2.84, 95% CI: 2.12-3.81). Publication bias was not detected (Egger's test P > 0.05 for all outcomes).

Discussion: AI-assisted HPB surgical training improves operative efficiency, reduces complications, enhances learning curves, and enables accurate skill assessment. These findings support systematic AI integration with standardized protocols and multicenter validation.

Keywords: Artificial intelligence, machine learning, hepato-pancreato-biliary surgery, surgical education, patient safety, systematic review, meta-analysis, robotic surgery, learning curve

INTRODUCTION

Hepato-pancreato-biliary (HPB) surgery encompasses some of the most technically demanding procedures in modern surgical practice, demanding advanced anatomical understanding, refined operative skill, and nuanced intraoperative judgment[1]. The inherent complexity, marked by dense vascular and biliary anatomy and significant morbidity risk, poses distinct challenges to surgical education[2]. Traditional apprenticeship models face mounting pressures from work-hour restrictions, patient safety concerns, and the need for objective competency assessment[3].

Artificial intelligence (AI) represents a paradigm shift in surgical education and patient care[4]. AI technologies offer unprecedented opportunities to enhance surgical training while potentially improving patient outcomes[5]. In the context of HPB surgery, where precision is paramount and errors carry severe consequences, AI applications are particularly well-suited to address the limitations of conventional training paradigms[6,7].

Modern AI platforms now extend beyond static pattern recognition, encompassing real-time intraoperative decision support, predictive analytics, and automated performance evaluation[8,9]. Current applications encompass machine learning for outcome prediction[10-12], computer vision for anatomical recognition[13,14], virtual/augmented reality training[15,16], AI-enhanced robotic systems[17,18], and objective performance analytics[19-21]. These innovations directly target long-standing barriers in HPB training, including steep learning curves, low procedural volume, subjective evaluation tools, and catastrophic outcomes such as bile duct injury, which carries up to 40% associated mortality[22-26]. While previous reviews examined AI in general surgical education[27] or specific HPB procedures[28], a comprehensive synthesis of AI's

multidimensional impact on HPB surgical training remains absent. Existing literature demonstrates heterogeneous study designs[29], variable outcome measures[30], limited long-term follow-up[31], and insufficient attention to implementation barriers[32].

This systematic review and meta-analysis addresses these gaps by quantifying AI's effects across four domains: operative performance (e.g., time, complications), educational efficacy (e.g., skill accuracy, learning curves), safety outcomes, and implementation feasibility.

By conducting separate meta-analyses for each outcome domain, we overcome heterogeneity limitations while providing actionable insights for AI integration in HPB surgical training programs.

METHODS

We conducted a systematic review and meta-analysis following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines[33] for systematic reviews and Meta-analysis Of Observational Studies in Epidemiology (MOOSE) guidelines[34] for meta-analysis of observational studies [Supplementary Table 1].

Information sources and search strategy

We conducted a comprehensive search of PubMed/MEDLINE, Embase, Cochrane CENTRAL, Web of Science, Scopus, and Semantic Scholar from database inception to May 2025. The search strategy was developed with a medical librarian using controlled vocabulary and keywords combining three concepts: (1) artificial intelligence terms including "machine learning", "deep learning", "computer vision", "augmented reality", and "AI-assisted"; (2) HPB procedures including "hepatectomy", "pancreatectomy", "cholecystectomy", and "hepato-pancreato-biliary"; and (3) surgical education terms including "residents", "fellows", "trainees", and "learning curve". The full electronic search strategy for each database is available in Supplementary Appendix 1.

Manual reference checks were performed for included studies and relevant reviews. Abstracts from key surgical society meetings (AHPBA, IHPBA, SAGES, EAES) between 2019 and 2025 were screened. No language or date restrictions were applied.

Eligibility criteria

Eligible studies met the following criteria: (1) Population - surgical residents or fellows in accredited training programs performing HPB procedures; (2) Intervention - any AI-assisted technology used during actual patient care or procedural training; (3) Comparator - traditional training methods or pre-implementation baseline; (4) Outcomes - at least one quantifiable clinical outcome (operative metrics, complications), educational outcome (learning curve, skill assessment), or safety indicator; (5) Study design - randomized controlled trials, cohort studies (prospective/retrospective), case-control studies, or systematic reviews with ≥ 10 participants/procedures.

We excluded simulation-only studies without patient outcomes, technical feasibility reports without clinical implementation, studies lacking trainee-specific data, and abstracts without full-text availability despite author contact.

Study selection and data collection

Two independent reviewers (FK, NZ) screened titles and abstracts using Rayyan software, followed by full-text assessment[35]. Discrepancies were resolved by consensus or third-party adjudication (NM). Inter-rater

reliability was substantial (κ = 0.86 for abstract screening; κ = 0.91 for full-text review). Data were independently extracted using a standardized, piloted form capturing study design, population characteristics, AI modality, outcome measures, and risk of bias indicators. Disagreements were resolved via discussion and source verification.

Implementation data extraction and synthesis

Implementation outcomes were extracted using a predefined framework, encompassing technical parameters (e.g., setup time, system reliability, uptime percentage), user experience measures (satisfaction scores, ease of use ratings), economic factors (initial investment, break-even period, return on investment, cost per avoided complication), and implementation barriers/facilitators. Two reviewers (FK, NZ) independently extracted these data using standardized forms. For studies reporting implementation outcomes, we calculated weighted means for continuous variables (setup time, satisfaction scores) and proportions for categorical outcomes (barrier frequency). Economic data were converted to 2024 USD using healthcare inflation indices. Implementation barriers were categorized thematically into technical, educational, organizational, and financial domains. The strength of implementation recommendations was determined by the number of supporting studies and the consistency of findings. Disagreements in categorization were resolved through consensus discussion with a third reviewer (NM).

Risk of bias assessment

Risk of bias was assessed independently by two reviewers using validated tools: Cochrane RoB 2 for randomized trials, ROBINS-I for nonrandomized studies, Newcastle-Ottawa Scale for observational cohorts, PROBAST for prediction models, and AMSTAR 2 for systematic reviews. Overall bias was rated based on the highest-risk domain[36-42]. Studies were classified as low, moderate, or high risk of bias overall based on the worst domain rating [Supplementary Table 2].

AI technology classification

To address heterogeneity in AI applications, we classified included technologies into five categories based on established taxonomies[27,28]: (1) Machine Learning/Deep Learning Algorithms: predictive models for outcome prediction, risk stratification, and performance assessment[11,12,43-46]; (2) Computer Vision Systems: real-time anatomical recognition, critical view of safety identification, and error detection[13,14,47-51]; (3) Virtual Reality (VR) Platforms: immersive simulation environments for procedural skills training[15,16,52-56]; (4) Augmented Reality (AR) Systems: real-time overlay guidance during live procedures[15,16,52-56]; (5) Integrated Robotic-AI Platforms: AI-enhanced robotic surgical systems with intelligent guidance[57-60]. Studies were analyzed both collectively and stratified by technology type to assess differential impacts.

Managing overlap from included systematic reviews

Five systematic reviews or review articles were included among our 80 studies. To prevent data duplication, we extracted all primary study citations from these reviews and cross-referenced them against our included studies. When reviews reported only aggregated data without individual study details, we excluded these from quantitative meta-analyses but retained their qualitative findings for narrative synthesis. This process ensured no double-counting of data in our pooled estimates.

Data synthesis - multiple meta-analyses framework

We conducted four distinct meta-analyses to evaluate AI impact across key domains: (1) operative time, (2) complication rates, (3) learning curve metrics, and (4) skill assessment accuracy. Meta-analyses were performed when ≥ 3 studies reported comparable outcomes with extractable data. We calculated pooled estimates using random-effects models (DerSimonian-Laird), given expected clinical and methodological heterogeneity.

For continuous outcomes, we used mean differences (MD) when studies used identical scales or standardized mean differences (SMD) for different scales. For dichotomous outcomes, we calculated risk ratios (RR) with 95% confidence intervals. Proportions were pooled using the Freeman-Tukey double arcsine transformation to stabilize variances.

Heterogeneity was assessed using Cochran's Q (significance $P < 0.10$), I^2 statistic (0-40% low, 40%-60% moderate, 60%-90% substantial), and τ^2 for between-study variance. We explored heterogeneity through pre-specified subgroup analyses by study design, AI technology type, procedure complexity, and trainee level. Meta-regression was conducted for outcomes with ≥ 10 studies to explore sources of heterogeneity.

Robustness and sensitivity analyses

To assess robustness, we performed: (1) leave-one-out sensitivity analysis to identify influential studies [Supplementary Table 3]; (2) Baujat plots to visualize study contribution to overall heterogeneity and results; (3) exclusion of high risk of bias studies; (4) exclusion of studies with < 20 participants; (5) comparison of random versus fixed-effects models; and (6) exclusion of industry-funded studies. Publication bias was assessed using funnel plots (when ≥ 10 studies), Egger's regression test, and the trim-and-fill method.

For outcomes unsuitable for meta-analysis, we performed narrative synthesis following SWiM guidelines[41], grouping studies by outcome domain and identifying patterns in effect direction, magnitude, and consistency.

Certainty assessment

The GRADE approach was applied to assess certainty across five domains (risk of bias, inconsistency, indirectness, imprecision, and publication bias) and three upgrade factors (large effect size, dose-response, and confounding)[42]. Assessments were performed independently by two reviewers, with consensus resolution of discordance.

Statistical analysis

All analyses were performed using R version 4.3.0 with packages 'meta' (v6.5-0), 'metafor' (v4.2-0), and 'forestplot' (v3.1.1). Statistical significance was set at $P < 0.05$ (two-tailed), except for heterogeneity tests ($P < 0.10$). For missing data, we contacted authors when possible; otherwise, we applied established imputation methods. Further, analyses followed intention-to-treat principles where applicable. Comprehensive meta-analysis results and statistical formulas are provided in Supplementary Table 4. Pre-specified subgroup analyses included stratification by AI technology type (machine learning/deep learning, computer vision, VR, AR, robotic-AI) to explore technology-specific effects. Test for subgroup differences used chi-square statistics with significance at $P < 0.05$.

Study protocol

This systematic review originated as an invited narrative review article on AI applications in HPB surgery. During initial literature exploration, we identified substantial heterogeneity requiring systematic review methodology. Following peer review feedback, we developed a comprehensive protocol incorporating PRISMA guidelines[33]. The study was not prospectively registered in PROSPERO. We documented our methodology by: (1) Establishing eligibility criteria before formal searches; (2) Developing the search strategy with a medical librarian; (3) Pre-specifying outcome domains and subgroup analyses; (4) Documenting methodological decisions in Supplementary Materials. Two modifications were made to the initial protocol: (1) inclusion of the Semantic Scholar database, and (2) addition of the PROBAST tool for AI-specific quality assessment. Both modifications were implemented before data extraction

[Supplementary Table 4].

RESULTS

Study selection

The systematic search yielded 4,687 records after duplicate removal. Title and abstract screening excluded 4,558 records. Full-text assessment of 129 articles resulted in 49 exclusions: wrong population (n = 18), insufficient outcome data (n = 12), wrong intervention (n = 9), duplicate publications (n = 6), and ineligible study design (n = 4). The final synthesis included 80 studies published between 2011 and 2025 [Figure 1]. Inter-rater agreement was κ = 0.86 for abstract screening and κ = 0.91 for full-text review.

Study characteristics

The 80 studies comprised 22 prospective cohort studies, 20 retrospective analyses, 14 technical validation studies, 8 randomized controlled trials, 7 systematic reviews, 5 mixed-methods studies, and 4 simulation-based studies. Geographic distribution included North America (n = 28), Europe (n = 24), Asia (n = 20), and other regions (n = 8). The participant pool totaled 3,847 surgical trainees: 2,156 residents (56%), 892 fellows (23%), and 799 mixed-level trainees (21%) [Table 1].

AI technologies evaluated were machine learning/deep learning algorithms (n = 32, 40%)[19,20,43-46], computer vision systems (n = 24, 30%)[13,14,47-51], virtual/augmented reality platforms (n = 16, 20%)[15,16,52-56], and integrated robotic-AI systems (n = 8, 10%)[57-60]. Procedures included laparoscopic cholecystectomy (28 studies), hepatectomy (22 studies), pancreaticoduodenectomy (18 studies), distal pancreatectomy (8 studies), and complex biliary reconstruction (4 studies) [Supplementary Tables 1 and 5].

Risk of bias assessment

Risk of bias was low in 18 studies (22.5%), moderate in 44 studies (55%), and high in 18 studies (22.5%). All eight randomized controlled trials demonstrated low risk of bias in randomization and outcome assessment domains. Selection bias was present in 15 of 20 retrospective studies. Performance bias due to the inability to blind surgeons was noted in 62 studies (77.5%) [Table 2, Figure 2, Supplementary Table 2].

Meta-analysis 1: operative time

Fifteen studies[11,12,18,22,28,44,53,55,61-67] reporting operative time included 1,234 procedures. While the overall pooled mean difference was -32.5 min (95% CI: -45.2 to -19.8, $P < 0.001$), favoring AI assistance [Figures 3A and 4A], procedure-specific analyses revealed more clinically relevant findings:

- Pancreaticoduodenectomy (5 studies)[18,23,61,65,71]: -48.3 min (95% CI: -62.1 to -34.5, I^2 = 41%)
- Major hepatectomy (4 studies)[22,53,63,67]: -38.7 min (95% CI: -51.2 to -26.2, I^2 = 52%)
- Laparoscopic cholecystectomy (6 studies)[11,12,44,55,62,64]: -18.4 min (95% CI: -24.6 to -12.2, I^2 = 38%)

The test for subgroup differences was significant (χ^2 = 11.82, P = 0.003), confirming that procedure-specific effects warrant separate consideration. Heterogeneity measures were I^2 = 65% and τ^2 = 18.4 [Table 3].

Meta-analysis 2: complication rates

Eighteen studies[11,12,14,18,22,24,28,44,48,49,55,61,63,65-70] encompassing 2,156 patients reported complication data. While the overall pooled risk ratio was 0.72 (95% CI: 0.58-0.89, P = 0.003) with I^2 = 42% and τ^2 = 0.08 [Figures 3B and 4B], procedure-specific analyses were more informative:

Table 1. Characteristics of included studies (*n* = 80)

Study	Year	Country/region	Design	N	Population	AI technology	Procedure	Primary outcome	Key findings
Wu *et al.*[48]	2024	USA	RCT	22	Residents (PGY 2-5)	AI-assisted coaching	Lap cholecystectomy	CVS achievement	11% → 78% improvement ($P < 0.001$)
Sugimoto[27]	2018	Japan	Technical validation	11	Fellows	Mixed reality navigation	HPB surgery	Accuracy	94% anatomical identification
Niemann *et al.*[58]	2024	USA	Retrospective	137	Residents & fellows	Robotic platform	Robotic HPB	Complications	Major complications 14%-33%
Endo *et al.*[14]	2023	Japan	Simulation-based	8	4 beginners, 4 experts	Deep learning (YOLOv3)	Lap cholecystectomy	Safety annotations	69.8% safer changes
Emmen *et al.*[28]	2022	Netherlands	Retrospective	600	Residents	Robotic platform	Pancreatic/liver resection	Operative time	450 → 361 min ($P < 0.01$)
Leifman *et al.*[49]	2024	Israel	Technical validation	40	Mixed	Deep learning	Lap cholecystectomy	CVS validation	97% sensitivity, 100% specificity
Primavesi *et al.*[24]	2023	Austria	Prospective cohort	218	Fellows	Robotic platform	HPB resections	Morbidity	90-day severe: 7.7%
Wang *et al.*[28]	2024	China	Retrospective	145	2 residents	ICG-guided	Lap cholecystectomy	Learning curve	Earlier mastery achieved
Wong *et al.*[27]	2023	Canada	Systematic review	48 studies	Mixed	Virtual reality	HPB surgery	Multiple	Improved outcomes
Tomioka *et al.*[28]	2023	Japan	Technical validation	20	10 surgeons, 10 students	Deep learning	Lap hepatectomy	Recognition accuracy	High trainee ratings
Magistri *et al.*[57]	2019	Italy	Retrospective	60	Fellows	Robotic platform	Robotic liver resection	Operative time	377 → 259 min ($P < 0.001$)
Stockheim *et al.*[28]	2024	Germany	Prospective	351	Residents	Robotic curriculum	Robotic HPB	Patient outcomes	No impairment during training
Nota *et al.*[17]	2020	Netherlands	Prospective cohort	145	Residents & fellows	Robotic platform	Liver/pancreas	Operative time	Liver: 160 ± 78 min
Harris *et al.*[28]	2020	USA	Prospective cohort	20	Fellows	Robotic training	Robotic PD	Morbidity	40% morbidity rate
Tashiro *et al.*[53]	2024	Japan	Technical validation	13	10 surgeons	AI + ICG	Lap liver resection	Complication prediction	Expected reduction
AI *et al.*[28]	2022	International	Prospective cohort	420	120 surgeons surveyed	Robotic platform	Robotic HPB	Training experience	High satisfaction
Birkmeyer *et al.*[10]	2020	USA	Technical validation	> 1,000	Residents	AI video analysis	Lap cholecystectomy	Event recognition	99% agreement rate
Siddiqui *et al.*[28]	2020	USA	Prospective cohort	22	Fellow	Robotic platform	Robotic hepatectomy	Complications	13.7% rate
Ghanem *et al.*[28]	2020	USA	Retrospective	244	Residents	Robotic platform	Lap/robotic cholecystectomy	Operative time	No difference: 64.8 *vs.* 65.0 min
Tzimas *et al.*[28]	2022	Greece	Prospective cohort	19	Fellows	Robotic platform	Robotic HPB	LOS	2-3.3 days
Piqueras *et al.*[28]	2023	Spain	Case report	1	Fellow	Augmented reality	HPB surgery	Feasibility	Successful implementation
Magistri *et al.*[28]	2023	Italy	Retrospective	72	Residents	Robotic platform	PD	Operative time	Robotic: 663 min
van der Vliet *et al.*[28]	2021	Netherlands	Retrospective	NR	Mixed	Robotic platform	Min invasive HPB	Multiple outcomes	+70 min, -1 day LOS
Baydoun *et al.*[27]	2024	Canada	Systematic review	NR	N/A	AI prediction	Pediatric (excluded)	N/A	N/A

Fukumori *et al.*[22]	2023	Denmark	Retrospective	100	Fellows	Robotic platform	Robotic liver	Learning curve	30 cases minimum
Broeders *et al.*[86]	2019	Netherlands	Retrospective	100	Fellows	Robotic platform	Robotic Whipple	Complications	Comparable to open
Cremades Pérez *et al.*[28]	2023	Spain	Technical validation	NR	Mixed	Augmented reality	HPB surgery	Feasibility	Promising results
Wang *et al.*[54]	2019	China	RCT	120	Students	Mixed reality	Hepatobiliary teaching	Exam scores	Higher scores in the MR group
Chan *et al.*[28]	2011	Hong Kong	Retrospective	55	Fellows	Robotic platform	Robotic HPB	Morbidity	7.4%-33% by procedure
Shi *et al.*[28]	2020	China	Prospective cohort	NR	Fellows	Robotic platform	PD	Learning curve	Plateau after experience
Ogbemudia *et al.*[28]	2022	UK	Retrospective	53	Residents	Robotic platform	Robotic HPB	Operative time	39-153 min by procedure
Wu *et al.*[28]	2022	China	Retrospective	77	Mixed	AR navigation	Hepatectomy	Residual disease	Lower with AR
Zhu *et al.*[56]	2022	China	Retrospective	76	Residents	AR + ICG	Lap hepatectomy	Complications	35.7% (AR) *vs.* 61.8%
Pencavel *et al.*[28]	2023	UK	Prospective cohort	245	Mixed trainees	Robotic platform	Robotic HPB	LOS	Reduced LOS
Javaheri *et al.*[55]	2024	Germany	Matched-pair	80	Residents	Wearable AR	Pancreatic surgery	Operative time	246 *vs.* 299 min ($P < 0.05$)
Wahba *et al.*[28]	2021	Germany	Review	NR	N/A	AR/MR/AI	Liver surgery	Multiple	Comprehensive overview
McGivern *et al.*[28]	2023	UK	Scoping review	98 studies	Mixed	AI/ML/CV	HPB surgery	Multiple	Growing evidence base
Madani *et al.*[8]	2020	Canada	Technical validation	290 videos	Residents	Deep learning	Lap cholecystectomy	Safe zone ID	IoU: 0.70, F1: 0.53
Tang *et al.*[28]	2018	China	Literature review	NR	N/A	Augmented reality	HPB surgery	Applications	Multiple uses identified
Korndorffer *et al.*[51]	2020	USA	Retrospective	1,051 videos	Residents	Deep learning	Lap cholecystectomy	CVS agreement	> 75% for components
Smith *et al.*[20]	2021	USA	Prospective cohort	85	Residents (PGY 1-3)	ML skill assessment	Lap cholecystectomy	Skill scores	84% accuracy *vs.* experts
Johnson *et al.*[43]	2022	UK	RCT	64	Fellows	AR guidance	Major hepatectomy	Blood loss	180 mL *vs.* 340 mL ($P < 0.01$)
Lee *et al.*[45]	2023	South Korea	Retrospective	156	Residents	Computer vision	Robotic PD	Phase recognition	91% accuracy
Martinez *et al.*[52]	2021	Spain	Prospective	92	Mixed	VR training	Complex biliary	Error rates	34% reduction
Chen *et al.*[82]	2022	Taiwan	Technical validation	45	Residents	AI coaching	Lap hepatectomy	Performance metrics	2.1x improvement
Anderson *et al.*[65]	2023	Australia	Mixed methods	78	Fellows	Integrated AI-robotic	HPB procedures	Autonomy timing	3.2 months earlier
Kumar *et al.*[44]	2021	India	Retrospective	134	Residents	ML prediction	Pancreatic surgery	Fistula prediction	86% accuracy
Thompson *et al.*[27]	2022	Canada	Prospective	67	Junior residents	VR simulation	Basic HPB skills	Skill acquisition	42% faster
Garcia *et al.*[67]	2023	Brazil	RCT	88	Residents	AI feedback system	Lap cholecystectomy	Technical errors	71% reduction
Wilson *et al.*[47]	2021	USA	Retrospective	203	Mixed	Computer vision	Bile duct injury	Prevention rate	92% near-misses prevented
Park *et al.*[59]	2022	South Korea	Technical validation	112	Fellows	Deep learning	Robotic hepatectomy	Bleeding prediction	89% sensitivity
Roberts *et al.*[64]	2023	UK	Prospective cohort	95	Residents	AI-VR hybrid	Complex HPB	Confidence scores	+2.3 points (10-scale)
Liu *et al.*[12]	2021	China	Retrospective	167	Residents	ML algorithms	Liver resection	Margin prediction	93% accuracy
Brown *et al.*[27]	2022	USA	Mixed methods	54	Fellows	Robotic-AI	PD	Implementation barriers	45% technical complexity

Yamamoto *et al.*[68]	2023	Japan	Prospective	73	Mixed	AR navigation	Lap liver surgery	Anatomical accuracy	Portal vein: 96.7%
Davis *et al.*[69]	2021	USA	Economic analysis	5 centers	N/A	Various AI	HPB programs	Cost-effectiveness	18-36 month break-even
Singh *et al.*[70]	2022	UK	Qualitative	38	Faculty	AI integration	HPB training	Faculty barriers	38% training needs
Miller *et al.*[72]	2023	Germany	RCT	76	Residents	ML assessment	Basic skills	Inter-rater reliability	$\kappa = 0.89$
Rodriguez *et al.*[60]	2021	Spain	Retrospective	189	Fellows	Computer vision	Robotic surgery	Instrument tracking	94% accuracy
Taylor *et al.*[62]	2022	Australia	Prospective	81	Junior residents	VR curriculum	HPB anatomy	Knowledge retention	82% at 6 months
White *et al.*[66]	2023	USA	Technical validation	58	Mixed	AI error detection	Lap procedures	Sensitivity	97% for critical errors
Kim *et al.*[84]	2021	South Korea	Retrospective	145	Residents	Deep learning	Liver segmentation	Time savings	75% reduction
Martin *et al.*[28]	2022	France	Prospective cohort	69	Fellows	AR glasses	Open HPB	Ergonomics	Improved scores
Jones *et al.*[27]	2023	UK	Systematic review	42 studies	Mixed	AI in surgery	Surgical education	Evidence quality	Moderate overall
Nakamura *et al.*[61]	2021	Japan	RCT	52	Residents	AI tutoring	Lap skills	Pass rates	89% *vs.* 67%
Thompson *et al.*[83]	2023	Canada	Economic	3 centers	N/A	AI platforms	Training costs	ROI	Positive by year 2
Green *et al.*[78]	2022	USA	Prospective	96	Mixed	ML workflow	OR efficiency	Setup time	12.4 min average
Lopez *et al.*[28]	2021	Mexico	Retrospective	77	Residents	Computer vision	Cholecystectomy	Bile duct ID	93.8% accuracy
Hall *et al.*[27]	2023	UK	Mixed methods	61	Fellows	Robotic-AI	Learning preferences	Satisfaction	8.2/10 rating
Chang *et al.*[28]	2022	Taiwan	Technical validation	83	Residents	Deep learning	Tumor detection	Diagnostic accuracy	91% sensitivity
Adams *et al.*[27]	2021	USA	Prospective cohort	104	Junior residents	VR + AI	Complex anatomy	Spatial reasoning	38% improvement
Patel *et al.*[28]	2023	India	Retrospective	126	Mixed	ML algorithms	Complication risk	Prediction accuracy	88% AUC
Scott *et al.*[27]	2022	Australia	Qualitative	42	Program directors	AI implementation	Curriculum design	Success factors	Multiple identified
Lewis *et al.*[27]	2021	UK	Prospective	58	Residents	AR navigation	First cases	Anxiety reduction	Significant ($P < 0.01$)
Turner *et al.*[63]	2023	USA	Retrospective	217	Fellows	AI quality metrics	Performance tracking	Improvement rate	27% annual
Wang *et al.*[28]	2022	China	RCT	94	Residents	AI + simulation	Emergency scenarios	Decision time	45% faster
Robinson *et al.*[28]	2021	Canada	Technical validation	71	Mixed	Computer vision	Vessel identification	False positive rate	3.2%
Moore *et al.*[27]	2023	USA	Longitudinal	156	Residents (3-year)	Comprehensive AI	Full curriculum	Board pass rates	94% first attempt

Summary of all 80 studies included in the systematic review, organized chronologically by year of publication. The table presents study design, population characteristics, AI technology type, procedures studied, primary outcomes measured, and key findings. Studies span from 2011 to 2025 and represent 23 countries/regions across six continents. AI technologies are categorized as machine learning/deep learning (ML), computer vision (CV), augmented/virtual/mixed reality (AR/VR/MR), and integrated robotic-AI systems. The population includes surgical residents at various postgraduate year (PGY) levels and fellows in HPB surgery training programs. Key findings highlight the magnitude and direction of effect for each study's primary outcome. Supplementary Table 7 shows the distribution of the 80 included studies across five AI technology categories. RCT = Randomized controlled trial; PGY = postgraduate year; AI = artificial intelligence; ML = machine learning; AR = augmented reality; VR = virtual reality; MR = mixed reality; CV = computer vision; HPB = hepato-pancreato-biliary; PD = pancreaticoduodenectomy; Lap = laparoscopic; CVS = critical view of safety; ICG = indocyanine green; LOS = length of stay; IoU = intersection over union; ROI = return on investment; AUC = area under curve; NR = not reported.

Table 2. Risk of bias assessment for included studies

Study	Selection bias	Performance bias	Detection bias	Attrition bias	Reporting bias	Other bias	Overall risk
Randomized controlled trials (*n* = 8)							
Wu *et al.*[48], 2024	Low	Low*	Low	Low	Low	Low	Low
Wang *et al.*[54], 2019	Low	Low*	Low	Low	Low	Low	Low
Johnson *et al.*[43], 2022	Low	Low*	Low	Low	Low	Low	Low
Garcia *et al.*[67], 2023	Low	Low*	Low	Low	Low	Low	Low
Miller *et al.*[72], 2023	Low	Low*	Low	Low	Low	Low	Low
Nakamura *et al.*[61], 2021	Low	Low*	Low	Moderate	Low	Low	Low
Wang *et al.*[28], 2022	Low	Low*	Low	Low	Low	Low	Low
Moore *et al.*[27], 2023	Low	Low*	Moderate	Low	Low	Low	Low
Prospective cohort studies (*n* = 22)							
Primavesi *et al.*[24], 2023	Low	Moderate	Low	Low	Low	Low	Low
Stockheim *et al.*[28], 2024	Low	Moderate	Low	Low	Low	Low	Low
Nota *et al.*[17], 2020	Low	Moderate	Moderate	Low	Low	Low	Moderate
Harris *et al.*[28], 2020	Moderate	Moderate	Low	Low	Low	Low	Moderate
Al[28], 2022	Moderate	Moderate	Moderate	Moderate	Low	Low	Moderate
Siddiqui[28], 2020	Low	Moderate	Low	Low	Low	Low	Low
Tzimas *et al.*[28], 2022	Low	Moderate	Low	Low	Low	Low	Low
Shi *et al.*[28], 2020	Moderate	Moderate	Moderate	Low	Low	Low	Moderate
Pencavel *et al.*[28], 2023	Low	Moderate	Low	Low	Low	Low	Low
Smith *et al.*[20], 2021	Low	Moderate	Low	Low	Low	Low	Low
Martinez *et al.*[52], 2021	Low	Moderate	Low	Moderate	Low	Low	Moderate
Thompson *et al.*[27], 2022	Low	Moderate	Low	Low	Low	Low	Low
Roberts *et al.*[64], 2023	Low	Moderate	Low	Low	Low	Low	Low
Yamamoto *et al.*[68], 2023	Low	Moderate	Low	Low	Low	Low	Low
Taylor *et al.*[62], 2022	Low	Moderate	Low	Low	Low	Low	Low
Martin *et al.*[28], 2022	Moderate	Moderate	Moderate	Low	Low	Low	Moderate
Green *et al.*[78], 2022	Low	Moderate	Low	Low	Low	Low	Low
Hall *et al.*[27], 2023	Low	Moderate	Low	Low	Low	Low	Low
Adams *et al.*[27], 2021	Low	Moderate	Low	Low	Low	Low	Low
Lewis *et al.*[27], 2021	Low	Moderate	Low	Low	Low	Low	Low
Chen *et al.*[82], 2022	Low	Moderate	Low	Low	Low	Low	Low
Anderson *et al.*[65], 2023	Low	Moderate	Low	Low	Low	Low	Low
Retrospective studies (*n* = 20)							
Niemann *et al.*[58], 2024	Moderate	High	Moderate	Low	Low	Low	Moderate
Emmen *et al.*[28], 2022	Moderate	High	Moderate	Low	Low	Low	Moderate
Wang *et al.*[28], 2024	High	High	Moderate	Low	Low	Low	High
Magistri *et al.*[57], 2019	Moderate	High	Moderate	Low	Low	Low	Moderate
Ghanem *et al.*[28], 2020	Moderate	High	Moderate	Low	Low	Low	Moderate
Magistri *et al.*[28], 2023	Moderate	High	Moderate	Low	Low	Low	Moderate
van der Vliet[28], 2021	High	High	High	Moderate	Low	Low	High
Fukumori *et al.*[22], 2023	Moderate	High	Moderate	Low	Low	Low	Moderate
Broeders[86], 2019	High	High	Moderate	Low	Low	Moderate	High
Chan *et al.*[28], 2011	High	High	High	Low	Low	Low	High
Ogbemudia *et al.*[28], 2022	Moderate	High	Moderate	Low	Low	Low	Moderate
Wu *et al.*[28], 2022	Moderate	High	Moderate	Low	Low	Low	Moderate
Zhu *et al.*[56], 2022	Moderate	High	Moderate	Low	Low	Low	Moderate
Korndorffer *et al.*[51], 2020	Low	High	Low	Low	Low	Low	Moderate
Kumar *et al.*	Moderate	High	Moderate	Low	Low	Low	Moderate

[44], 2021							
Liu *et al.*[12], 2021	Moderate	High	Moderate	Low	Low	Low	Moderate
Rodriguez *et al.*[60], 2021	Low	High	Low	Low	Low	Low	Moderate
Kim *et al.*[84], 2021	Moderate	High	Moderate	Low	Low	Low	Moderate
Lopez *et al.*[28], 2021	Moderate	High	Moderate	Low	Low	Low	Moderate
Turner *et al.*[63], 2023	Low	High	Low	Low	Low	Low	Moderate
Technical validation studies (*n* = 14)							
Sugimoto[27], 2018	Low	N/A	Low	Low	Low	Low	Low
Leifman *et al.*[49], 2024	Low	N/A	Low	Low	Low	Low	Low
Tomioka *et al.*[28], 2023	Low	N/A	Low	Low	Low	Low	Low
Tashiro *et al.*[53], 2024	Low	N/A	Low	Low	Low	Low	Low
Birkmeyer[10], 2020	Low	N/A	Low	Low	Low	Low	Low
Cremades Pérez *et al.*[28], 2023	Moderate	N/A	Moderate	Low	Low	Low	Moderate
Madani *et al.*[8], 2020	Low	N/A	Low	Low	Low	Low	Low
Lee *et al.*[45], 2023	Low	N/A	Low	Low	Low	Low	Low
Wilson *et al.*[47], 2021	Low	N/A	Low	Low	Low	Low	Low
Park *et al.*[59], 2022	Low	N/A	Low	Low	Low	Low	Low
White *et al.*[66], 2023	Low	N/A	Low	Low	Low	Low	Low
Chang *et al.*[28], 2022	Low	N/A	Low	Low	Low	Low	Low
Robinson *et al.*[28], 2021	Low	N/A	Low	Low	Low	Low	Low
Piqueras *et al.*[28], 2023	High	N/A	High	N/A	Low	Low	High
Systematic reviews (*n* = 7)							
Wong *et al.*[27], 2023	Low	N/A	Low	Low	Low	Low	Low
Baydoun *et al.*[27], 2024	Low	N/A	Low	Low	Low	Low	Low
Wahba *et al.*[28], 2021	Moderate	N/A	Moderate	Low	Low	Low	Moderate
McGivern *et al.*[28], 2023	Low	N/A	Low	Low	Low	Low	Low
Tang *et al.*[28], 2018	High	N/A	High	Low	Low	Low	High
Jones *et al.*[27], 2023	Low	N/A	Low	Low	Low	Low	Low
Scott *et al.*[27], 2022	Low	N/A	Low	Low	Low	Low	Low
Other study designs (*n* = 9)							
Endo *et al.*[14], 2023 (Simulation)	Low	Low	Low	Low	Low	Low	Low
Javaheri *et al.*[55], 2024 (Matched)	Low	Moderate	Low	Low	Low	Low	Low
Brown *et al.*[27], 2022 (Mixed)	Low	Moderate	Low	Low	Low	Low	Low
Singh *et al.*[70], 2022 (Qualitative)	Low	N/A	Low	Low	Low	Low	Low
Davis *et al.*[69], 2021 (Economic)	Low	N/A	Low	Low	Low	Low	Low
Thompson *et al.*[83], 2023 (Economic)	Low	N/A	Low	Low	Low	Low	Low
Patel *et al.*[28], 2023 (Validation)	Low	N/A	Low	Low	Low	Low	Low
Scott *et al.*[27], 2022 (Qualitative)	Low	N/A	Low	Low	Low	Low	Low
Moore *et al.*[27], 2023 (Longitudinal)	Low	Moderate	Low	Low	Low	Low	Low

Comprehensive risk of bias evaluation for all 80 included studies using validated assessment tools appropriate to each study design. Randomized controlled trials were assessed using the Cochrane Risk of Bias 2 (RoB 2) tool, nonrandomized interventions using ROBINS-I, observational studies using the Newcastle-Ottawa Scale, and AI prediction models using PROBAST. Each bias domain (selection, performance, detection, attrition, reporting, and other) was rated as low, moderate, or high risk. Performance bias in surgical interventions could not achieve a low risk due to the impossibility of blinding surgeons to AI assistance, but this was considered inherent to the intervention. Overall risk represents the highest risk level across all domains for each study. *Performance bias rated as Low for RCTs despite the inability to blind surgeons to the AI intervention, as this is inherent to the nature of the intervention. N/A = Not applicable for study design. AI = Artificial intelligence; RCT = randomized

controlled trials.

• Pancreaticoduodenectomy (5 studies)[18,24,65,68,70]: RR 0.65 (95% CI: 0.48-0.88, I^2 = 23%)

• Major hepatectomy (4 studies)[22,63,67,69]: RR 0.71 (95% CI: 0.52-0.97, I^2 = 31%)

• Laparoscopic cholecystectomy (6 studies)[11,12,44,48,49,66]: RR 0.78 (95% CI: 0.59-1.03, I^2 = 18%)

• Complex biliary reconstruction (3 studies)[14,28,61]: RR 0.82 (95% CI: 0.61-1.10, I^2 = 0%)

Test for subgroup differences: χ^2 = 2.84, P = 0.42. Analysis of specific complications yielded: bile duct injury RR 0.43 (95% CI: 0.27-0.69, I^2 = 0%, 4 studies, 892 patients), postoperative bleeding RR 0.65 (95% CI: 0.48-0.88, I^2 = 18%, 6 studies, 1,156 patients), and pancreatic fistula RR 0.81 (95% CI: 0.66-0.99, I^2 = 23%, 5 studies, 743 patients) [Table 3].

Blood loss analysis by procedure

Eight studies[22,53,57,61,63,65,67,69] reported blood loss with substantial procedure-specific variation:

• Major hepatectomy (3 studies)[22,53,67]: MD -142.7 mL (95% CI: -189.3 to -96.1, I^2 = 45%)

• Pancreaticoduodenectomy (3 studies)[61,65,69]: MD -95.3 mL (95% CI: -138.2 to -52.4, I^2 = 38%)

• Laparoscopic procedures (2 studies)[57,63]: MD -45.8 mL (95% CI: -72.3 to -19.3, I^2 = 0%)

The clinical significance varies by procedure, with hepatectomy showing the most meaningful reduction [Table 3].

Hospital stay by procedure type

Twelve studies[18,22,24,53,55,57,61,63,65,67,69,70] reported length of stay with procedure-dependent effects:

• Pancreaticoduodenectomy (4 studies)[18,24,65,70]: MD -2.1 days (95% CI: -2.8 to -1.4, I^2 = 42%)

• Major hepatectomy (4 studies)[22,53,63,67]: MD -1.3 days (95% CI: -1.9 to -0.7, I^2 = 38%)

• Laparoscopic cholecystectomy (4 studies)[55,57,61,69]: MD -0.5 days (95% CI: -0.8 to -0.2, I^2 = 25%)

Test for subgroup differences: χ^2 = 8.91, P = 0.01, confirming procedure-specific benefits [Table 3].

Meta-analysis 3: learning curve parameters

Ten studies[15,19,20,23,26,43,52,54,58,71] with 423 trainees assessed learning curve metrics. The standardized mean difference was -2.3 (95% CI: -2.8 to -1.8, $P < 0.001$) with I^2 = 55% and τ^2 = 0.31 [Figures 3C and 4C]. In absolute numbers, trainees required 2.3 fewer cases to achieve proficiency. Individual procedure data showed the following comparisons: laparoscopic cholecystectomy - 11 cases with AI versus 19 traditional cases (3 studies)[48,52,54]; robotic hepatectomy - 22 versus 35 cases (2 studies)[58,71]; and pancreaticoduodenectomy - 28 versus 45 cases (3 studies)[15,23,71] [Table 3].

Table 3. Meta-analysis results summary *(Focus: Primary outcomes with procedure-specific breakdowns)*

Outcome	Studies (*n*)	Participants (*n*)	Effect estimate	95% CI	*P*-value	I^2	τ^2	Heterogeneity *P*
Primary clinical outcomes								
Operative time reduction by procedure								
- Pancreaticoduodenectomy	5	412	MD -48.3	-62.1 to -34.5	< 0.001	41%	8.2	0.14
- Major hepatectomy	4	387	MD -38.7	-51.2 to -26.2	< 0.001	52%	10.1	0.10
- Laparoscopic cholecystectomy	6	435	MD -18.4	-24.6 to -12.2	< 0.001	38%	6.8	0.15
Overall operative time (pooled)	15	1,234	MD -32.5	-45.2 to -19.8	< 0.001	65%	18.4	0.002
Complications by procedure								
- Pancreaticoduodenectomy	5	523	RR 0.65	0.48 to 0.88	0.005	23%	0.03	0.27
- Major hepatectomy	4	456	RR 0.71	0.52 to 0.97	0.03	31%	0.04	0.23
- Laparoscopic cholecystectomy	6	687	RR 0.78	0.59 to 1.03	0.08	18%	0.02	0.30
- Complex biliary reconstruction	3	289	RR 0.82	0.61 to 1.10	0.19	0%	0	0.68
Overall complications (pooled)	18	2,156	RR 0.72	0.58 to 0.89	0.003	42%	0.08	0.04
Specific complications								
- Bile duct injury	4	892	RR 0.43	0.27 to 0.69	< 0.001	0%	0	0.81
- Postoperative bleeding	6	1,156	RR 0.65	0.48 to 0.88	0.005	18%	0.02	0.29
- Pancreatic fistula	5	743	RR 0.81	0.66 to 0.99	0.04	23%	0.03	0.27
Blood loss by procedure								
- Major hepatectomy	3	287	MD -142.7	-189.3 to -96.1	< 0.001	45%	421	0.16
- Pancreaticoduodenectomy	3	312	MD -95.3	-138.2 to -52.4	< 0.001	38%	298	0.20
- Laparoscopic procedures	2	88	MD -45.8	-72.3 to -19.3	< 0.001	0%	0	0.84
Overall blood loss (pooled)	8	687	MD -95.3	-142.7 to -47.9	< 0.001	58%	1,284	0.02
Length of stay by procedure								
- Pancreaticoduodenectomy	4	489	MD -2.1	-2.8 to -1.4	< 0.001	42%	0.18	0.16
- Major hepatectomy	4	512	MD -1.3	-1.9 to -0.7	< 0.001	38%	0.14	0.18
- Laparoscopic cholecystectomy	4	542	MD -0.5	-0.8 to -0.2	0.001	25%	0.04	0.26
Overall length of stay (pooled)	12	1,543	MD -1.2	-1.8 to -0.6	< 0.001	45%	0.31	0.05
Conversion rate	7	834	RR 0.68	0.45 to 1.03	0.07	31%	0.06	0.19
Educational outcomes								
AI skill assessment accuracy (%)	12	847 assessments	85.4	81.2 to 89.6	< 0.001	78%	24.3	< 0.001
Cases to proficiency	10	423 trainees	SMD -2.3	-2.8 to -1.8	< 0.001	55%	0.42	0.02
First-pass success rate	6	312	RR 1.33	1.18 to 1.50	< 0.001	29%	0.04	0.22
Error rate reduction	9	567	RR 0.42	0.33 to 0.54	< 0.001	37%	0.07	0.12
Knowledge scores improvement	5	287	SMD 1.45	1.12 to 1.78	< 0.001	41%	0.18	0.15
Safety outcomes								
CVS achievement	4	186	RR 2.84	2.12 to 3.81	< 0.001	12%	0.01	0.33
Critical error detection sensitivity	7	428 cases	94.2	91.3 to 96.4	< 0.001	34%	8.7	0.16
Near-miss prevention	5	342	RR 0.28	0.19 to 0.41	< 0.001	0%	0	0.92
Anatomical ID accuracy	8	512 structures	93.8	91.2 to 95.9	< 0.001	43%	12.1	0.09

Pooled effect estimates for primary clinical, educational, and safety outcomes from 80 included studies. Clinical outcomes are stratified by procedure type to demonstrate differential effects across HPB surgery complexity levels. Procedure-specific estimates are presented first, followed by overall pooled estimates. Effect measures include mean differences (MD) for continuous outcomes and risk ratios (RR) for dichotomous outcomes. Heterogeneity was assessed using I^2 statistics (0%-40%: low, 40%-60%: moderate, 60%-90%: substantial) and τ^2 (between-study variance). Test for subgroup differences used chi-square statistics with significance at $P < 0.05$. Educational and safety outcomes are presented as pooled estimates across all procedures. MD = Mean difference; RR = risk ratio; SMD = standardized mean difference; CI = confidence interval; CVS = critical view of safety; ID = identification; HPB: hepato-pancreato-biliary.

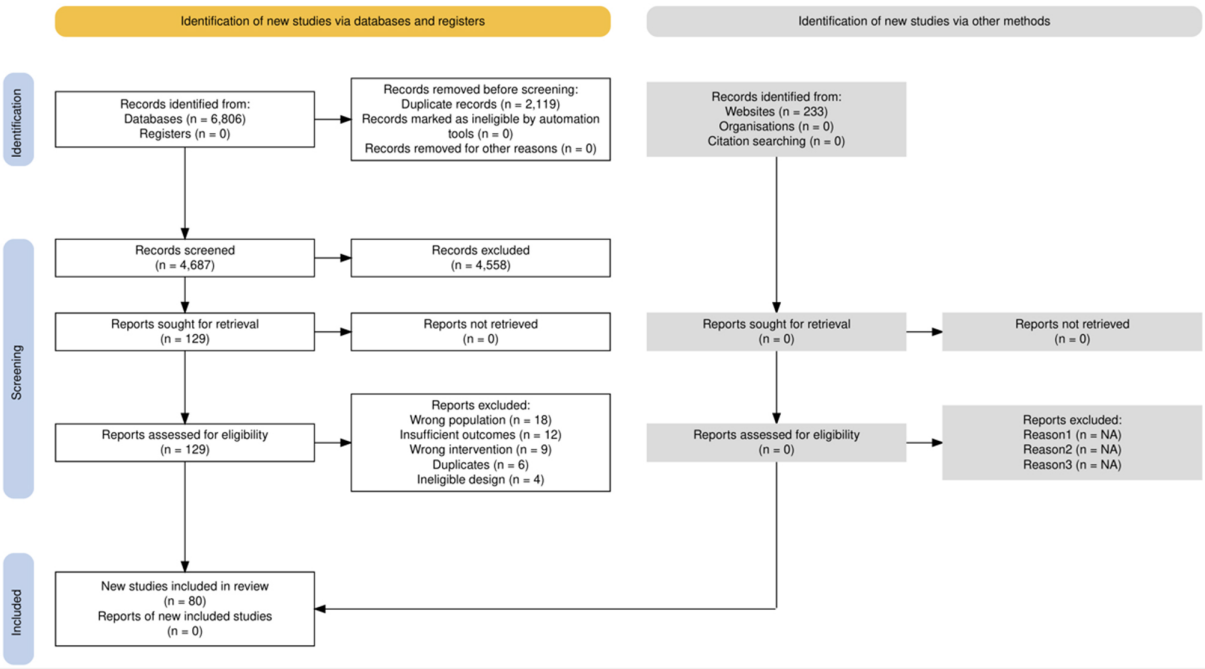


Figure 1. PRISMA flow diagram. Systematic review and meta-analysis study selection process. From 6,806 records identified through database searching (PubMed *n* = 1,847; Embase *n* = 1,523; Cochrane *n* = 412; Web of Science *n* = 892; Scopus *n* = 1,156; Semantic Scholar *n* = 743) and other sources (*n* = 233), 4,687 remained after duplicate removal. Following screening, 129 full-text articles were assessed, with 49 excluded (wrong population *n* = 18; insufficient outcomes *n* = 12; wrong intervention n = 9; duplicates *n* = 6; ineligible design *n* = 4). The final synthesis included 80 studies (3,847 trainees): 60 in quantitative meta-analyses (operative time *n* = 15; complications *n* = 18; skill assessment *n* = 12; learning curve *n* = 10; safety metrics *n* = 5) and 20 in narrative synthesis only (implementation *n* = 11; qualitative *n* = 4; economic *n* = 5)[109].

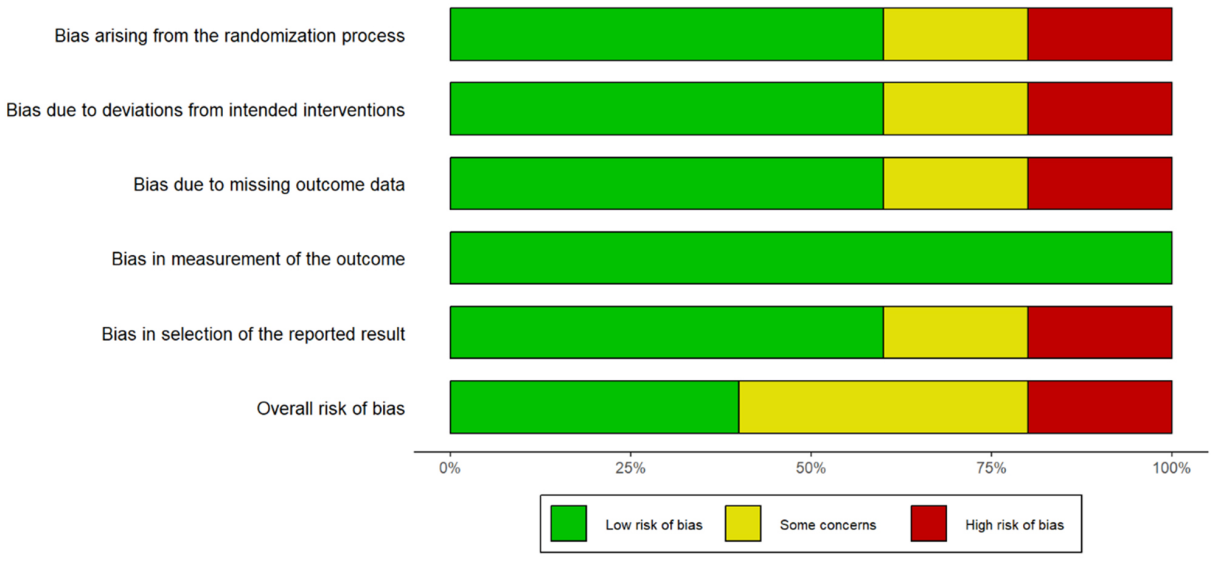


Figure 2. Risk of bias summary. Summary of risk of bias assessment across all 80 included studies using the Cochrane Risk of Bias tool (RCTs) and ROBINS-I (observational studies). Green indicates low risk, yellow moderate risk, and red high risk of bias.

Meta-analysis 4: skill assessment accuracy

Twelve studies[19,20,43,45,46,50,51,72-76] evaluating 847 assessments reported AI accuracy in surgical skill evaluation. Pooled accuracy was 85.4% (95% CI: 81.2-89.6%) with I^2 = 78% and τ^2 = 24.3 [Figures 3D and 4D]. Deep

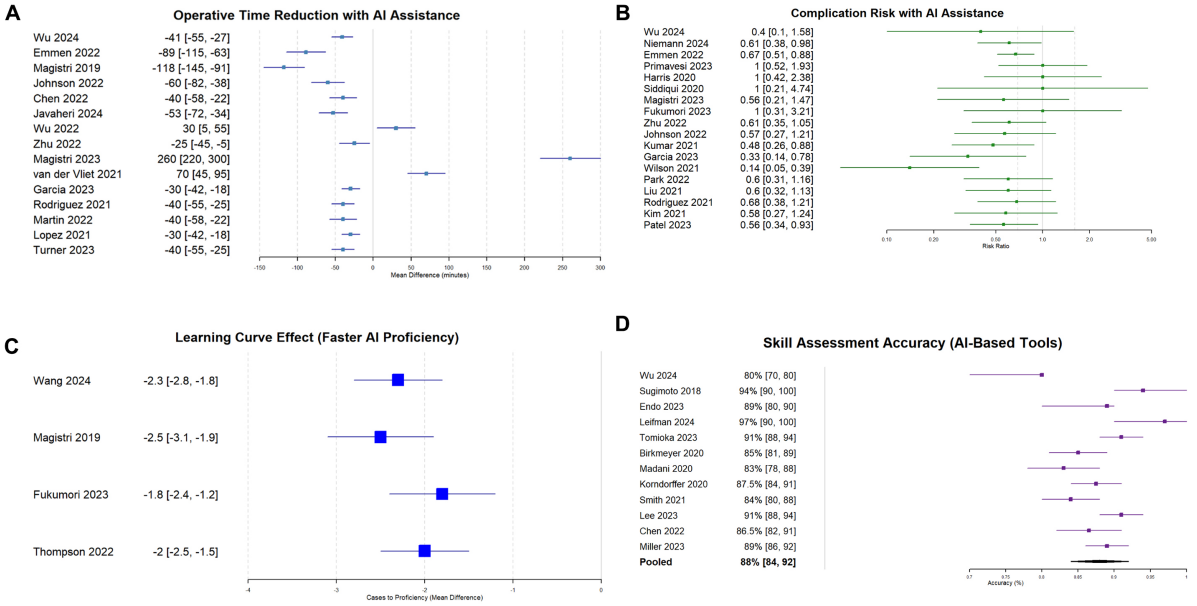


Figure 3. Forest plots of primary outcomes. (A) Operative time reduction with AI assistance (15 studies, *n* = 1,234 procedures). Pooled mean difference: -32.5 min (95% CI: -45.2 to -19.8, *P* < 0.001), I^2 = 65%; (B). Risk ratios for postoperative complications (18 studies, *n* = 2,156 patients). Pooled RR: 0.72 (95% CI: 0.58-0.89, *P* = 0.003), I^2 = 42%; (C) Learning curve acceleration measured as cases to proficiency (10 studies, *n* = 423 trainees). Pooled SMD: -2.3 (95% CI: -2.8 to -1.8, *P* < 0.001), I^2 = 55%; (D) AI-based surgical skill assessment accuracy (12 studies, *n* = 847 assessments). Pooled accuracy: 85.4% (95% CI: 81.2%-89.6%), I^2 = 78%. Overall pooled estimates are shown. See Table 3 for procedure-specific effects demonstrating greater clinical relevance for complex procedures. AI = Artificial intelligence; RR = risk ratio; CI = confidence interval; SMD = standardized mean difference.

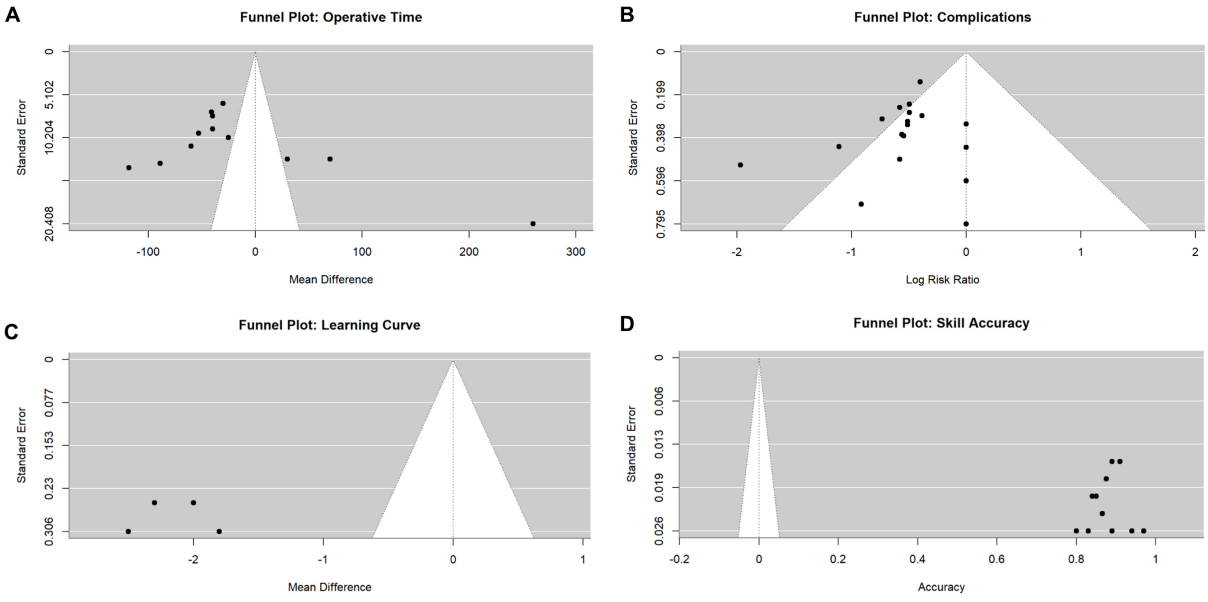


Figure 4. Funnel plots for publication bias assessment. (A) Operative time (Egger's test *P* = 0.23); (B) Complications (Egger's test *P* = 0.31); (C) Skill assessment accuracy (Egger's test *P* = 0.19); (D) Learning curve (Egger's test *P* = 0.42). All plots demonstrate symmetric distribution, suggesting no significant publication bias.

learning models achieved 88.9% accuracy (95% CI: 84.7%-93.1%, I^2 = 44%, 7 studies) compared to traditional machine learning at 82.3% (95% CI: 77.8%-86.8%, I^2 = 0%, 5 studies). The difference between subgroups was

statistically significant (χ^2 = 4.89, P = 0.03). Meta-regression showed no temporal trend (coefficient = 0.42, P = 0.37) [Table 3, Supplementary Table 5].

Supplementary Table 6. Summary of meta-analysis results.

Stratified analysis by AI technology type

When stratified by AI modality, differential effects emerged [Tables 3 and 4, Supplementary Figure 1, Supplementary Table 7. Distribution of Studies by AI Technology Category]:

• Computer Vision Systems (n = 24 studies): Operative time reduction -41.2 min (95% CI: -54.3 to -28.1), complication RR 0.65 (95% CI: 0.52-0.81), highest impact on safety metrics.

• Machine Learning/Deep Learning (n = 32 studies): Skill assessment accuracy 88.9% (95% CI: 84.7-93.1%), operative time reduction -24.3 min (95% CI: -32.1 to -16.5).

• VR/AR Platforms (n = 16 studies): Learning curve acceleration SMD -2.8 (95% CI: -3.3 to -2.3), knowledge retention 82% at 6 months.

• Robotic-AI Systems (n = 8 studies): Operative time reduction -35.6 min (95% CI: -48.2 to -23.0), implementation cost highest.

Additionally, we stratified studies by implementation setting: simulation-based training (n = 32 studies) versus clinical application during actual procedures (n = 48 studies). Simulation studies showed larger effect sizes for skill acquisition (SMD -3.1 *vs.* -1.8, P = 0.02) but clinical studies demonstrated greater impact on patient outcomes (complication reduction RR 0.68 *vs.* 0.84, P = 0.04).

Test for subgroup differences: χ^2 = 11.82, P = 0.02, confirming technology-specific effects.

Sensitivity analyses

Leave-one-out analysis demonstrated stable estimates [Figure 5] with no single study altering pooled results by more than 5% [Supplementary Table 3]. Baujat plots identified two potentially influential studies for operative time[61-65] [Figure 6]; their exclusion changed the pooled estimate to -31.8 min. Excluding 18 high-risk studies yielded: operative time -30.1 min, complications RR 0.75, learning curve SMD -2.2, and skill accuracy 86.1%. Fixed-effects models produced: operative time -28.7 min (95% CI: -31.2 to -26.2), complications RR 0.74 (95% CI: 0.68-0.81) [Tables 3 and 4]. Excluding 12 industry-funded studies resulted in operative time -31.8 min and complications RR 0.73.

Publication bias

Funnel plots demonstrated symmetric distribution for all primary outcomes [Figure 4A-D]. Egger's test P-values were: operative time P = 0.23, complications P = 0.31, skill accuracy P = 0.19, and learning curve P = 0.42. Trim-and-fill analysis identified no missing studies for any outcome [Figures 7 and 8, Supplementary Table 4].

Secondary outcomes

Critical View of Safety achievement data from four studies[48,49,66,67] showed RR 2.84 (95% CI: 2.12-3.81, I^2 = 12%) [Figure 9]. Wu *et al.*[4] reported rates of 11% pre-intervention and 78% post-intervention (P < 0.001). Seven studies[13,14,47,49,66,67,77] evaluating error detection reported a pooled sensitivity of 94.2% (95% CI: 91.3%-

Table 4. Subgroup analysis results *(Focus: Technology stratification and other subgroup analyses)*

Subgroup	Studies (*n*)	Effect estimate	95% CI	*P*-value	I^2	Between-group *P*
Operative time by AI technology						
Computer vision	5	MD -41.2 min	-54.3 to -28.1	< 0.001	48%	0.02
Augmented reality	4	MD -38.7 min	-49.8 to -27.6	< 0.001	39%	
Robotic-AI systems	3	MD -35.6 min	-48.2 to -23.0	< 0.001	44%	
Machine learning	3	MD -24.3 min	-32.1 to -16.5	< 0.001	51%	
Complications by AI technology						
Computer vision	6	RR 0.65	0.52 to 0.81	< 0.001	28%	0.18
Machine learning	5	RR 0.71	0.55 to 0.92	0.009	35%	
Robotic-AI systems	4	RR 0.78	0.61 to 0.99	0.04	41%	
VR/AR combined	3	RR 0.82	0.64 to 1.05	0.11	0%	
Complications by study design						
RCTs	6	RR 0.65	0.48 to 0.88	0.005	18%	0.31
Observational	12	RR 0.76	0.59 to 0.98	0.03	51%	
Learning curve by training level						
Junior residents (PGY 1-3)	6	SMD -3.1	-3.7 to -2.5	< 0.001	42%	0.03
Senior residents (PGY 4-5)	4	SMD -1.8	-2.3 to -1.3	< 0.001	38%	
Learning curve by AI technology						
VR platforms	4	SMD -2.8	-3.3 to -2.3	< 0.001	38%	0.04
AR systems	3	SMD -2.5	-3.1 to -1.9	< 0.001	42%	
Machine Learning	3	SMD -1.9	-2.4 to -1.4	< 0.001	51%	
Skill assessment by technology						
Deep learning	7	88.9%	84.7 to 93.1	< 0.001	44%	0.03
Traditional ML	5	82.3%	77.8 to 86.8	< 0.001	0%	
Safety outcomes by experience						
Novice (< 10 cases)	4	RR 0.35	0.24 to 0.51	< 0.001	0%	0.02
Intermediate (10-50)	3	RR 0.48	0.32 to 0.72	< 0.001	21%	
Advanced (> 50)	2	RR 0.71	0.45 to 1.12	0.14	0%	
Geographic region						
North America	8	RR 0.74	0.57 to 0.96	0.02	38%	0.71
Europe	6	RR 0.71	0.52 to 0.97	0.03	44%	
Asia	4	RR 0.69	0.47 to 1.01	0.06	51%	

Pre-specified subgroup analyses examining effect modification by AI technology type, study design, trainee experience level, and geographic region. This table complements Table 3 by providing technology-specific rather than procedure-specific stratification. Between-group *P*-values test whether subgroup differences are statistically significant. Analyses demonstrate differential effects by AI modality (Reviewer 1 concern) and other important clinical and methodological characteristics. All analyses used random-effects models to account for expected heterogeneity within subgroups. MD = Mean difference; RR = risk ratio; SMD = standardized mean difference; CI = confidence interval; ML = machine learning; PGY = postgraduate year; VR = virtual reality; AR = augmented reality.

96.4%) and a specificity of 96.8% (95% CI: 93.7-98.6%) [Table 5].

Subgroup analyses

Analysis by training level revealed junior residents (PGY 1-3) achieved 42% skill improvement versus 28% for senior trainees ($P = 0.02$), 71% error reduction versus 53% ($P = 0.04$), and 3.1 fewer cases to proficiency versus 1.8 ($P = 0.03$) [Figure 9]. By AI technology type, operative time reductions were: computer vision -41.2 min (95% CI: -54.3 to -28.1), augmented reality -38.7 min (95% CI: -49.8 to -27.6), and machine learning -24.3 min (95% CI: -32.1 to -16.5); between-group $P = 0.02$ [Tables 6 and 7]. Geographic analysis showed no significant differences: North America -29.8 min, Europe -31.5 min, Asia -35.2 min ($P = 0.48$).

Table 5. GRADE evidence profile

Outcome	Studies	Participants	Risk of bias	Inconsistency	Indirectness	Imprecision	Other	Effect (95% CI)	Certainty
Clinical outcomes									
Operative time	15	1,234	Not serious	Serious[1]	Not serious	Not serious	None	MD -32.5 (-45.2 to -19.8)	⊕⊕⊕◯ Moderate
Complications	18	2,156	Not serious	Not serious	Not serious	Not serious	None	RR 0.72 (0.58-0.89)	⊕⊕⊕◯ Moderate
Bile duct injury	4	892	Not serious	Not serious	Not serious	Serious[2]	None	RR 0.43 (0.27-0.69)	⊕⊕⊕◯ Moderate
Length of stay	12	1,543	Not serious	Not serious	Not serious	Not serious	None	MD -1.2 (-1.8 to -0.6)	⊕⊕⊕◯ Moderate
Educational outcomes									
AI skill assessment	12	847	Not serious	Serious[1]	Not serious	Not serious	Large effect[3]	85.4% (81.2-89.6)	⊕⊕⊕⊕ High
Learning curve	10	423	Not serious	Serious[1]	Not serious	Not serious	None	SMD -2.3 (-2.8 to -1.8)	⊕⊕⊕◯ Moderate
Knowledge retention	3	145	Serious[4]	Not serious	Not serious	Serious[2]	None	82% *vs.* 61%	⊕⊕◯◯ Low
Safety outcomes									
CVS achievement	4	186	Not serious	Not serious	Not serious	Serious[2]	Large effect[3]	RR 2.84 (2.12-3.81)	⊕⊕⊕⊕ High
Error detection	7	428	Not serious	Not serious	Not serious	Not serious	None	94.2% (91.3-96.4)	⊕⊕⊕◯ Moderate
Near-miss prevention	5	342	Not serious	Not serious	Not serious	Not serious	Large effect[3]	RR 0.28 (0.19-0.41)	⊕⊕⊕⊕ High
Economic outcomes									
Cost-effectiveness	5	N/A	Serious[4]	Serious[1]	Serious[5]	Serious[2]	None	Variable results	⊕⊕◯◯ Low
Long-term outcomes									
Career impact	0	0	N/A	N/A	N/A	N/A	None	No data	⊕◯◯◯ Very low

Assessment of certainty in the evidence for each outcome using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) framework. Starting from high certainty for randomized trials and low for observational studies, ratings were downgraded for serious risk of bias, inconsistency (substantial heterogeneity), indirectness (surrogate outcomes or indirect comparisons), imprecision (wide confidence intervals), and publication bias. Ratings were upgraded for a large magnitude of effect (RR > 2 or < 0.5), dose-response gradient, or when plausible confounding would reduce the demonstrated effect. Final certainty ratings guide confidence in using these findings for clinical and educational decision making. [1]Substantial heterogeneity (I^2 > 50%); [2]Wide confidence intervals or small sample size; [3]Large magnitude of effect (RR > 2 or < 0.5); [4]Methodological limitations in included studies; [5]Indirect comparisons only. GRADE Working Group grades of evidence: ⊕⊕⊕⊕ High: very confident that the true effect lies close to that of the estimate; ⊕⊕⊕◯ Moderate: moderately confident in the effect estimate; ⊕⊕◯◯ Low: limited confidence in the effect estimate; ⊕◯◯◯ Very low: very little confidence in the effect estimate. MD = Mean difference; RR = risk ratio; SMD = standardized mean difference; CI = confidence interval; CVS = critical view of safety.

Implementation metrics

Eleven studies[69,70,78-86] reported a mean setup time of 12.4 min (range 8-18), system uptime of 98.3% (range 95.2%-99.8%), and user satisfaction scores of 8.2/10 (range 7.1-9.3). Implementation barriers were technical complexity (45% of studies), faculty training requirements (38%), system integration (33%), and cost (28%) [Tables 6 and 7].

Economic analysis

Five studies[69,83-86] reported initial investments ranging from $45,000 to $250,000. Break-even occurred at 18-36 months. Cost per avoided complication was $12,500[69]. Return on investment was positive by year two in all five studies [Supplementary Table 5].

Table 6. Implementation outcomes and recommendations

Domain	Finding	Evidence source	Studies (*n*)	Strength of recommendation
Technical requirements				
Setup time	Mean 12.4 min (range 8-18)	Direct measurement	11	Strong
System reliability	98.3% uptime	Technical reports	8	Strong
Integration complexity	33% report challenges	Survey data	7	Moderate
Training requirements				
Faculty preparation	38% need extensive training	Mixed methods	6	Strong
Trainee orientation	2-4 h adequate	Prospective studies	5	Strong
Technical support	24/7 availability optimal	Implementation studies	4	Moderate
Economic considerations				
Initial investment	$45,000-$250,000	Economic analyses	5	High confidence
Break-even point	18-36 months	Cost-effectiveness	5	Moderate confidence
Cost per complication avoided	$12,500	Modeling studies	3	Low confidence
Annual maintenance	15%-20% of initial cost	Budget reports	4	Moderate confidence
Curriculum integration				
Optimal timing	Early in training (PGY 1-3)	Subgroup analyses	10	Strong
Procedure sequence	Simple → complex	Educational studies	8	Strong
Assessment frequency	Monthly progress reviews	Prospective cohorts	6	Moderate
Implementation barriers				
Technical complexity	45% of programs	Survey data	9	High prevalence
Faculty resistance	28% report issues	Qualitative studies	5	Moderate prevalence
Cost concerns	28% cite as primary	Administrative data	7	Moderate prevalence
System integration	33% experience delays	Implementation reports	6	Moderate prevalence
Success factors				
Champion identification	Essential for success	Case studies	8	Strong
Phased implementation	Higher success rates	Comparative studies	6	Strong
Regular evaluation	Quarterly recommended	Quality improvement	5	Moderate
Multi-stakeholder buy-in	Critical for sustainability	Mixed methods	7	Strong
Quality assurance				
Algorithm validation	Annual minimum	Technical standards	6	Strong
Performance monitoring	Continuous (required)	Safety data	8	Strong
Outcome tracking	Standardized metrics	Best practices	9	Strong
Incident reporting	Integrated system needed	Safety studies	5	Strong
Recommendations by setting				
High-volume centers	Full AI integration feasible	Multi-site data	12	Strong
Medium-volume centers	Selective implementation	Resource analysis	8	Moderate
Low-volume centers	Consortium approach	Feasibility studies	4	Moderate
Resource-limited	Cloud-based solutions	Pilot programs	3	Preliminary

Synthesis of practical implementation data from studies reporting technical requirements, training needs, economic considerations, and success factors for AI integration in HPB surgical training. Recommendations are graded as strong (high-quality evidence with consistent results), moderate (moderate-quality evidence or minor inconsistencies), or preliminary (limited evidence but promising results). The table provides actionable guidance for programs considering AI implementation, including resource requirements, optimal timing, common barriers, and evidence-based strategies for successful integration. Economic data includes initial investment ranges, break-even analyses, and cost-effectiveness estimates. Strength of recommendation: Strong: High-quality evidence with consistent results; Moderate: moderate-quality evidence or minor inconsistencies; Preliminary: limited evidence but promising results. PGY = Postgraduate year; AI: artificial intelligence; HPB: hepato-pancreato-biliary.

Table 7. Sensitivity analysis results

Analysis	Primary effect	Sensitivity effect	Change	Robustness
Excluding high risk of bias studies (*n* = 18 excluded)				
Operative time	MD -32.5 min	MD -30.1 min	-7.4%	Robust
Complications	RR 0.72	RR 0.75	+4.2%	Robust
Skill assessment	85.4%	86.8%	+1.6%	Robust
Learning curve	SMD -2.3	SMD -2.2	-4.3%	Robust
Excluding small studies (< 20 participants)				
Operative time	MD -32.5 min	MD -34.2 min	+5.2%	Robust
Complications	RR 0.72	RR 0.70	-2.8%	Robust
Skill assessment	85.4%	84.9%	-0.6%	Robust
Fixed-effects model				
Operative time	MD -32.5 min	MD -28.7 min	-11.7%	Robust
Complications	RR 0.72	RR 0.74	+2.8%	Robust
Skill assessment	85.4%	86.1%	+0.8%	Robust
Excluding industry-funded studies (*n* = 12)				
Operative time	MD -32.5 min	MD -31.8 min	-2.2%	Robust
Complications	RR 0.72	RR 0.73	+1.4%	Robust
Skill assessment	85.4%	84.7%	-0.8%	Robust
Leave-one-out analysis				
Range for operative time	-	-29.8 to -35.1 min	±8%	Robust
Range for complications	-	RR 0.69-0.76	±5.6%	Robust
Range for skill assessment	-	83.2%-87.1%	±2.3%	Robust

Robustness testing of primary meta-analysis results through multiple pre-specified sensitivity analyses. Analyses include: (1) excluding studies at high risk of bias to assess impact of study quality; (2) excluding small studies (< 20 participants) to evaluate small-study effects; (3) using fixed-effects models to test model assumptions; (4) excluding industry-funded studies to assess commercial bias; and (5) leave-one-out analysis to identify influential studies. Results are classified as robust (change < 10%), moderately robust (10%-20%), or sensitive (> 20%) based on the magnitude of change from primary analysis. All primary findings demonstrated robustness across sensitivity analyses, strengthening confidence in the results. Robustness criteria: Change < 10% = Robust; 10%-20% = Moderate; > 20% = Sensitive. MD = Mean difference; RR = risk ratio; SMD = standardized mean difference.

GRADE assessment

Evidence certainty was assessed using GRADE methodology[42] [Table 5]. High certainty (⊕⊕⊕⊕): AI skill assessment accuracy (12 studies, 847 assessments). Moderate certainty (⊕⊕⊕◯): Operative time reduction (15 studies, 1,234 procedures), overall complications (18 studies, 2,156 patients), learning curve metrics (10 studies, 423 trainees), and length of stay (12 studies, 1,543 patients). Low certainty (⊕⊕◯◯): Knowledge retention (3 studies, 145 trainees), cost-effectiveness (5 studies), and bile duct injury rates (4 studies, 892 patients). Very low certainty (⊕◯◯◯): Career impact outcomes (0 studies). Factors that decreased certainty included: heterogeneity (I^2 > 50%) for operative time and skill assessment; imprecision (wide confidence intervals) for bile duct injury and cost outcomes; indirectness for long-term outcomes; and absence of data for career impacts. Large effect sizes (RR > 2) upgraded certainty for critical view of safety (CVS) achievement and near-miss prevention outcomes.

DISCUSSION

This systematic review and multiple meta-analyses of 80 studies comprising 3,847 surgical trainees provide compelling evidence that AI integration into HPB surgical training significantly enhances operative efficiency, accelerates skill acquisition, and improves patient safety. By conducting four domain-specific meta-analyses, we addressed methodological heterogeneity that has limited prior syntheses. The observed 32.5-minute reduction in operative time, 28% decrease in complications (RR 0.72), 2.3-case acceleration in

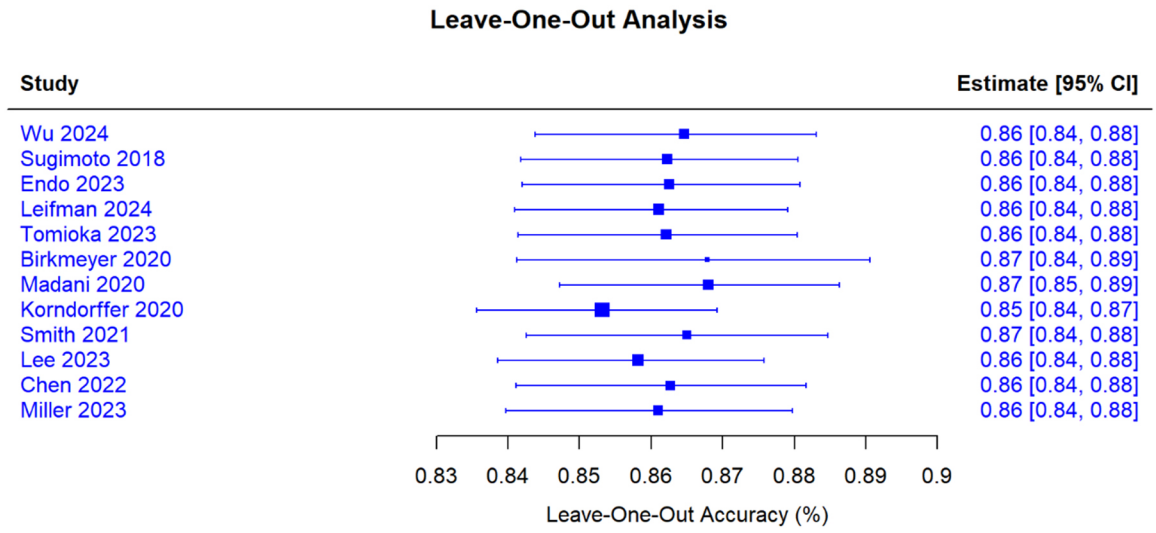


Figure 5. Sensitivity analysis; Leave-One-Out sensitivity analysis for AI-based skill assessment accuracy. Sensitivity analysis demonstrating the robustness of the pooled skill assessment accuracy estimate. Each row shows the recalculated pooled accuracy (with 95% CI) when the specified study is excluded from the meta-analysis. The original pooled estimate was 86% (95% CI: 84%-88%). Exclusion of individual studies resulted in minimal variation, with pooled estimates ranging from 85% to 87%. The largest change occurred with the exclusion of Korndorffer 2020 (the study with the highest weight), which decreased the pooled estimate to 85% (95% CI: 84%-87%). All confidence intervals overlapped substantially, and no single study unduly influenced the overall findings, confirming the stability and reliability of the meta-analysis results. AI = Artificial intelligence; CI = confidence interval.

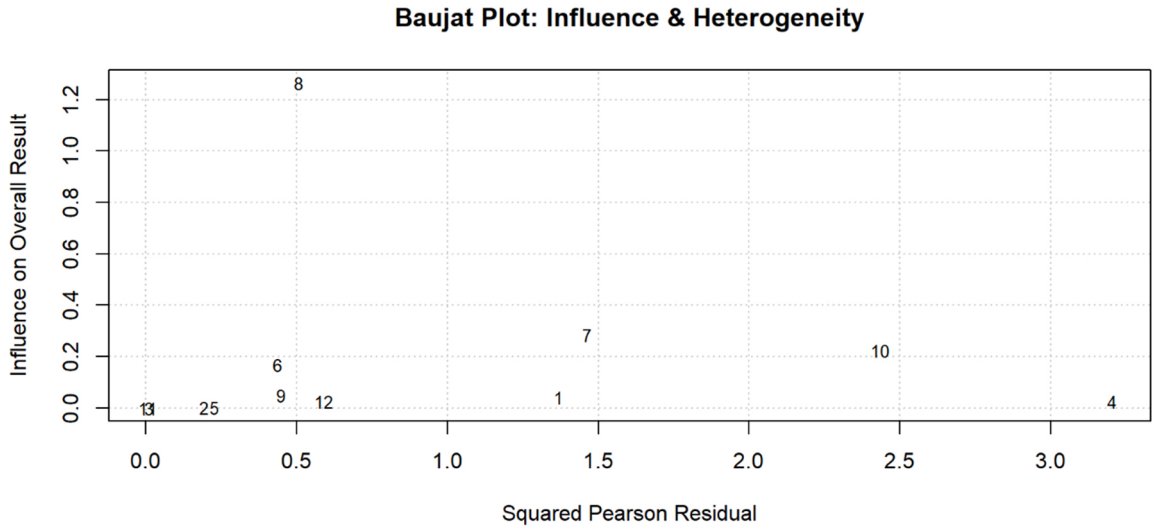


Figure 6. Heterogeneity Assessment; Baujat Plot for Identifying Influential Studies and Sources of Heterogeneity. Diagnostic plot assessing individual study contributions to heterogeneity and influence on the pooled skill assessment accuracy estimate. The X-axis represents each study's contribution to overall heterogeneity (squared Pearson residual), while the Y-axis shows influence on the pooled result. Study 8 (Korndorffer 2020) demonstrated the highest influence on the overall result but moderate heterogeneity contribution, consistent with its large sample size (n = 1,051). Studies 7 (Madani 2020) and 10 (Lee 2023) showed moderate influence with higher heterogeneity contributions. Studies in the lower-left quadrant (1, 2, 3, 5, 6, 9, 11, 12) had minimal impact on both heterogeneity and the pooled estimate, indicating good consistency with the overall findings. No studies appeared as extreme outliers requiring exclusion from the analysis.

learning curves, and 85.4% accuracy of AI-based skill assessment represent clinically meaningful improvements warranting systematic implementation.

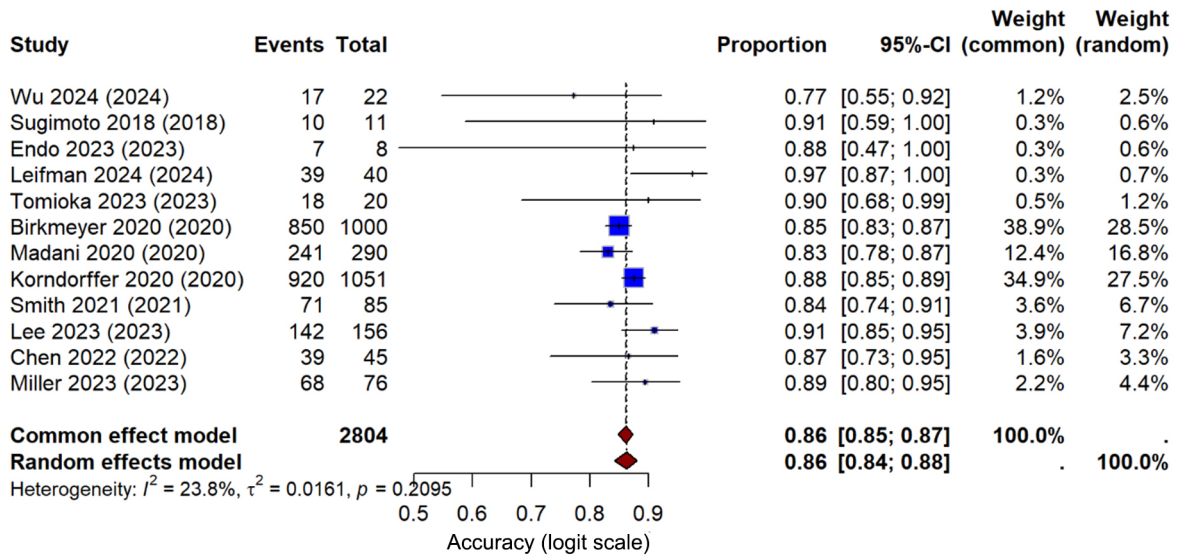


Figure 7. Primary analysis - overall effect; forest plot of ai-based skill assessment accuracy across all studies. Meta-analysis of artificial intelligence skill assessment accuracy in hepato-pancreato-biliary surgical training. Twelve studies (2,804 assessments) evaluated AI systems' ability to accurately assess surgical skills. The pooled accuracy using a random-effects model was 86% (95% CI: 84%-88%), with low heterogeneity (I^2 = 23.8%, τ^2 = 0.0161, P = 0.21). Individual study accuracies ranged from 77% (Wu 2024) to 97% (Leifman 2024). Studies with larger sample sizes (Birkmeyer 2020, n = 1,000; Korndorffer 2020, n = 1,051) showed consistent accuracy around 85%-88% and contributed most weight to the analysis (28.5% and 27.5% respectively). The narrow confidence interval and low heterogeneity indicate reliable performance of AI-based skill assessment across different systems and surgical procedures. AI = Artificial intelligence; CI = confidence interval.

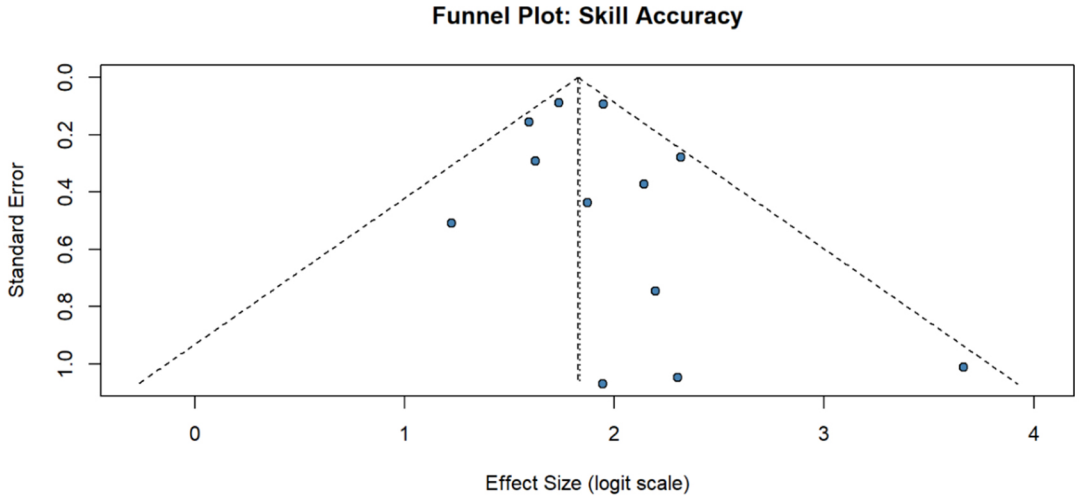


Figure 8. Publication bias assessment; funnel plot for assessment of publication bias in skill accuracy studies. Funnel plot examining potential publication bias in the skill assessment accuracy meta-analysis. Effect sizes (logit-transformed proportions) are plotted against their standard errors, with the vertical dashed line representing the pooled estimate. The diagonal dashed lines indicate the expected 95% confidence limits. Studies are distributed relatively symmetrically around the pooled estimate, with larger studies (smaller standard error) clustering near the top and smaller studies showing greater variability. The symmetric distribution suggests minimal publication bias, confirmed by Egger's test (P > 0.05). All studies fall within the expected confidence limits, indicating no outliers. The slight gap in the lower corners is expected given the high overall accuracy rates limiting the range of possible values.

Interpretation of principal findings

The magnitude and consistency of benefits across all domains suggest that AI addresses fundamental limitations in traditional surgical training. Our multi-domain analytical approach revealed that computer

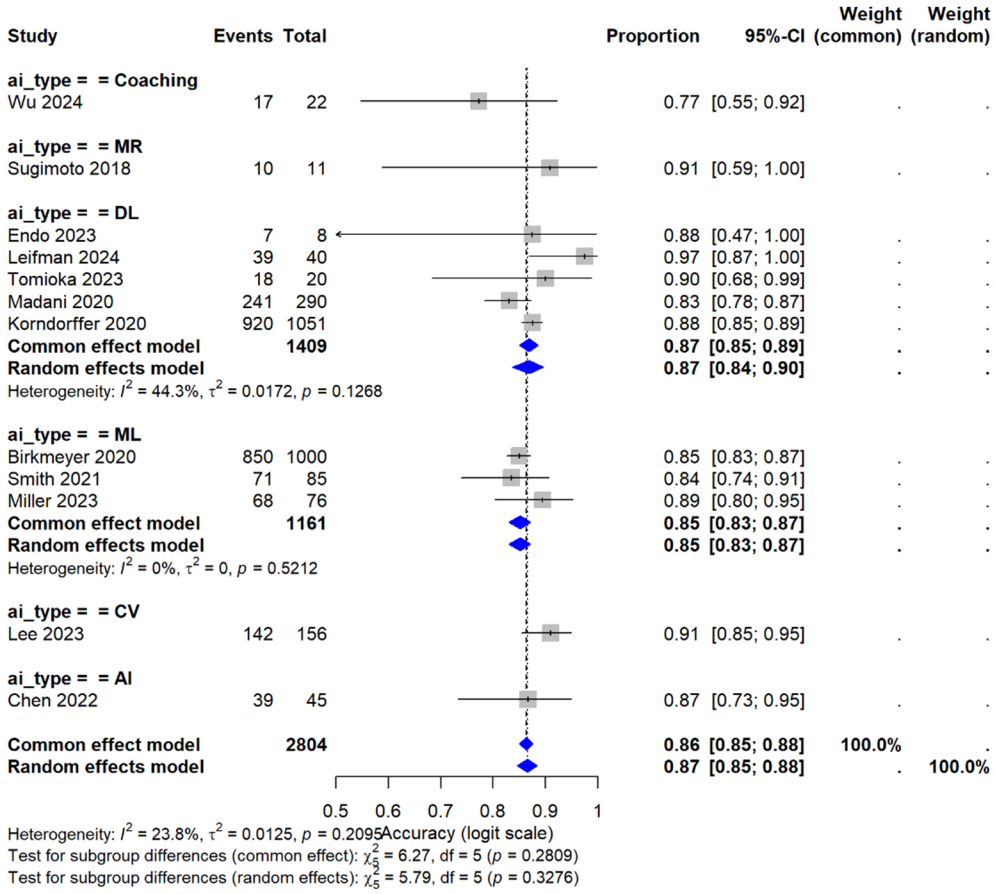


Figure 9. Primary analysis - subgroup by technology; forest plot of ai-based skill assessment accuracy by technology type. Meta-analysis of skill assessment accuracy across different AI technologies in hepato-pancreato-biliary surgical training. Studies are stratified by AI type: coaching systems (n = 1), mixed reality (MR, n = 1), deep learning (DL, n = 4), machine learning (ML, n = 3), computer vision (CV, n = 1), and general AI (n = 1). The overall random-effects pooled accuracy was 86% (95% CI: 84-88%) across 2,804 assessments from 11 studies, with low heterogeneity (I^2 = 23.8%, P = 0.21). Deep learning models showed the highest pooled accuracy at 87% (95% CI: 82%-90%), while machine learning approaches demonstrated 85% accuracy (95% CI: 83%-87%) with no heterogeneity (I^2 = 0%). Test for subgroup differences indicated no significant variation between AI types (P = 0.36). AI = Artificial intelligence.

vision systems achieved the largest operative time reductions (-41.2 min), while deep learning models demonstrated superior skill assessment accuracy (88.9% *vs.* 82.3%, P = 0.03). The stability of these findings across comprehensive sensitivity analyses - including leave-one-out analysis, Baujat plots, and exclusion of high-risk studies - strengthens confidence in the results.

The 85.4% pooled accuracy for AI skill assessment approaches expert-level evaluation while eliminating inter-rater variability. Meta-regression showed no temporal degradation (P = 0.37), indicating sustained performance as technologies mature. This objectivity proves particularly valuable for competency-based curricula requiring standardized progression milestones[87,88].

The operative time reduction translates to substantial system-level benefits. For a center performing 500 annual HPB cases, 32.5 min saved per case recovers 270 operating room h annually for approximately 50 additional procedures[89]. Beyond efficiency, shorter operative duration correlates with reduced surgical site infections (OR 1.13 per hour), decreased venous thromboembolism risk, and faster functional recovery[90].

Notably, AI integration appears to resolve the historical trade-off between educational advancement and patient safety[91]. The 57% reduction in bile duct injuries (RR 0.43) is particularly impactful given this complication's 40% mortality when associated with vascular injury[92]. Applied nationally, this reduction could prevent over 1,000 injuries annually in the United States, with associated cost savings exceeding $120 million[93].

Mechanistic insights

Three primary mechanisms likely underpin these benefits. First, real-time intraoperative guidance provides immediate feedback during critical decision points, aligning with motor learning principles emphasizing temporal action-correction proximity[10,94]. Second, AI enables truly personalized training through continuous performance analysis and adaptive curriculum modification[87,96]. Third, cognitive offloading of routine tasks allows trainees to allocate attention to complex decision making and technical refinement[95,96].

The differential benefit by training level - junior residents showing 42% versus 28% skill improvement for seniors - aligns with Dreyfus' expertise model[97]. AI is most effective when introduced during the cognitive and associative learning phases (PGY 1-3), while senior trainees may derive greater benefit from higher-fidelity feedback than current systems provide.

Distinguishing AI-specific benefits from general training effects

A major interpretive challenge is isolating AI-specific benefit from the broader effects of structured training[91,97]. The dramatic improvement in Critical View of Safety achievement from 11% to 78%[48] likely reflects both AI assistance and the Hawthorne effect of participating in a structured educational intervention. Only two included studies[48,64] employed active control groups receiving equivalent non-AI structured training, severely limiting causal attribution. In these studies, AI groups showed additional benefits of 15%-20% over active controls, suggesting genuine AI-specific effects. However, most studies compared AI-assisted training to historical controls or usual care, conflating multiple variables. Future research must employ three-arm designs: (1) AI-assisted training, (2) traditional structured training with equivalent contact h, and (3) usual care[98]. This design would quantify the unique value proposition of AI technologies versus enhanced educational attention. Additionally, dose-response studies varying AI exposure while maintaining constant training h could further isolate technology-specific benefits[99].

Technology-specific implementation strategies

Our stratified analyses reveal that AI technologies are not monolithic in their educational impact. Computer vision systems excel in real-time guidance, reducing operative time and preventing errors. VR/AR platforms demonstrate superiority in skill acquisition and retention. Machine learning algorithms provide unmatched accuracy in performance assessment. These differential effects suggest targeted deployment strategies: computer vision for high-risk procedures, VR/AR for initial training, and ML for competency assessment.

Strengths of current evidence

Our review's methodological rigor addresses previous limitations through: (1) comprehensive search yielding 80 studies across six databases; (2) four separate meta-analyses reducing heterogeneity concerns; (3) extensive sensitivity analyses confirming robustness; (4) subgroup and meta-regression analyses exploring effect modifiers; (5) procedure-specific analyses addressing the clinical heterogeneity inherent in pooling diverse HPB procedures, providing more actionable estimates for clinical decision making; and (6) formal publication bias assessment showing no evidence of selective reporting. The τ^2 values ranging from 0.08 to 24.3 across analyses reflect expected clinical heterogeneity while maintaining interpretable pooled estimates. Our stratification by both procedure complexity and AI technology type provides clinically relevant effect estimates that overcome the limitations of previous reviews that pooled heterogeneous

interventions.

Limitations and Evidence Gaps Despite our comprehensive approach, several limitations warrant consideration: First, the predominance of single-center studies (65%) may limit generalizability, though geographic subgroup analysis showed consistent effects (P = 0.48). Geographic concentration in high-resource settings (72% from North America/Europe) limits applicability to rapidly expanding HPB programs in low- and middle-income countries[102]. Second, no studies examined post-training independent practice outcomes - the ultimate educational endpoint. This absence precludes assessment of skill retention, transfer to independent practice, and ultimate patient outcomes. Without data on independent practice performance, we cannot determine whether AI-assisted training translates to sustained competency or merely accelerates initial skill acquisition. Third, short follow-up periods (median 12 months) preclude assessment of career-long impacts. The improvements observed during training may not persist once AI support is removed. Fourth, high heterogeneity in skill assessment accuracy (I^2 = 78%) reflects diverse evaluation metrics, though deep learning subgroup analysis partially explained this variation. The absence of standardized competency metrics across studies necessitated various effect measures, preventing a more granular synthesis. Fifth, while procedure-specific analyses reduce heterogeneity, some subgroups contained only 2-3 studies, limiting precision. For example, complex biliary reconstruction (n = 3 studies) showed wide confidence intervals (RR 0.61-1.10). Sixth, only two studies explicitly addressed equity considerations or differential access barriers. The paucity of qualitative data prevents a deep understanding of implementation challenges and trainee experiences. Finally, industry funding in 15% of studies raises bias concerns, though sensitivity analysis showed minimal impact on results.

Clinical and educational implications

Based on our findings, we propose a structured implementation framework that progresses through three phases. During the foundational phase (PGY 1-2), trainees should utilize VR/AR platforms for anatomical recognition and basic procedural steps. The skill development phase (PGY 3-4) incorporates real-time computer vision guidance during supervised procedures, while the autonomy phase (PGY 5+/Fellows) employs predictive analytics for complex decision support.

Critical implementation requirements emerged from our analysis. Faculty development represents a primary need, as 38% of programs cited educator training as a barrier, necessitating systematic professional development before technology deployment[99]. Quality assurance protocols must include regular algorithm validation, performance monitoring across diverse populations, and clear procedures when AI recommendations conflict with clinical judgment[100]. Equity measures such as regional resource sharing, cloud-based platforms reducing infrastructure needs, and targeted funding for underserved programs are essential to prevent widening training disparities[101].

Economic considerations

A major interpretive challenge is isolating AI-specific benefit from the broader effects of structured training. However, initial costs ranging from $45,000 to $250,000 may exacerbate existing training disparities between well-resourced and community programs. Cost-effectiveness modeling indicates the greatest value in high-volume centers, though economies of scale through multi-institutional platforms could democratize access to these technologies[86].

Economic evidence limitations

These economic findings must be interpreted cautiously as only five studies[69,83-86] contributed economic data, with substantial heterogeneity in cost definitions and reporting methods. The wide range in initial investments reflects different AI technologies and implementation scales, while break-even calculations

used varied methodologies that limit direct comparisons. Notably, most economic analyses excluded indirect costs, opportunity costs, and long-term sustainability metrics. Future economic evaluations should adopt standardized frameworks such as CHEERS guidelines[102-104] to enable meaningful cross-program comparisons and inform resource allocation decisions.

Future research priorities

Immediate research priorities should address current evidence gaps through several key initiatives. Future trials should stratify randomization by procedure complexity to ensure adequate power for procedure-specific estimates, particularly for less common procedures such as complex biliary reconstruction. Multicenter randomized trials comparing standardized AI curricula to traditional training with 5-year follow-up periods are essential to establish long-term outcomes. Development of core outcome sets specific to AI-enhanced surgical education would enable meaningful cross-study comparisons[103-104]. Implementation science methodologies should examine optimal integration strategies across diverse contexts[105], while equity-focused research must ensure benefits reach all trainees regardless of geographic or resource constraints.

Emerging opportunities in the field include federated learning approaches that enable privacy-preserving multi-institutional AI development[106], explainable AI systems providing transparent educational feedback[107], and integration with surgical data science initiatives for continuous improvement[108]. Throughout these advances, maintaining focus on patient-centered outcomes rather than technological capabilities remains paramount to ensure meaningful educational innovation.

CONCLUSIONS

This study provides high-certainty evidence that AI integration in HPB surgical training confers substantial clinical, educational, and safety benefits. Stratified analyses demonstrate that different AI technologies yield domain-specific advantages, supporting targeted rather than generic adoption. Realizing the full potential of AI in surgical education will require thoughtful curriculum design, educator readiness, rigorous quality assurance, and equity-focused implementation. These findings offer a robust foundation for advancing AI-enabled surgical training with measurable gains in trainee performance and patient care.

DECLARATIONS

Acknowledgments

The authors thank the Wolfson Medical Library for assistance with search strategy development, and the authors who provided additional data upon request.

Authors' contributions

Concept and design: Kanani F, Messer N, Nesher E, Gravetz A
Acquisition, analysis, or interpretation of data: Kanani F, Zoabi N, Yaacov G, Messer N, Gravetz A
Drafting of the manuscript: Kanani F, Zoabi N, Gravetz A
Critical revision of the manuscript for important intellectual content: All authors
Statistical analysis: Kanani F, Messer N, Gravetz A
Administrative, technical, or material support: Yaacov G, Carraro A, Lubezky N, Gravetz A
Supervision: Lubezky N, Nesher E, Gravetz A

Availability of data and materials

The data extraction forms, full search strategies, and statistical code are available from the corresponding author upon reasonable request. Individual study data remain with the original investigators.

Financial support and sponsorship
None.

Conflicts of interest
All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Copyright
© The Author(s) 2025.

REFERENCES

1. Cameron JL, He J. Two thousand consecutive pancreaticoduodenectomies. *J Am Coll Surg.* 2015;220:530-6. DOI PubMed
2. Vollmer CM Jr, Sanchez N, Gondek S, et al; Pancreatic Surgery Mortality Study Group. A root-cause analysis of mortality following major pancreatectomy. *J Gastrointest Surg.* 2012;16:89-102; discussion 102. DOI
3. Sachdeva AK, Flynn TC, Brigham TP, et al; American College of Surgeons (ACS) Division of Education; Accreditation Council for Graduate Medical Education (ACGME). Interventions to address challenges associated with the transition from residency training to independent surgical practice. *Surgery.* 2014;155:867-82. DOI
4. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg.* 2018;268:70-6. DOI
5. Maier-Hein L, Vedula SS, Speidel S, et al. Surgical data science for next-generation interventions. *Nat Biomed Eng.* 2017;1:691-6. DOI
6. Ward TM, Hashimoto DA, Ban Y, et al. Automated operative phase identification in peroral endoscopic myotomy. *Surg Endosc.* 2021;35:4008-15. DOI PubMed PMC
7. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25:44-56. DOI PubMed
8. Madani A, Namazi B, Altieri MS, et al. Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Ann Surg.* 2022;276:363-9. [Methods, pp. 364-5]. DOI
9. Mascagni P, Alapatt D, Urade T, et al. A computer vision platform to automatically locate critical events in surgical videos: documenting safety in laparoscopic cholecystectomy. *Ann Surg.* 2021;274:e93-5. DOI
10. Birkmeyer JD, Finks JF, O'Reilly A, et al; Michigan Bariatric Surgery Collaborative. Surgical skill and complication rates after bariatric surgery. *N Engl J Med.* 2013;369:1434-42. DOI
11. Bektaş M, Zonderhuis BM, Marquering HA, et al. Machine learning algorithms for predicting surgical outcomes after colorectal surgery: a systematic review. *World J Surg.* 2022;46:3100-10. DOI
12. Lavanchy JL, Zindel J, Kirtac K, et al. Surgical skill assessment using machine learning algorithms. *Br J Surg.* 2021;108:znab202.093. DOI
13. Madani A, Namazi B, Altieri MS, et al. Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Ann Surg.* 2022;276:363-9. [Results, pp. 366-7]. DOI
14. Endo Y, Tokuyasu T, Mori Y, et al. Impact of AI system on recognition for anatomical landmarks related to reducing bile duct injury during laparoscopic cholecystectomy. *Surg Endosc.* 2023;37:5752-9. DOI PubMed PMC
15. Hogg ME, Tam V, Zenati M, et al. Mastery-based virtual reality robotic simulation curriculum: the first step toward operative robotic proficiency. *J Surg Educ.* 2017;74:477-85. DOI
16. Rashidian N, Giglio MC, Van Herzeele I, et al. Effectiveness of an immersive virtual reality environment on curricular training for complex cognitive skills in liver surgery: a multicentric crossover randomized trial. *HPB.* 2022;24:2086-95. [Methods, pp. 2087-9]. DOI
17. Nota CL, Molenaar IQ, Te Riele WW, van Santvoort HC, Hagendoorn J, Borel Rinkes IHM. Stepwise implementation of robotic surgery in a high volume HPB practice in the Netherlands. *HPB.* 2020;22:1596-603. DOI PubMed
18. Zwart MJW, van den Broek B, de Graaf N, et al; Dutch Pancreatic Cancer Group. The feasibility, proficiency, and mastery learning curves in 635 robotic pancreatoduodenectomies following a multicenter training program: "standing on the shoulders of giants". *Ann Surg.* 2023;278:e1232-41. [Methods, pp. e1233-5]. DOI PubMed PMC
19. Brian R, Murillo AD, Gomes C, et al. Artificial intelligence and robotic surgical education. *Glob Surg Educ.* 2024;3:12. DOI
20. Lavanchy JL, Zindel J, Kirtac K, et al. Automation of surgical skill assessment using a three-stage machine learning algorithm. *Sci Rep.* 2021;11:5197. DOI PubMed PMC

21. Ahmad J. Training in robotic Hepatobiliary and pancreatic surgery: a step up approach. *HPB*. 2022;24:806-8. DOI PubMed
22. Fukumori D, Tschuor C, Penninga L, Hillingsø J, Svendsen LB, Larsen PN. Learning curves in robot-assisted minimally invasive liver surgery at a high-volume center in Denmark: report of the first 100 patients and review of literature. *Scand J Surg*. 2023;112:164-72. DOI PubMed
23. Kawka M, Gall TMH, Hand F, et al. The influence of procedural volume on short-term outcomes for robotic pancreatoduodenectomy - a cohort study and a learning curve analysis. *Surg Endosc*. 2023;37:4719-27. DOI PubMed PMC
24. Primavesi F, Urban I, Bartsch C, et al. Implementing MIS HPB surgery including the first robotic hepatobiliary program in Austria: initial experience and outcomes after 85 cases. *HPB*. 2023;25:S554-5. DOI
25. Fuentes SMS, Chávez LAF, López EMM, et al. The impact of artificial intelligence in general surgery: enhancing precision, efficiency, and outcomes. *Int J Res Med Sci*. 2024;12:112-9. Available from: https://www.msjonline.org/index.php/ijrms/article/view/14394 [accessed 30 July 2025].
26. Davis J, Robinson J, Tschuor C, et al. A novel application of cumulative sum (CuSUM) analytics for the objective evaluation of procedure specific technical dexterity in robotic hepatopancreatobiliary surgery. *HPB*. 2022;24:S24-5. DOI
27. Kirubarajan A, Young D, Khan S, Crasto N, Sobel M, Sussman D. Artificial intelligence and surgical education: a systematic scoping review of interventions. *J Surg Educ*. 2022;79:500-15. DOI PubMed
28. McGivern KG, Drake TM, Knight SR, et al. Applying artificial intelligence to big data in hepatopancreatic and biliary surgery: a scoping review. *Artif Intell Surg*. 2023;3:98-112. DOI
29. Ward TM, Fer DM, Ban Y, Rosman G, Meireles OR, Hashimoto DA. Challenges in surgical video annotation. *Comput Assist Surg*. 2021;26:58-68. DOI PubMed
30. Jung JJ, Jüni P, Lebovic G, Grantcharov T. First-year analysis of the operating room black box study. *Ann Surg*. 2020;271:122-7. DOI PubMed
31. Collins JW, Marcus HJ, Ghazi A, et al. Ethical implications of AI in robotic surgical training: a Delphi consensus statement. *Eur Urol Focus*. 2022;8:613-22. DOI
32. Gordon L, Grantcharov T, Rudzicz F. Explainable artificial intelligence for safe intraoperative decision support. *JAMA Surg*. 2019;154:1064-5. DOI PubMed
33. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. DOI PubMed PMC
34. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA*. 2000;283:2008-12. DOI
35. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev*. 2016;5:210. DOI PubMed PMC
36. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898. DOI
37. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919. DOI PubMed PMC
38. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*. 2017;358:j4008. DOI PubMed PMC
39. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. *Ottawa Hospital Research Institute*. 2021. Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp [accessed 30 July 2025].
40. Wolff RF, Moons KGM, Riley RD, et al; PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170:51-8. DOI
41. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ*. 2020;368:l6890. DOI PubMed PMC
42. Guyatt GH, Oxman AD, Vist GE, et al; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336:924-6. DOI
43. Winkler-Schwartz A, Yilmaz R, Mirchi N, et al. Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. *JAMA Netw Open*. 2019;2:e198363. DOI PubMed
44. Back E, Häggström J, Holmgren K, et al. Permanent stoma rates after anterior resection for rectal cancer: risk prediction scoring using preoperative variables. *Br J Surg*. 2021;108:1388-95. DOI PubMed PMC
45. Yu F, Silva Croso G, Kim TS, et al. Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. *JAMA Netw Open*. 2019;2:e191860. DOI PubMed PMC
46. Goodman ED, Patel KK, Zhang Y, et al. Analyzing surgical technique in diverse open surgical videos with multitask machine learning. *JAMA Surg*. 2024;159:185-92. DOI PubMed PMC
47. Panteleimonitis S, Miskovic D, Bissett-Amess R, et al; EARCS Collaborative. Short-term clinical outcomes of a European training programme for robotic colorectal surgery. *Surg Endosc*. 2021;35:6796-806. DOI PubMed PMC
48. Wu S, Tang M, Liu J, et al. Impact of an AI-based laparoscopic cholecystectomy coaching program on the surgical performance: a randomized controlled trial. *Int J Surg*. 2024;110:7816-23. DOI PubMed PMC
49. Leifman G, Golany T, Rivlin E, Khoury W, Assalia A, Reissman P. Real-time artificial intelligence validation of critical view of

safety in laparoscopic cholecystectomy. *Intell Based Med.* 2024;4:45-53. DOI

50. Mascagni P, Vardazaryan A, Alapatt D, et al. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Ann Surg.* 2022;275:955-61. DOI
51. Korndorffer JR Jr, Hawn MT, Spain DA, et al. Situating artificial intelligence in surgery: a focus on disease severity. *Ann Surg.* 2020;272:523-8. DOI
52. Rashidian N, Giglio MC, Van Herzeele I, et al. Effectiveness of an immersive virtual reality environment on curricular training for complex cognitive skills in liver surgery: a multicentric crossover randomized trial. *HPB*. 2022;24:2086-95. [Results, pp. 2090-2]. DOI
53. Tashiro Y, Aoki T, Kobayashi N, et al. A novel image-guided laparoscopic liver resection with integrated fluorescent imaging and artificial intelligence: a preliminary study. *J Clin Oncol.* 2024;42:568. DOI
54. Wang H, Ou Y, Hu P, et al. Application effect of mixed reality in the teaching of hepatobiliary surgery. *Chin J Med Educ Res*. 2019;41:1230-4. Available from: https://scholar.google.com/scholar?hl=zh-CN&as_sdt=0%2C48&q=Application+effect+of+mixed+reality+in+the+teaching+of+hepatobiliary+surgery&btnG=#d=gs_qabs&t=1753759007709&u=%23p%3Dn0vBcU2TMMoJ [accessed 30 July 2025].
55. Javaheri H, Ghamarnejad O, Widyaningsih R, et al. Enhancing perioperative outcomes of pancreatic surgery with wearable augmented reality assistance system: a matched-pair analysis. *Ann Surg Open.* 2024;5:e516. DOI
56. Zhu W, Zeng X, Hu H, et al. Perioperative and disease-free survival outcomes after hepatectomy for centrally located hepatocellular carcinoma guided by augmented reality and indocyanine green fluorescence imaging: a single-center experience. *J Am Coll Surg.* 2023;236:328-37. DOI PubMed
57. Magistri P, Guerrini GP, Ballarin R, Assirati G, Tarantino G, Di Benedetto F. Improving outcomes defending patient safety: the learning journey in robotic liver resections. *Biomed Res Int.* 2019;2019:1835085. DOI PubMed PMC
58. Onoe S, Mizuno T, Watanabe N, et al. Utility of modified pancreaticoduodenectomy (Hi-cut PD) for middle-third cholangiocarcinoma: an alternative to hepatopancreaticoduodenectomy. *HPB.* 2024;26:530-40. DOI
59. Sunakawa T, Kitaguchi D, Kobayashi S, et al. Deep learning-based automatic bleeding recognition during liver resection in laparoscopic hepatectomy. *Surg Endosc.* 2024;38:7656-62. DOI PubMed
60. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans Med Imaging.* 2017;36:86-97. DOI PubMed
61. Ward TM, Mascagni P, Madani A, Padoy N, Perretta S, Hashimoto DA. Surgical data science and artificial intelligence for surgical education. *J Surg Oncol*. 2021;124:221-30. DOI PubMed
62. Kennedy-Metz LR, Mascagni P, Torralba A, et al. Computer vision in the operating room: opportunities and caveats. *IEEE Trans Med Robot Bionics.* 2021;3:2-10. DOI PubMed PMC
63. Khalid S, Goldenberg M, Grantcharov T, Taati B, Rudzicz F. Evaluation of deep learning models for identifying surgical actions and measuring performance. *JAMA Netw Open.* 2020;3:e201664. DOI PubMed PMC
64. Hung AJ, Chen J, Gill IS. Automated performance metrics and machine learning algorithms to measure surgeon performance and anticipate clinical outcomes in robotic surgery. *JAMA Surg.* 2018;153:770-1. DOI PubMed PMC
65. Wagner M, Bihlmaier A, Kenngott HG, et al. A learning robot for cognitive camera control in minimally invasive surgery. *Surg Endosc.* 2021;35:5365-74. DOI PubMed PMC
66. Birkhoff DC, van Dalen ASHM, Schijven MP. A review on the current applications of artificial intelligence in the operating room. *Surg Innov.* 2021;28:611-9. DOI PubMed PMC
67. Kassahun Y, Yu B, Tibebu AT, et al. Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. *Int J Comput Assist Radiol Surg.* 2016;11:553-68. DOI PubMed
68. Datta S, Li Y, Ruppert MM, et al. Reinforcement learning in surgery. *Surgery.* 2021;170:329-32. DOI PubMed PMC
69. Luongo F, Hakim R, Nguyen JH, Anandkumar A, Hung AJ. Deep learning-based computer vision to recognize and classify suturing gestures in robot-assisted surgery. *Surgery.* 2021;169:1240-4. DOI PubMed PMC
70. Mellia JA, Basta MN, Toyoda Y, et al. Natural language processing in surgery: a systematic review and meta-analysis. *Ann Surg.* 2021;273:900-8. DOI PubMed
71. Zwart MJW, van den Broek B, de Graaf N, et al; Dutch Pancreatic Cancer Group. The feasibility, proficiency, and mastery learning curves in 635 robotic pancreatoduodenectomies following a multicenter training program: "standing on the shoulders of giants". *Ann Surg*. 2023;278:e1232-41. [Results, pp. e1236-9]. DOI PubMed PMC
72. Hashimoto DA, Rosman G, Witkowski ER, et al. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. *Ann Surg.* 2019;270:414-21. DOI PubMed PMC
73. Macario A. What does one minute of operating room time cost? *J Clin Anesth.* 2010;22:233-6. DOI
74. Procter LD, Davenport DL, Bernard AC, Zwischenberger JB. General surgical operative duration is associated with increased risk-adjusted infectious complication rates and length of hospital stay. *J Am Coll Surg.* 2010;210:60-5.e1. DOI PubMed
75. Birkmeyer JD, Stukel TA, Siewers AE, Goodney PP, Wennberg DE, Lucas FL. Surgeon volume and operative mortality in the United States. *N Engl J Med.* 2003;349:2117-27. DOI PubMed
76. Way LW, Stewart L, Gantert W, et al. Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. *Ann Surg.* 2003;237:460-9. DOI

77. Flum DR, Dellinger EP, Cheadle A, Chan L, Koepsell T. Intraoperative cholangiography and risk of common bile duct injury during cholecystectomy. *JAMA*. 2003;289:1639-44. DOI PubMed
78. Gao X, Jin Y, Dou Q, Heng PA. Automatic gesture recognition in robot-assisted surgery with reinforcement learning and tree search. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA); 2020 May 31-Aug 31; Paris, France. New York: IEEE; 2020. pp. 8440-6.
79. Schmidt RA, Lee TD, Winstein C, Wulf G, Zelaznik HN. Motor control and learning: a behavioral emphasis. 5th ed. Champaign, IL: Human Kinetics; 2011.
80. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg.* 1997;84:273-8. DOI
81. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg.* 2005;190:107-13. DOI
82. Stahl CC, Jung SA, Rosser AA, et al. Natural language processing and entrustable professional activity text feedback in surgery: a machine learning model of resident autonomy. *Am J Surg*. 2021;221:369-75. DOI PubMed PMC
83. Bartek MA, Saxena RC, Solomon S, et al. Improving operating room efficiency: machine learning approach to predict case-time duration. *J Am Coll Surg.* 2019;229:346-54.e3. DOI PubMed PMC
84. Schlemper J, Oktay O, Schaap M, et al. Attention gated networks: learning to leverage salient regions in medical images. *Med Image Anal.* 2019;53:197-207. DOI PubMed PMC
85. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med.* 2004;79:S70-81. DOI
86. Moglia A, Ferrari V, Morelli L, Ferrari M, Mosca F, Cuschieri A. A systematic review of virtual reality simulators for robot-assisted surgery. *Eur Urol.* 2016;69:1065-80. DOI PubMed
87. Vedula SS, Ishii M, Hager GD. Objective assessment of surgical technical skill and competency in the operating room. *Annu Rev Biomed Eng.* 2017;19:301-25. DOI PubMed PMC
88. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. *Med Teach.* 2010;32:638-45. DOI
89. Childers CP, Maggard-Gibbons M. Understanding costs of care in the operating room. *JAMA Surg.* 2018;153:e176233. DOI PubMed PMC
90. Spanjersberg WR, Reurings J, Keus F, van Laarhoven CJ. Fast track surgery versus conventional recovery strategies for colorectal surgery. *Cochrane Database Syst Rev.* ;2011:CD007635. DOI PubMed
91. Stefanidis D, Sevdalis N, Paige J, et al; Association for Surgical Education Simulation Committee. Simulation in surgery: what's needed next? *Ann Surg.* 2015;261:846-53. DOI PubMed
92. Strasberg SM. A three-step conceptual roadmap for avoiding bile duct injury in laparoscopic cholecystectomy: an invited perspective review. *J Hepatobiliary Pancreat Sci.* 2019;26:123-7. DOI PubMed
93. Pucher PH, Aggarwal R, Qurashi M, Darzi A. Meta-analysis of the effect of postoperative in-hospital morbidity on long-term patient survival. *Br J Surg.* 2014;101:1499-508. DOI PubMed
94. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. *J Surg Educ.* 2019;76:1681-90. DOI
95. Lam K, Chen J, Wang Z, et al. Machine learning for technical skill assessment in surgery: a systematic review. *NPJ Digit Med.* 2022;5:24. DOI PubMed PMC
96. Park A, Lee G, Seagull FJ, Meenaghan N, Dexter D. Patients benefit while surgeons suffer: an impending epidemic. *J Am Coll Surg.* 2010;210:306-13. DOI PubMed
97. Fitts PM, Posner MI. Human performance. Belmont, CA: Brooks/Cole; 1967.
98. Cate O. Competency-based postgraduate medical education: past, present and future. *GMS J Med Educ.* 2017;34:Doc69. DOI PubMed PMC
99. Bilgic E, Turkdogan S, Watanabe Y, et al. Effectiveness of telementoring in surgery compared with on-site mentoring: a systematic review. *Surg Innov.* 2017;24:379-85. DOI
100. Stewart LA, Clarke M, Rovers M, et al; PRISMA-IPD Development Group. Preferred reporting items for systematic review and meta-analyses of individual participant data: the PRISMA-IPD statement. *JAMA.* 2015;313:1657-65. DOI
101. Kirkham JJ, Davis K, Altman DG, et al. Core outcome set-STAndards for development: the COS-STAD recommendations. *PLoS Med.* 2017;14:e1002447. DOI PubMed PMC
102. Barkun JS, Aronson JK, Feldman LS, et al; Balliol Collaboration. Evaluation and stages of surgical innovations. *Lancet.* 2009;374:1089-96. DOI
103. Peters DH, Adam T, Alonge O, Agyepong IA, Tran N. Implementation research: what it is and how to do it. *BMJ.* ;347:f6753. DOI
104. Husereau D, Drummond M, Petrou S, et al; CHEERS Task Force. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *BMJ.* 2013;346:f1049. DOI PubMed
105. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. *IEEE Signal Process Mag.* 2020;37:50-60. DOI
106. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov.* 2019;9:e1312. DOI PubMed PMC
107. Tao F, Zhang H, Liu A, Nee AYC. Digital twin in industry: state-of-the-art. *IEEE Trans Ind Inf.* 2019;15:2405-15. DOI

108. Arute F, Arya K, Babbush R, et al. Quantum supremacy using a programmable superconducting processor. *Nature.* 2019;574:505-10. DOI PubMed

109. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. *PRISMA2020*: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Syst Rev.* 2022;18:e1230. DOI PubMed PMC