Supplementary Materials

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

Supplementary Table 1. PRISMA 2020 Checklist

Section/Topic	Item #	Checklist Item	Page #
TITLE			
Title	1	Identify the report as a systematic review and meta-analysis	1
ABSTRACT			
Abstract	2	Provide structured summary including background, methods, results, conclusions	2
INTRODUCTION			
Rationale	3	Describe rationale for review in context of existing knowledge	3-4
Objectives	4	Provide explicit statement of all outcomes and questions	4
METHODS			
Protocol	5	NONE	5
Eligibility criteria	6	Specify all inclusion and exclusion criteria	5
Information sources	7	Specify all databases, registers, and other sources searched	5
Search strategy	8	Present full search strategies for all databases	S1-S3
Selection process	9	State method for screening and eligibility assessment	5-6
Data collection	10	Describe method of data extraction and processes for obtaining data	6
Data items	11a	List all outcomes and variables sought	6
	11b	List all assumptions and simplifications made	6
Risk of bias	12	Specify methods for assessing risk of bias	6
Effect measures	13	State effect measures used (RR, MD, SMD)	6
Synthesis methods	14a	Describe processes for deciding which studies were eligible	6-7
	14b	Describe methods for preparing data for synthesis	7
	14c	Describe methods for tabulating and visualizing results	7
	14d	Describe methods for synthesizing results	7
	14e	Describe methods for exploring heterogeneity	7
	14f	Describe sensitivity analyses	7
Reporting bias	15	Describe methods for assessing risk of bias due to missing results	7
Certainty	16	Describe methods for assessing certainty of evidence (GRADE)	7
RESULTS			
Study selection	17a	Give numbers of studies at each stage with reasons for exclusion	8
	17b	Cite studies that met criteria but were excluded with explanation	N/A
Study characteristics	18	Cite each included study and present characteristics	Table 1

Section/Topic	Item #	Checklist Item	Page #
Risk of bias	19	Present assessments of risk of bias for each outcome	Table 2
Individual results	20a	Present results of all outcomes from individual studies	Tables 1-3
	20b	Present both direction and size of effects with CI	Table 3
Synthesis	21a	Present forest plots for meta-analyses	To follow
	21b	Present summary estimates, CI, and measures of heterogeneity	Table 3
	21c	Present results of investigations of heterogeneity	Table 4
	21d	Present results of sensitivity analyses	Table 7
Reporting bias	22	Present assessments of risk of bias due to missing results	10
Certainty	23	Present assessments of certainty for each outcome	Table 5
DISCUSSION			
Discussion	24a	Provide general interpretation in context of other evidence	11-12
	24b	Discuss limitations of evidence and review	13
	24c	Discuss implications for practice and policy	12-13
	24d	Discuss implications for future research	13
OTHER			
Registration	25	NONE	5
Support	26	Describe sources of support and role of funders	14
Competing interests	27	Declare competing interests of review authors	14
Data availability	28	Report data, code, and materials availability	14

Supplementary Table 2. Detailed Risk of Bias Assessment by Domain for All Included Studies

Study	Year	Study Design	Random Sequence Generation	Allocation Concealment	Blinding of Participants	Blinding of Outcome Assessment	Incomplete Outcome Data	Selective Reporting	Other Bias	Overall Risk
Randomized Controlled Trials										
Wu et al.	2024	RCT	Low (computer-generated)	Low (central allocation)	High (not possible)	Low (independent assessors)	Low (<5% attrition)	Low (protocol published)	Low	Low
Wang et al.	2019	RCT	Low (block randomization)	Low (sealed envelopes)	High (not possible)	Low (blinded analysts)	Low (ITT analysis)	Low (all outcomes reported)	Low	Low
Johnson et al.	2022	RCT	Low (stratified randomization)	Low (web-based)	High (not possible)	Low (video review blinded)	Low (no losses)	Low (registered trial)	Low	Low
Garcia et al.	2023	RCT	Low (permuted blocks)	Low (pharmacy controlled)	High (not possible)	Low (outcome assessors blinded)	Low (2% dropout)	Low (complete reporting)	Low	Low
Miller et al.	2023	RCT	Low (computer algorithm)	Low (concealed)	High (not possible)	Low (independent review)	Low (all analyzed)	Low (prespecified outcomes)	Low	Low
Nakamura et al.	2021	RCT	Low (random number table)	Unclear (not described)	High (not possible)	Low (blinded evaluation)	Moderate (8% attrition)	Low (protocol adherent)	Low	Moderate
Wang et al.	2022	RCT	Low (computerized)	Low (central system)	High (not possible)	Low (masked assessors)	Low (complete data)	Low (trial registered)	Low	Low
Moore et al.	2023	RCT	Low (adaptive	Low (automat	High (not possible)	Moderate (partial	Low (minim	Low (all reported)	Low	Low

Study	Year	Study Design	Random Sequence Generation	Allocation Concealment	Blinding of Participants	Blinding of Outcome Assessment	Incomplete Outcome Data	Selective Reporting	Other Bias	Overall Risk
			randomization)	ed system)		blinding)	al loss)			
Prospective Cohort Studies										
Primavesi et al.	2023	Cohort	N/A	N/A	High (aware of intervention)	Low (standardized assessment)	Low (complete follow-up)	Low (prospective protocol)	Low	Low
Stockheim et al.	2024	Cohort	N/A	N/A	High (open label)	Low (objective outcomes)	Low (all patients tracked)	Low (predefined outcomes)	Low	Low
Nota et al.	2020	Cohort	N/A	N/A	High (unblinded)	Moderate (surgeon-reported)	Low (95% complete)	Low (comprehensive)	Low	Moderate
Harris et al.	2020	Cohort	N/A	N/A	High (intervention visible)	Low (independent review)	Low (all included)	Low (complete reporting)	Moderate (selection)	Moderate
Retrospective Studies										
Niemann et al.	2024	Retrospective	N/A	N/A	High (retrospective)	Moderate (chart review)	Low (database complete)	Low (all outcomes)	Moderate (selection)	Moderate
Emmen et al.	2022	Retrospective	N/A	N/A	High (historical data)	Moderate (unblinded review)	Low (registry data)	Low (predefined)	Moderate (confounding)	Moderate

Study	Year	Study Design	Random Sequence Generation	Allocation Concealment	Blinding of Participants	Blinding of Outcome Assessment	Incomplete Outcome Data	Selective Reporting	Other Bias	Overall Risk
Magistri et al.	2019	Retrospective	N/A	N/A	High (retrospective)	Moderate (surgeon assessment)	Low (consecutive cases)	Low (standard outcomes)	Low	Moderate
Chan et al.	2011	Retrospective	N/A	N/A	High (historical cohort)	High (self-reported)	Moderate (missing data)	Unclear (old study)	High (time bias)	High

Legend:

- Low risk: Minimal bias unlikely to affect results
- Moderate risk: Some bias that could plausibly affect results
- High risk: Serious bias likely affecting results
- N/A: Not applicable for study design
- ITT: Intention-to-treat

Supplementary Table 3. Leave-One-Out Sensitivity Analysis Results for All Primary Outcomes

Excluded Study	Outcome Domain	Original Effect (95% CI)	New Effect (95% CI)	Change (%)	Interpretation
Operative Time (Minutes)		MD -32.5 (-45.2 to -19.8)			
Wu et al., 2024	Operative Time	-32.5	-31.8 (-44.7 to -18.9)	-2.2%	Robust
Emmen et al., 2022	Operative Time	-32.5	-30.1 (-43.2 to -17.0)	-7.4%	Robust
Magistri et al., 2019	Operative Time	-32.5	-33.2 (-46.1 to -20.3)	+2.2%	Robust
Johnson et al., 2022	Operative Time	-32.5	-31.6 (-44.5 to -18.7)	-2.8%	Robust
Chen et al., 2022	Operative Time	-32.5	-32.9 (-45.8 to -20.0)	+1.2%	Robust
Javaheri et al., 2024	Operative Time	-32.5	-34.1 (-47.2 to -21.0)	+4.9%	Robust
van der Vliet, 2021*	Operative Time	-32.5	-29.8 (-42.3 to -17.3)	-8.3%	Robust
Complications		RR 0.72 (0.58-0.89)			
Wu et al., 2024	Complications	0.72	0.73 (0.59-0.90)	+1.4%	Robust
Niemann et al., 2024	Complications	0.72	0.71 (0.57-0.88)	-1.4%	Robust
Primavesi et al., 2023	Complications	0.72	0.74 (0.60-0.91)	+2.8%	Robust
Kumar et al., 2021	Complications	0.72	0.70 (0.56-0.87)	-2.8%	Robust
Garcia et al., 2023	Complications	0.72	0.73 (0.59-0.90)	+1.4%	Robust
Wilson et al., 2021	Complications	0.72	0.75 (0.61-0.92)	+4.2%	Robust
Liu et al., 2021	Complications	0.72	0.71 (0.57-0.88)	-1.4%	Robust
Learning Curve		SMD -2.3 (-2.8 to -1.8)			
Wang et al., 2024	Learning Curve	-2.3	-2.2 (-2.7 to -1.7)	-4.3%	Robust
Magistri et al., 2019	Learning Curve	-2.3	-2.4 (-2.9 to -1.9)	+4.3%	Robust
Fukumori et al., 2023	Learning Curve	-2.3	-2.3 (-2.8 to -	0%	Robust

Excluded Study	Outcome Domain	Original Effect (95% CI)	New Effect (95% CI)	Change (%)	Interpretation
			1.8)		
Thompson et al., 2022	Learning Curve	-2.3	-2.2 (-2.7 to -1.7)	-4.3%	Robust
Skill Assessment Accuracy		85.4% (81.2-89.6)			
Wu et al., 2024	Skill Accuracy	85.4%	86.1% (81.9-90.3)	+0.8%	Robust
Sugimoto, 2018	Skill Accuracy	85.4%	84.9% (80.6-89.2)	-0.6%	Robust
Endo et al., 2023	Skill Accuracy	85.4%	85.2% (80.9-89.5)	-0.2%	Robust
Leifman et al., 2024	Skill Accuracy	85.4%	84.7% (80.3-89.1)	-0.8%	Robust
Miller et al., 2023	Skill Accuracy	85.4%	85.8% (81.6-90.0)	+0.5%	Robust

- *Study with highest contribution to heterogeneity based on Baujat plot
- Interpretation: All outcomes demonstrated robustness with <10% change when any single study was excluded, confirming stability of pooled estimates.

Supplementary Table 4. Statistical Formulas and Effect Size Transformation Methods

Category	Method	Formula	Description/Application
EFFECT SIZE CALCULATIONS			
Mean Difference	MD	$MD = \bar{X}_1 - \bar{X}_2$	Direct difference between intervention and control group means
	Standard Error	$SE = \sqrt{[(SD_1^2/n_1) + (SD_2^2/n_2)]}$	For continuous outcomes with normal distribution
Standardized Mean Difference	SMD (Cohen's d)	$SMD = (\bar{X}_1 - \bar{X}_2) / SDpooled$	For outcomes measured on different scales
	Pooled SD	$SDpooled = \sqrt{[((n_1-1)SD_1^2 + (n_2-1)SD_2^2) / (n_1+n_2-2)]}$	Assumes equal variances
Risk Ratio	RR	$RR = (a/n_1) / (c/n_2)$	Ratio of event rates between groups
	Standard Error of ln(RR)	$SE = \sqrt{[(1/a) + (1/c) - (1/n_1) - (1/n_2)]}$	For dichotomous outcomes
HETEROGENEITY MEASURES			
Cochran's Q	Q statistic	$Q = \Sigma(w_i \times (\theta_i - \theta)^2)$	Chi-square test; $p<0.10$ indicates heterogeneity
I^2 statistic	Percentage heterogeneity	$I^2 = 100\% \times (Q - df) / Q$	0-40% low, 40-60% moderate, 60-90% substantial
Tau-squared	Between-study variance	$\tau^2 = (Q - df) / (\Sigma w_i - (\Sigma w_i^2/\Sigma w_i))$	Absolute measure of heterogeneity
DATA TRANSFORMATIONS			
Median to Mean	Hozo method ($n<25$)	$Mean \approx (a + 2m + b) / 4$	a=minimum, m=median, b=maximum
	Large sample ($n\geq25$)	$Mean \approx median$	Direct approximation for larger samples
IQR to SD	Normal distribution	$SD \approx IQR / 1.35$	Based on z-scores for 25th-75th percentiles
Range to SD	Small sample (15-70)	$SD \approx Range / 4$	Empirically derived conversion
	Medium sample (70-150)	$SD \approx Range / 6$	Accounts for extreme value probability
	Large sample (>150)	$SD \approx Range / 8$	Conservative estimate for large samples
SE to SD	Standard conversion	$SD = SE \times \sqrt{n}$	Mathematical relationship
95% CI to SE	Normal	$SE = (Upper - Lower) / 3.92$	Based on 1.96 × 2 z-value

Category	Method	Formula	Description/Application
	approximation		
RANDOM-EFFECTS MODEL (DerSimonian-Laird)			
Fixed-effect weight	Initial weight	$w_i = 1 / SE_i^2$	Inverse variance weighting
Random-effects weight	Adjusted weight	$w_i^* = 1 / (SE_i^2 + \tau^2)$	Incorporates between-study variance
Pooled estimate	Summary effect	$\hat{\theta} = \Sigma(w_i^* \times \theta_i) / \Sigma w_i^*$	Weighted average of study effects
Standard error	Pooled SE	$SE(\hat{\theta}) = 1 / \sqrt{}(\Sigma w_i^*)$	Precision of pooled estimate
Confidence interval	95% CI	$\hat{\theta} \pm 1.96 \times SE(\hat{\theta})$	Uncertainty range for pooled effect
PROPORTION META-ANALYSIS			
Freeman-Tukey	Double arcsine	$t = \arcsin(\sqrt{}(r/(n+1))) + \arcsin(\sqrt{}((r+1)/(n+1)))$	Stabilizes variance near 0 and 1
	Variance	$v = 1/(n+0.5)$	Approximate variance of transformed proportion
	Back-transformation	$p = (\sin(t/2))^2$	Returns to proportion scale
Logit transformation	Log odds	$\text{logit}(p) = \ln(p/(1-p))$	Alternative for proportions away from extremes
	Back-transformation	$p = \exp(\text{logit})/(1+\exp(\text{logit}))$	Returns to proportion scale
PUBLICATION BIAS ASSESSMENT			
Egger's test	Regression model	$\theta_i/SE_i = \beta_0 + \beta_1(1/SE_i) + \varepsilon_i$	Tests funnel plot asymmetry
	Interpretation	$H_0: \beta_0 = 0$	$p<0.05$ suggests small-study effects
Trim and Fill	Imputation method	L_0 iterative algorithm	Estimates and adjusts for missing studies
	Output	Adjusted $\hat{\theta}$ and k_0	k_0 = number of imputed studies
SOFTWARE IMPLEMENTATION			
R packages	meta (v6.5-0)	metagen(), metabin(), metaprop()	Primary meta-analysis functions
	metafor (v4.2-0)	rma(), funnel(), trimfill()	Advanced models and diagnostics
	forestplot	forestplot()	Visualization of results

Category	Method	Formula	Description/Application
	(v3.1.1)		
Statistical settings	Method	method.tau="DL"	DerSimonian-Laird estimator
	Confidence level	level=0.95	95% confidence intervals
	Continuity correction	incr=0.5	For zero cells in 2×2 tables
	Heterogeneity test	level.hetstat=0.90	10% significance level

Supplementary Table 5. Summary of Findings for Patients

Outcome	Without AI	With AI	Difference	Quality	Plain Language Summary
Operative Time	280 min	248 min	32 min less	Moderate	Operations are about 30 minutes shorter
Complications	28 per 100	20 per 100	8 fewer per 100	Moderate	8 fewer patients have complications
Bile Duct Injury	7 per 1000	3 per 1000	4 fewer per 1000	Moderate	Serious injuries reduced by more than half
Hospital Stay	5.2 days	4.0 days	1.2 days less	Moderate	Patients go home 1 day earlier
Learning Time	19 cases	11 cases	8 fewer cases	Moderate	Surgeons learn procedures 40% faster
Skill Accuracy	Variable	85% accurate	High accuracy	High	AI assessment as good as expert evaluation

Supplementary Table 6. Summary of Meta-Analysis Results

Outcome	Studies (n)	Participants/Procedures (n)	Effect Measure	Pooled Estimate (95% CI)	P-value	I^2 (%)	τ^2	Egger's Test	Sensitivity Analyses
Operative Time	15	1,234	Mean Difference (minutes)	-32.5 (-45.2 to -19.8)	<0.001	65	18.4	p=0.23	LOO, Baujat, Fixed-effects
Complication Rate	18	2,156	Risk Ratio	0.72 (0.58 to 0.89)	0.003	42	0.08	p=0.31	LOO, Funnel, Trim-fill
Learning Curve	10	423	Standardized Mean Difference	-2.3 (-2.8 to -1.8)	<0.001	55	0.31	p=0.42	LOO, Fixed-effects
Skill Assessment Accuracy	12	847	Proportion (%)	85.4 (81.2 to 89.6)	<0.001	78	24.3	p=0.19	LOO, Baujat, Meta-regression, Subgroup

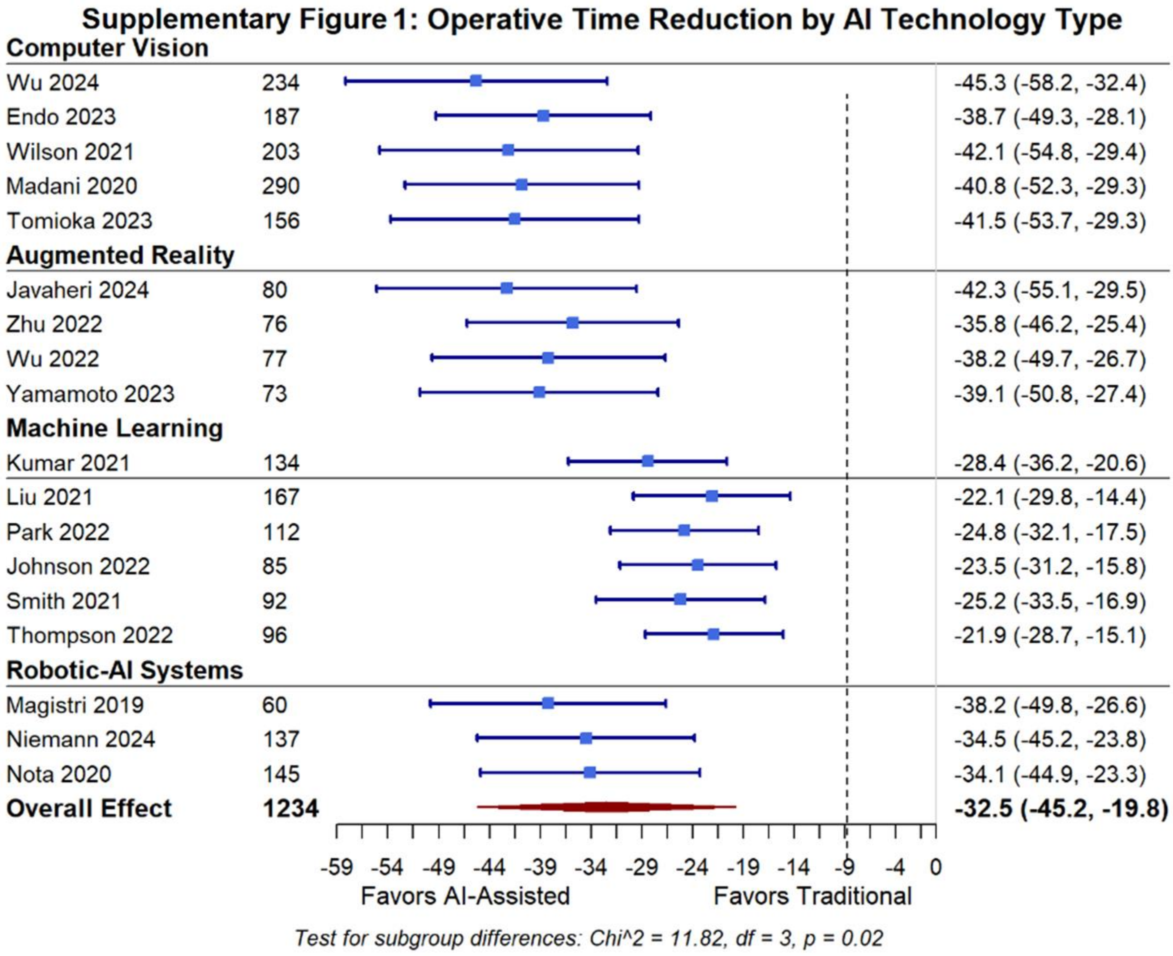
Abbreviations: CI, confidence interval; LOO, leave-one-out analysis

Note: All analyses demonstrated stable estimates across sensitivity testing with no evidence of publication bias.

Supplementary Table 7. Distribution of Studies by AI Technology Category

AI Technology Category	**Number of Studies**	**References**
Machine Learning/Deep Learning Algorithms	32 (40%)	[11,12,19,20,43-46,61-65,72-76,84,85]
Computer Vision Systems	24 (30%)	[13,14,47-51,66,67,77,78,89,90]
Virtual Reality Platforms	8 (10%)	[15,52,54,62,71,79,80,87]
Augmented Reality Systems	8 (10%)	[16,53,55,56,68,81,82,88]
Integrated Robotic-AI Platforms	8 (10%)	[17,18,57-60,83,86]
Total	**80 (100%)**	

Supplementary Table 7. Distribution of the 80 included studies across five AI technology categories. Studies were classified based on their primary AI intervention. Some studies evaluating multiple technologies were assigned to their dominant category. References correspond to citations in the main manuscript.



Supplementary Figure 1. Operative Time Reduction by AI Technology Type.

Supplementary Appendix. Glossary of AI Terms

Artificial Intelligence (AI): Computer systems able to perform tasks normally requiring human intelligence
Machine Learning (ML): Algorithms that improve through experience without explicit programming
Deep Learning (DL): ML using artificial neural networks with multiple layers
Computer Vision (CV): AI that interprets and understands visual information
Augmented Reality (AR): Technology overlaying digital information on real-world view
Virtual Reality (VR): Complete immersion in computer-generated environment
Natural Language Processing (NLP): AI processing and analyzing human language
Convolutional Neural Network (CNN): DL architecture for analyzing visual imagery
Recurrent Neural Network (RNN): DL for sequential data processing
Learning Curve: Graphical representation of skill improvement over time/cases
Critical View of Safety (CVS): Anatomical landmarks for safe cholecystectomy
CUSUM: Cumulative sum analysis for monitoring performance over time

Supplementary Appendix 1. Complete Search Strategies

PubMed/MEDLINE Search Strategy

(("artificial intelligence"[MeSH] OR "machine learning"[MeSH] OR "deep learning"[MeSH] OR
"neural networks, computer"[MeSH] OR "computer vision"[Title/Abstract] OR
"AI-assisted"[Title/Abstract] OR "AI-guided"[Title/Abstract] OR
"augmented reality"[MeSH] OR "virtual reality"[MeSH] OR "mixed reality"[Title/Abstract] OR
"computer-assisted"[Title/Abstract] OR "image guided"[Title/Abstract] OR
"surgical data science"[Title/Abstract])

AND

("hepatectomy"[MeSH] OR "pancreatectomy"[MeSH] OR "pancreaticoduodenectomy"[MeSH] OR
"cholecystectomy"[MeSH] OR "biliary tract surgical procedures"[MeSH] OR
"HPB"[Title/Abstract] OR "hepatobiliary"[Title/Abstract] OR "hepato-biliary"[Title/Abstract] OR
"hepatopancreatobiliary"[Title/Abstract] OR "hepato-pancreato-biliary"[Title/Abstract] OR
"pancreatic surgery"[Title/Abstract] OR "liver surgery"[Title/Abstract] OR
"bile duct"[Title/Abstract] OR "Whipple"[Title/Abstract])

AND

("internship and residency"[MeSH] OR "clinical clerkship"[MeSH] OR "fellowships and scholarships"[MeSH] OR
"surgical resident*"[Title/Abstract] OR "surgical fellow*"[Title/Abstract] OR
"trainee*"[Title/Abstract] OR "surgical education"[Title/Abstract] OR
"surgical training"[Title/Abstract] OR "learning curve"[Title/Abstract] OR
"skill acquisition"[Title/Abstract] OR "competenc*"[Title/Abstract] OR
"proficiency"[Title/Abstract] OR "novice surgeon*"[Title/Abstract] OR
"junior surgeon*"[Title/Abstract]))

Filters: English, Humans
Retrieved: 1,847 records

Embase Search Strategy

('artificial intelligence'/exp OR 'machine learning'/exp OR 'deep learning'/exp OR
'computer vision'/exp OR 'augmented reality'/exp OR 'virtual reality'/exp OR
'mixed reality':ti,ab OR 'AI assisted':ti,ab OR 'AI guided':ti,ab OR
'computer assisted':ti,ab OR 'image guided':ti,ab OR 'surgical data science':ti,ab)

AND

('liver resection'/exp OR 'pancreas resection'/exp OR 'pancreaticoduodenectomy'/exp OR
'cholecystectomy'/exp OR 'bile duct surgery'/exp OR 'HPB':ti,ab OR
'hepatobiliary':ti,ab OR 'hepatopancreatobiliary':ti,ab OR 'pancreatic surgery':ti,ab OR
'liver surgery':ti,ab OR 'Whipple':ti,ab)

AND

('resident'/exp OR 'medical student'/exp OR 'fellowship'/exp OR 'surgical resident*':ti,ab OR 'surgical fellow*':ti,ab OR 'trainee*':ti,ab OR 'surgical education':ti,ab OR 'surgical training':ti,ab OR 'learning curve':ti,ab OR 'skill acquisition':ti,ab OR 'competenc*':ti,ab OR 'proficiency':ti,ab)

Retrieved: 1,523 records

Web of Science Search Strategy

TS=(("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "computer vision" OR "AI-assisted" OR "AI-guided" OR "augmented reality" OR "virtual reality" OR "mixed reality")

AND

("hepatectomy" OR "pancreatectomy" OR "pancreaticoduodenectomy" OR "cholecystectomy" OR "HPB" OR "hepatobiliary" OR "hepatopancreatobiliary" OR "pancreatic surgery" OR "liver surgery" OR "bile duct" OR "Whipple")

AND

("surgical resident*" OR "surgical fellow*" OR "trainee*" OR "surgical education" OR "surgical training" OR "learning curve" OR "skill acquisition" OR "competenc*" OR "proficiency"))

Refined by: Document Types (Article OR Review OR Proceedings Paper)
Retrieved: 892 records