



THE EVIDENCE BASE FOR IMPORTANCE OF ARTIFICIAL INTELLIGENCE IN GENERAL SURGERY

Dr A.Swathi Reddy^{1*}, Dr Yanala Sheker², Dr Janipalli Mounica³

^{1*} Assistant professor General surgery, Arundhati institute of medical sciences Hyderabad

² Assistant professor, General surgery Arundhati institute of medical sciences Hyderabad

³ Assistant professor general surgery, Arundhati institute of medical sciences Hyderabad

*Corresponding author – Email - swathi.anumula4@gmail.com

Abstract

Background: Publication output on artificial intelligence (AI) in general surgery has grown steeply, and it is widely assumed that prospective clinical evaluation has failed to keep pace. That assumption has not been tested. We quantified the relationship between published output and registered prospective evaluation over eleven years.

Methods: Cross-sectional analysis of Clinical Trials.gov and PubMed covering complete calendar years 2015 to 2025. A pre-specified vocabulary of 21 AI terms, 48 general surgery terms and 24 exclusion terms was applied identically to both sources so that numerator and denominator were matched. Annual percentage change was estimated by Poisson regression with heteroscedasticity-consistent standard errors. The pre-specified primary hypothesis, that publications were growing faster than registered trials, was tested as a year-by-source interaction, the exponentiated coefficient giving the ratio of growth rates. A stricter exclusion screen was applied as a sensitivity analysis.

Results: Of 447 registry records retrieved, 145 met criteria (sensitivity analysis, 135). Over the period there were 6,818 publications, 145 registrations, 38 interventional trials and 24 randomised trials, a ratio of 179 publications Per interventional trial and 284 per randomized trial. No trial was registered before 2018, by which time 476 papers had appeared. The primary hypothesis was not supported: publications grew by 37.3% per year (95%CI 34.8 to 39.8) and interventional trials by 43.7% per year (95%CI 27.4 to 62.2), giving a growth -rate Ratio of 0.955 (95%CI 0.845 to 1.079; $p=0.461$). Registered studies were predominantly observational (107/145, 73.8%), single-centre (125/145, 86.2%) and non-industry (142/145, 97.9%); four (2.8%) recruited in more than one country and two of 38 interventional trials (5.3%) declared a phase. Median planned enrolment in interventional trials was 110 (IQR 50–266). Of 49 completed studies, one (2.0%) had posted results, at a median of 25 months after completion (IQR 19–31); 39 registrations (26.9%) had lapsed to unknown status.

Conclusion: The gap between publication and prospective evaluation in surgical AI is not widening. It is fixed at an extreme level and has been for a decade. Registered evaluation grew in proportion to the literature, but from a base so small, and of such uniform design weakness, that eleven years produced 24 randomised trials, none of which has reported results. The constraint on this field is not the volume of research but that almost none of it is designed to change practice, and almost none of what finishes is ever reported.

KeyWords: artificial intelligence; machine learning; general surgery; clinical trials as topic; translational research; research reporting.

Introduction

Artificial intelligence has become one of the most productive topics in surgical research. Models that recognise anatomy in laparoscopic video, that predict anastomotic leak from preoperative variables, or that draft operative documentation now appear in every major surgical journal, and several bibliometric analyses have documented the growth in output. Those analyses agree on the shape of the curve: steep, accelerating, and concentrated in a small number of countries.

A bibliometric study, however, counts what has been written. It cannot say whether anything has been evaluated. The distinction carries more weight in artificial intelligence than in most fields because the cost of producing a publishable surgical AI paper is unusually low. A retrospective dataset, a standard architecture and a favourable internal split will yield an area under the curve above 0.90 and a manuscript. No patient is prospectively enrolled, no protocol is committed to in advance, and no sample size is declared.

The steps that require those commitments are recorded elsewhere. Prospective registration on a public registry is the first point at which a hypothesis is fixed before the data are seen. It is also the point at which a study becomes traceable, so that one which finishes without reporting can be identified as such. Registrations and publications can be counted using the same vocabulary, which converts a rhetorical concern about hype into a testable quantity.

Work in neighbouring specialties suggests where the answer might lie. Systematic reviews of deep learning in medical imaging have repeatedly found sparse external validation and poor reporting. Analyses of regulatory submissions have found that most cleared algorithms were assessed retrospectively at a small number of sites. Reporting guidelines written for this precise problem, including TRIPOD+AI, DECIDE-AI, and the CONSORT-AI and SPIRIT-AI extensions, exist because the underlying evidence was judged inadequate. General surgery has not been examined separately, despite features that make translation harder than in radiology: the intervention is a procedure, the operator forms part of the system, and the outcome that matters is a complication rate rather than a classification accuracy.

We set out to test a hypothesis stated in advance, that publication output in surgical artificial intelligence was growing faster than prospective registered evaluation and that the gap between them was therefore widening. What we found was not what we expected, and there as on it matters is set out below.

Methods

Design, registration and reporting

This was a cross-sectional analysis of publicly available registry and bibliographic records. No individual patient data were accessed and no human participants were involved, so ethical approval was not required. The protocol, including all three search vocabularies, the eligibility rules, the primary hypothesis and the analysis plan, was registered on the Open Science Framework before extraction. Reporting follows the STROBE recommendations for cross-sectional studies where these apply to a study without human participants. Data and code are openly available.

Data sources and study period

Registered studies were obtained from ClinicalTrials.gov and annual publication counts from PubMed. The period was restricted to complete calendar years 2015 to 2025. The partial year in progress at the time of analysis was excluded, because an incomplete final year depresses the terminal data point and biases trend estimation downwards.

Search strategy and screening

Three vocabularies were fixed in advance and are reproduced verbatim in Supplementary File 1. The artificial intelligence vocabulary contained 21 terms spanning classical machine learning, deep learning, computer vision, natural language processing, radiomics and generative models. The general surgery vocabulary contained 48 terms covering the core operative scope of the specialty: alimentary tract, abdomen and abdominal wall, breast, endocrine, and surgical critical care. The exclusion vocabulary contained 24 terms removing cardiothoracic, neurological, orthopaedic, urological, obstetric, ophthalmic and dental contexts.

The identical three-vocabulary logic was applied to both sources so that the publication numerator and the trial denominator describe the same field. This matters, because comparing a broadly defined publication count against a narrowly defined trial count would inflate any ratio between them. The registry search was executed through the ClinicalTrials.gov interface using the artificial intelligence and general surgery vocabularies, and the exclusion vocabulary was applied to the exported records against the study title and conditions fields. A record was included if an artificial intelligence term and a general surgery term appeared in the title, conditions, brief summary or interventions, no exclusion term appeared in the title or conditions, and the record was first posted within the study period.

Because the boundary of general surgery is a matter of judgement, a second and deliberately more aggressive exclusion screen was applied as a pre-planned sensitivity analysis, additionally removing thoracic, plastic, transplant and gynaecological oncology contexts. Both sets of results are reported.

Variables and outcomes

Extracted variables were registration number, title, dates of first posting and completion, recruitment status, study type, phase, allocation, masking, planned enrolment, funder type, recruiting locations, and whether summary results had been posted. Each record was assigned to one general surgery sub domain using an ordered regular-expression hierarchy. Countries were derived from the location field and classified against World Bank income groupings. The primary outcome was the ratio of growth in publications to growth in registered interventional trials. Secondary outcomes were the absolute ratio of publications to trials, the proportion of completed studies with posted results, and the distribution of design, funder type, planned enrolment, number of sites and geography.

Statistical analysis

Counts were modelled by Poisson regression on calendar year with heteroscedasticity-consistent (HC0) standard errors, annual percentage change being obtained by exponentiating the year coefficient. The primary hypothesis was tested by fitting a single Poisson model to the stacked annual counts with year, source and a year-by-source interaction term. The exponentiated interaction coefficient is the ratio of the two growth rates and its confidence interval provides the test; a value above unity indicates publications outpacing trials. This approach yields a direct p value for the difference, which comparison of two separately fitted slopes does not. Continuous variables are reported as median and interquartile range and categorical variables as counts and percentages, with denominators stated. No values were imputed. Two-sided p values below 0.05 were treated as evidence against the null. Analyses used Python 3.11 with pandas, numpy, statsmodels and matplotlib.

Results

Study selection

The registry search retrieved 447 records. Of these, 232 were excluded because no general surgery term was present, 27 because the surgical context belonged to another specialty, and 43 because the record was first posted outside the study period, leaving 145 registrations for the primary analysis. The stricter sensitivity screen retained 135. Findings were materially unchanged between the two, and the primary set is reported throughout unless stated otherwise.

Publications, registrations and the ratio between them

PubMed returned 6,818 publications over the eleven years, rising from 88 in 2015 to 1,898 in 2025. Across the same period the registry contained 145 studies of any design, of which 38 were interventional and 24 randomised. No trial of any kind was registered before 2018, by which time 476 papers had already been published. The overall ratio was 179 publications for every registered interventional trial and 284 for every randomised trial (Table 3, Figure 1).

Table 3. Annual publication output and registered studies, 2015–2025.

Year	Publications	All registrations	Interventional	Randomised	Publications per trial
2015	88	0	0	0	—
2016	90	0	0	0	—
2017	117	0	0	0	—
2018	181	3	1	0	181
2019	273	9	4	2	68
2020	397	7	2	2	199
2021	634	9	2	1	317
2022	845	14	4	4	211
2023	1,034	27	4	2	259
2024	1,261	40	10	6	126
2025	1,898	36	11	7	173
Total	6,818	145	38	24	179

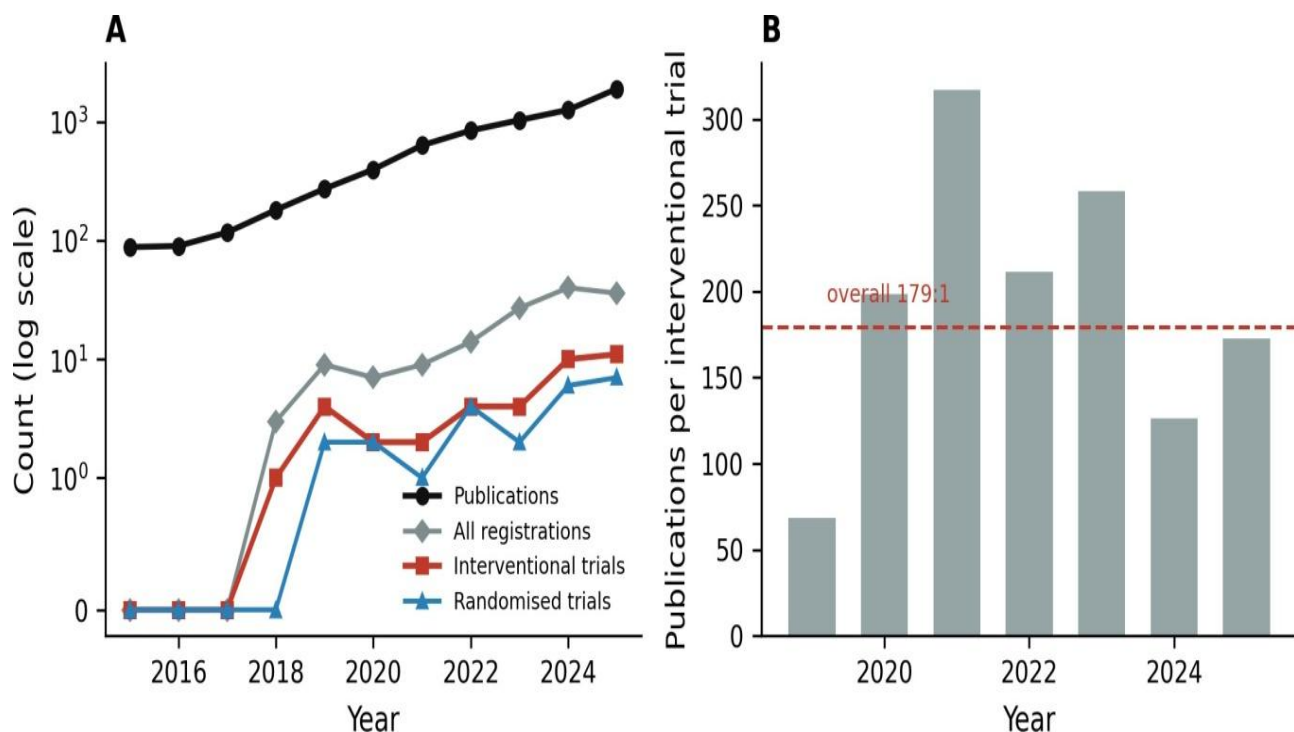


Figure 1. Publication output and registered evaluation, 2015–2025. (A) Annual counts on a symmetric logarithmic scale. (B) Publications per registered interventional trial by year, from 2018 when the first trial was registered; the dashed line marks the overall ratio of 179 to 1.

Publications grew by 37.3% per year (95% CI 34.8 to 39.8; $p < 0.001$). Registered interventional trials grew by 43.7% per year (95% CI 27.4 to 62.2; $p < 0.001$), randomised trials by 46.0% per year (95% CI 29.4 to 64.8; $p < 0.001$), and all registrations by 47.7% per year (95% CI 33.9 to 62.8; $p < 0.001$).

The pre-specified primary hypothesis was not supported. The ratio of growth rates for publications against interventional trials was 0.955 (95% CI 0.845 to 1.079; $p = 0.461$), the point estimate lying below unity and indicating if anything that trials grew marginally faster than publications. The same held against randomised trials (0.940, 95% CI 0.832 to 1.063; $p = 0.324$) and against all registrations (0.930, 95% CI 0.842 to 1.027; $p = 0.151$). The two series have therefore risen in proportion and the ratio between them has not narrowed.

Design of registered studies

Registered work was predominantly observational: 107 of 145 studies (73.8%) against 38 Interventional (26.2%). Among interventional studies, 24 (63.2%) used randomized allocation and 16 (42.1%) reported any masking, but only two (5.3%) declared a phase. Median planned enrolment was 110 participants in interventional studies (IQR 50–266) against 353 across all registrations (IQR 100–1,154), the difference reflecting large retrospective observational cohorts. One hundred and twenty-five studies (86.2%) listed a single recruiting site and only four (2.8%) recruited in more than one country. Industry was the lead funder in three registrations (2.1%), the remainder being academic or governmental (Table 1, Figure 2).

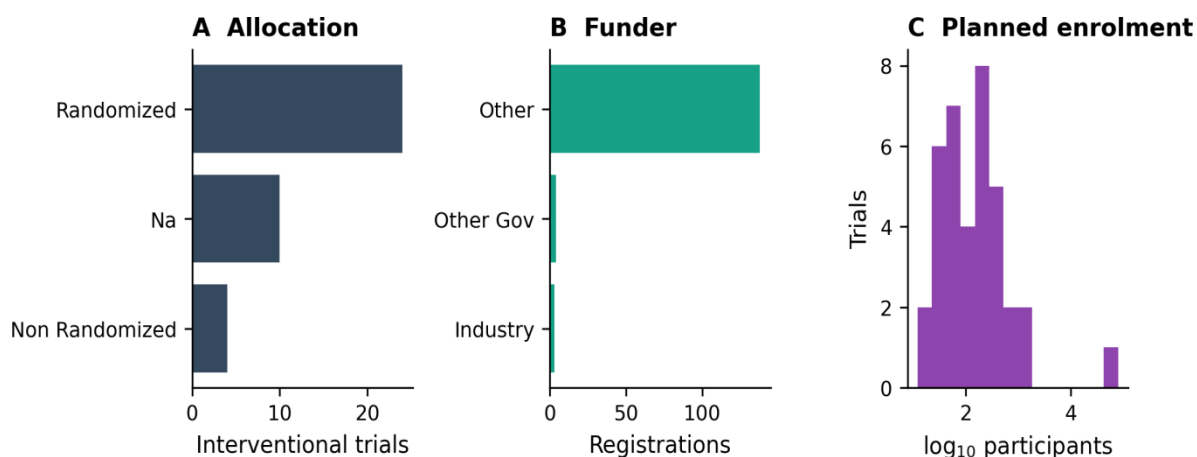


Figure 2. Design and funding of registered studies. (A) Allocation among interventional studies. (B) Funder type across all registrations. (C) Planned enrolment in interventional studies, logarithmic scale.

Table 1. Characteristics of registered studies of artificial intelligence in general surgery, 2015–2025.

Characteristic	Primary analysis (n=145)	Sensitivity analysis (n=135)
Interventional	38 (26.2%)	37 (27.4%)
Observational	107 (73.8%)	98 (72.6%)
Randomised, of interventional	24/38 (63.2%)	24/37 (64.9%)
Any masking, of interventional	16/38 (42.1%)	16/37 (43.2%)
Phase declared, of interventional	2/38 (5.3%)	2/37 (5.4%)
Single recruiting site	125 (86.2%)	115 (85.2%)
Recruiting in more than one country	4 (2.8%)	4 (3.0%)

Industry-funded	3(2.1%)	3(2.2%)
Atleastonesiteinalow/middle-incomecountry	58(40.0%)	53(39.3%)
Medianplannedenrolment,all(IQR)	353(100–1154)	342(100–1086)
Medianplannedenrolment,interventional (IQR)	110(50–266)	100(50–281)
Completed	49(33.8%)	45(33.3%)
Completedwithresults posted	1(2.0%)	1(2.2%)
Lapsedtounknownstatus	39(26.9%)	36(26.7%)

IQR,interquartile range.Percentagesare oftheanalysis set unlessa denominator is shown.

Completion and reporting

Forty-nine studies (33.8%)were recorded as completed. One(2.0%)had posted summary results.The 48 completed studies without posted results had been finished for a median of 25 months (IQR 19–31). A further 39 registrations (26.9%)had lapsed to unknown status, meaning the entry had not been updatedwithintheverificationwindowandthefateofthestudy cannotbedeterminedfromthepublic record. Two studies (1.4%) were terminated or withdrawn.

Sub domain and geographical distribution

The largest group comprised perioperative and general work not specific to one organ system (52 studies,35.9%),followed by hepatopancreatobiliary (20, 13.8%),upper gastrointestinal and bariatric (18, 12.4%) and colorectal (13, 9.0%). Endocrine surgery contributed four studies (2.8%) and no interventional trial at all (Table 2). Locations were listed for 125 studies (86.2%), spanning 31 countries. China accounted for 47 (37.6% of located studies), more than six times the United States (six, 4.8%)and more than the next five countries combined. Fifty-eight studies(40.0%)listed atleast one site in a low-or middle-income country, a figure driven almost entirely by China(Table4,Figure 3).

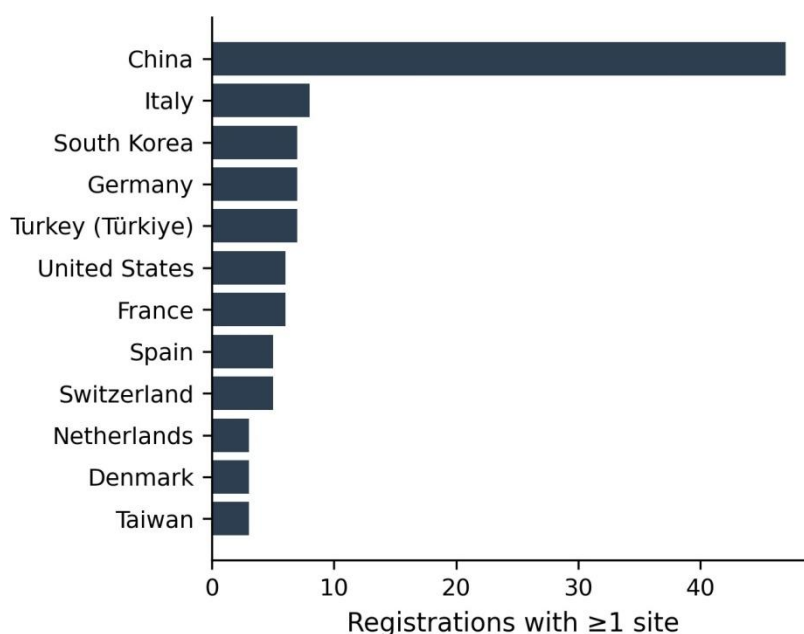
Table2.Distribution by general surgery subdomain, primary analysis.

Subdomain	Allstudies, n(%)	Interventional, n	Randomised,n
Perioperative/general	52(35.9)	18	11
Hepatopancreatobiliary	20(13.8)	4	1
Uppergastrointestinal/bariatric	18(12.4)	2	1
Colorectal	13(9.0)	2	1
Emergency/acute care	10(6.9)	4	3
Breast	10(6.9)	5	4
Hernia/abdominal wall	9(6.2)	1	1
Other/unspecified	9(6.2)	2	2
Endocrine	4(2.8)	0	0
Total	145(100)	38	24

Table4.Countries hosting atleast one recruiting site, primary analysis.

Country	Studies,n	%of locatedstudies
China	47	37.6
Italy	8	6.4
SouthKorea	7	5.6
Germany	7	5.6
Türkiye	7	5.6
UnitedStates	6	4.8
France	6	4.8
Spain	5	4.0
Switzerland	5	4.0
Netherlands	3	2.4
Denmark	3	2.4
Taiwan	3	2.4

Denominator Is the 125 studies (86.2%) listing atleast one location.Studies recruiting in more than one country are counted once per country. Thirty-one countries appeared in total.

**Figure3.Countries hosting atleast one recruiting site,twelve most frequent.**

Discussion

We expected to find a widening gap and did not. Publications and registered trials in surgical artificial intelligence have grown at statistically indistinguishable rates for a decade, and the point estimate favours trials rather than papers. On the hypothesis we set out to test, the field is not deteriorating.

That is the less important finding. The more important one is the level at which the two series have been running in parallel. One hundred and seventy-nine publications for every registered interventional trial, and 284 for every randomized trial, is not a gap that needs to widen in order to be a problem. It has been extreme from the outset and has stayed there. Proportional growth that that ratio means the stock of untested claims expands at the same rate as the literature itself, and the expectation that evaluation will eventually catch up finds no support in these data. Nothing in an eleven-year series suggests convergence.

The starting point deserves emphasis. The first prospective registration of any design appeared in 2018, after 476 papers had already been published on the subject. Three years of literature accumulated before a single study was committed to in advance.

What the registered studies contain compounds this. Three-quarters are observational. Of the interventional minority, a bare majority randomise, only two of 38 declare a phase, and median planned enrolment is 110 patients. Eighty-six per cent are single-centre. For a technology whose characteristic failure is distribution shift, in which performance collapses when a model meets a scanner, a surgeon, a case mix or a population it was not trained on, single-centre evaluation is close to uninformative about deployment. A model validated at one hospital has been shown to work at one hospital. Four studies in eleven years recruited across more than one country. The field is systematically failing to test the single property that determines whether these systems are safe outside the unit that built them.

The reporting figures are the most serious result in this paper. Of 49 completed studies, one has posted results. Not a low proportion but one study. The remaining 48 enrolled patients, consumed institutional Resources and ethics committee time, and have returned nothing to the evidence base after a median of two years. A further 27% of registrations have lapsed into unknown status, so their fate cannot be established at all. Registries exist partly to make non-reporting visible, and the mechanism has worked here exactly as designed: it has made visible a near-total failure. In a field already skewed towards favourable findings, the studies that disappear are disproportionately likely to be those that did not work.

Two structural observations follow. First, industry led three registrations out of 145. Whatever is driving surgical artificial intelligence research, it is not a commercial development pipeline, and the near-absence of industry sponsorship is difficult to reconcile with the volume of claims about imminent clinical deployment. Second, the geography is extraordinarily concentrated, with China alone accounting for 37.6% of located studies. This carries a direct methodological consequence, since ClinicalTrials.gov does not capture trials registered solely with the Chinese Clinical Trial Registry. Our denominator is therefore an undercount, and most plausibly an undercount concentrated in the country contributing most of the activity.

That undercount cuts against our own ratio and we state it plainly rather than defend the number. It does not, however, touch the reporting finding, which is computed within the registry and does not depend on completeness, nor the design findings, which describe the studies that exist rather than those that are missing. Nor does it explain the absence of multinational recruitment, which would be visible in ClinicalTrials.gov whichever additional registry a study used.

What would change this? Three things, in ascending order of difficulty. Journals could require a registration statement for any prospective evaluation of a surgical artificial intelligence system, as most already do for randomised trials, which costs nothing and would make non-registration visible at submission. Ethics committees and funders could adopt DECIDE-AI as the default framework for early clinical evaluation, which would push work towards the small, well-reported prospective studies the field is currently skipping. And the incentive structure that rewards a retrospective performance paper as highly as a multi-centre prospective study would have to change, which is the hardest of the three and the one that actually determines the ratio we measured.

These findings sit alongside work in other specialties, where external validation is sparse and reporting standards poorly met. The surgical picture is worse on the reporting dimension than most published estimates from general medicine, where results-posting rates are typically reported in the tens of percent rather than at 2%.

Limitations

ClinicalTrials.gov is the largest registry but not a complete one. Studies registered exclusively with the Chinese Clinical Trial Registry, the Clinical Trials Registry –India or other national registries are absent, and given the geographical concentration we observed this omission is likely to be substantial and non-random. Our absolute ratio should therefore be read as an upper bound rather than a point estimate. The growth-rate comparison is more robust to this, since it would be biased only if registry coverage changed systematically over time.

Publication counting by title and abstract terms will misclassify in both directions and we did not verify the corpus at record level; the count indexes output rather than enumerating it exactly. The registry search was executed through the web interface, so the exclusion vocabulary was applied to exported records rather than at query time, which is why the sensitivity analysis was pre-planned. Screening was performed by pre-specified vocabulary rather than duplicate manual review. This is a

genuine weakness, since automated screening cannot resolve the borderline cases a surgeon would adjudicate, and the small difference between the two screens (145 against 135) bounds but does not eliminate that uncertainty. Date of first posting is not date of study start, although it is the field most consistently populated. Finally, results posted on a registry is a conservative measure of reporting, because some of these studies will have been published in journals without the registry entry being updated. The true reporting rate is therefore higher than 2%, and establishing how much higher would require a manual publication search that we did not perform.

Conclusion

Between 2015 and 2025 the general surgery artificial intelligence literature and its prospective evaluation grew at the same rate, and the pre-specified hypothesis of a widening gap was not supported. That parallel growth occurred, however, at a ratio of 179 publications per interventional trial and 284 per randomised trial, with no registered study of any design before 2018. Registered work was overwhelmingly observational, single-centre and small; only four studies in eleven years recruited across borders; and of 49 completed studies one has reported results. The problem in this field is not that research is scarce. It is that almost none of it is built to change practice, and almost none of what finishes is ever heard from again.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56.
2. Hashimoto DA, Rosman G, RusD ,Meireles OR. Artificial intelligence in surgery: promises and perils. *Ann Surg.* 2018;268(1):70–6.
3. Maier-Hein L, Vedula SS, Speide IS, et al. Surgical data science for next-generation interventions. *Nat Biomed Eng.* 2017;1:691–6.
4. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ.* 2020;368:m689.
5. WuE, WuK, DaneshjouR, OuyangD, HoDE, ZouJ. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med.* 2021;27(4):582–4.
6. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378.
7. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med.* 2022;28(5):924–33.
8. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med.* 2020;26(9):1364–74.
9. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med.* 2020;26(9):1351–63.
10. 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(5):955–61.
11. Meara JG, Leather AJM, Hagander L, et al. Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. *Lancet.* 2015;386(9993):569–624.
12. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* 2015;350:g7594.
13. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) statement. *Lancet.* 2007;370(9596):1453–7.
14. Ioannidis JPA. Why most published research findings are false. *PLoS Med.* 2005;2(8):e124.