Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Intraoperative failure of the robotic workflow in spine surgery: incidence, failure modes and predictors of conversion to freehand or navigated technique. A systematic review and meta-analysis of proportions

Aashish Ghodke

aashishghodke@gmail.com

Royal Gwent Hospital, Aneurin Bevan University

Andre Chou

University Hospital of Wales, Cardiff and Vale University Health Board

Rishi Agarwal

Royal Gwent Hospital, Aneurin Bevan University

Sanjit Singh

Royal Gwent Hospital, Aneurin Bevan University

Birender Balain

The Robert Jones and Agnes Hunt Orthopaedic Hospital NHS Foundation Trust

Research Article

Keywords: Robot-assisted spine surgery, Conversion to freehand, Robot abandonment, Registration failure, Pedicle screw, Meta-analysis of proportions

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11128619/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations:
No competing interests reported.

Tables 1 to 6 are available in the Supplementary Files section.

Intraoperative failure of the robotic workflow in spine surgery: incidence, failure modes and predictors of conversion to freehand or navigated technique. A systematic review and meta-analysis of proportions

ABSTRACT

Purpose. Robot-assisted spine surgery has been assessed almost entirely through pedicle screw accuracy. Intraoperative conversion to freehand or navigated technique has not been synthesised as a primary outcome. We estimated its incidence, applied a pre-specified failure-mode taxonomy and examined predictors of conversion.

Methods. Systematic review and meta-analysis of proportions, registered on PROSPERO before searching. MEDLINE, Europe PMC and Cochrane CENTRAL were searched on 18 September 2026 using a robot-and-spine strategy with a workflow-failure overlay. Clinical series of ten or more patients reporting a conversion numerator with a matched denominator were eligible; overlapping cohorts were collapsed. Random-effects logit DerSimonian–Laird models estimated per-case any conversion (Grade 2 or 3) and Grade 3 total abandonment, with Clopper–Pearson study intervals and 95% prediction intervals. Series silent on conversion were not treated as reporting zero.

Results. Thirteen independent sources (2,728 robot-attempted cases) entered the per-case models. Pooled any-conversion incidence was 3.6% (95% CI 2.1–6.3) with a 95% prediction interval of 0.5–23.6% (I^2 81%). Most contributing numerators capture total abandonment only, so 3.6% is a lower bound on the primary outcome rather than an estimate of it. Grade 3 abandonment was 2.8% (1.7–4.6; 95% PI 0.6–13.2). Sources including the first institutional cases pooled at 7.1% (4.5–11.0) against 1.4% (0.8–2.6) in later or larger series. First-

generation Mazor platforms pooled higher than third-generation Mazor and floor-mounted ExcelsiusGPS. Registration and tracking accounted for 57% and the instrument–bone interface for 23% of 74 classified mechanisms; that denominator is stated mechanisms and not the 76 case-level conversion events, and a single source contributes 20 of the 74, so these percentages must not be applied to case counts.

Conclusion. A single pooled percentage is not sufficient for procurement planning. The prediction intervals, and not the pooled values of 3.6% and 2.8%, are the planning figures: conversion in a new series is compatible with anything between approximately 1 in 200 cases and 1 in 4, and most of that spread reflects platform generation and operator experience. Certainty in the pooled incidence is very low. Accuracy estimates that exclude converted screws overstate what the device delivers in practice.

Keywords. Robot-assisted spine surgery; Conversion to freehand; Robot abandonment; Registration failure; Pedicle screw; Meta-analysis of proportions

INTRODUCTION

Robot-assisted spinal instrumentation has been evaluated almost entirely through pedicle screw accuracy [1]. Systematic and umbrella reviews have pooled Gertzbein–Robbins grades, facet joint violation, radiation exposure, blood loss and length of stay against freehand and navigated techniques [2–6]. The outcome that determines whether a robot can be used at all in a given case has not been synthesised: how often the robotic workflow fails during surgery, obliging the surgeon to revert to freehand or navigated placement.

Conversion is a within-paper outcome that is not indexed. It rarely appears in a title, is often absent from the abstract and has no MeSH term. Individual series report registration failure, cannula skiving, guide-wire slippage, displacement of the reference array and abandonment of the robot, usually within a single sentence of the complications paragraph [7–13]. High-

volume multicentre series now report abandonment as a named endpoint [14–16], and learning-curve analyses have shown conversion falling after the first thirty cases [8]. These data have not been pooled, no common taxonomy has been applied, and predictors have not been examined.

Units preparing a capital business case need to know whether the robot is abandoned in 1 in 20 cases or 1 in 200; accuracy meta-analyses generally exclude converted screws, removing the most difficult screws and raising reported accuracy; and credentialling depends on whether conversion is a learning-curve effect or is driven by platform and anatomy. This paper reports the incidence synthesis, failure-mode taxonomy and predictor analysis specified in a protocol registered before any search was run.

METHODS

Protocol, registration and reporting

The review follows a protocol registered on PROSPERO before any search was run. Reporting follows PRISMA 2020 [17], PRISMA-S for the search [18] and SWiM for the narrative elements [19]. Certainty was assessed using GRADE adapted for incidence questions [20]. Screening and full-text assessment against locked eligibility criteria were carried out independently by two reviewers, with disagreements resolved by a third. Data were extracted in duplicate.

Outcome definitions

Definitions were locked before extraction because the primary literature does not use consistent terminology. Grade 0 denotes a case completed with robotic guidance throughout and Grade 1 a recoverable interruption (re-registration, re-scan, re-docking, arm repositioning, replanning or re-attachment of the reference array) with robotic technique

retained for all planned screws. Grade 2 denotes partial conversion, in which at least one planned screw is abandoned from the robotic workflow and completed freehand or by navigation while the robot is used for the remainder. Grade 3 denotes total abandonment for the rest of the case.

The primary outcome is the per-case proportion with Grade 2 or Grade 3 failure; Grade 3 is a pre-specified secondary outcome. Grade 1 is expected to be substantially under-captured and was not pooled, and per-case and per-screw denominators were not combined in any model. Each event was also coded to a six-part mechanism taxonomy, from A, registration and tracking, to F, unclassifiable, set out in full in Table 5. Authors' own labels were recorded verbatim but did not determine the grade.

Eligibility criteria

Human clinical studies of robot-assisted spinal surgery were eligible in any region, for any indication, at any age, on any platform with a robotic end-effector, with designs limited to randomised trials, comparative and single-arm cohorts and consecutive series of ten or more patients or fifty or more robotically planned screws. A study had to report conversion, abandonment or workflow failure as a numerator with a matched denominator, including an explicit statement of zero events. Cadaveric, phantom, animal and simulation studies, case reports, reviews, navigation without a robotic end-effector, and studies in which the robot-arm denominator could not be separated were excluded. No language restriction was applied; the date limit was 1 January 2005.

Information sources and search

Searches were run on 18 September 2026 in MEDLINE via PubMed (n = 1,902), Europe PMC (n = 1,406) and Cochrane CENTRAL with the randomised-trial subset via PubMed (n = 182). OpenAlex was queried as a proxy for Embase- and Scopus-indexed journals but

returned HTTP 429, a protocol deviation likely to have under-sampled Embase-only records; CNKI, Wanfang and KoreaMed were not searched. Identification used protocol Strand B, combining robot-and-spine terms with a workflow-failure overlay, with a hand-assembled seed set and citation chasing from included papers and the major accuracy reviews [2–6]. Strand A (robot AND spine, no outcome terms) was not run, which over-samples series that named conversion and makes a literature-wide reporting-completeness percentage impossible to calculate. Corresponding authors of silent series were not contacted and no PRESS review was obtained. Final strategies are in the supplementary file.

Selection, extraction and management of overlap

Records were deduplicated by PMID and DOI, and 96 clinical reports were read in full for a matched conversion numerator. Overlapping cohorts were identified by authorship, institution, platform, study dates and sample size, and the largest or most completely reported source retained: Hu 2013 within Hu 2014 [7,8]; Wallace 2020, reporting the same first 106 ExcelsiusGPS cases as Jain 2019 [9]; three Lee reports overlapping the four-centre series of Lee 2022 (n = 722), which was retained for the primary models, its matched platform arms being used only in the platform subgroup [14–16,21]; and Pojskić 2021 within Pojskić 2025 [22,23]. PRoGRSS may share centres with the Lee Stealth cohort and was dropped in a sensitivity analysis [24]. Table 2 maps the overlap. An explicit statement of zero events with a matched denominator was eligible; a missing statement was not treated as zero and was not imputed. Studies with a numerator but no matched case denominator were used only in the per-screw display.

Risk of bias

RoB 2, ROBINS-I and the JBI case-series checklist [25–27] do not address ascertainment of an unindexed outcome. A bespoke domain was applied to every source, recording whether

conversion was pre-specified, whether a definition was given, whether the cohort was consecutive, whether collection was prospective, whether the denominator was explicit, and whether industry relationships were declared. Industry relationships were recorded separately from the risk-of-bias judgement.

Synthesis

The protocol specified a random-intercept logistic generalised linear mixed model (GLMM) as the primary mixed model and logit DerSimonian–Laird for reporting [28]. Analyses used Python 3. The reported model is inverse-variance DerSimonian–Laird on the logit scale with a 0.5 continuity correction confined to zero-event and all-event studies; study-level intervals are Clopper–Pearson and uncorrected [29,30]. A Breslow–Clayton penalised quasi-likelihood GLMM is reported as the protocol-specified mixed model, and the Freeman–Tukey double-arcsine transformation as sensitivity only, because back-transformation at the harmonic mean sample size is a recognised source of bias and no conclusion rests on it [31].

For each model we report the pooled proportion, its 95% confidence interval, τ^2, I^2, Q with its p value, and a 95% Higgins–Thompson prediction interval [32]. The protocol specified that if I^2 exceeded 85% and subgroups did not explain it, the forest plot would be shown without a pooled diamond. Pre-specified subgroups were platform family, inclusion of the learning curve, design, outcome pre-specification, era, geography and industry funding. Leave-one-out analysis and exclusion of industry-related and overlapping-alternative sources were pre-specified sensitivity analyses.

RESULTS

Study selection

After deduplication, 2,022 unique records were screened; 96 were read in full and 24 provided a matched count. After overlap mapping, 13 independent sources contributing 2,728 robot-attempted cases entered the per-case models [8–13,16,22,24,33–36]. Six sources contributing 2,745 planned screws entered the per-screw display [9,12,37–40]. Lonjon 2016 reported four of 40 screws placed manually after failed robotic assistance but did not state how many cases these occupied, so only the per-screw figure was used [37]. Devito 2010 and van Dijk 2015, the latter having excluded aborted cases, could not yield an unbiased rate and were not pooled [41,42], and large accuracy series silent on conversion were not treated as reporting zero [43,44]. Figure 1 shows the PRISMA flow.

Characteristics of included sources

The 13 sources were published between 2012 and 2026 and cover eight platform designations from four manufacturers. One is a randomised trial [12], two are prospective non-randomised series [24,34] and ten are retrospective. Ten pre-specified conversion or abandonment as an outcome and eight include first robotic cases. Two have declared industry relationships or are industry-supported registries [8,24]. Seven sources are from the United States and three from Germany, with one each from Italy, China and the United Kingdom, and case numbers range from 30 to 722. Tables 1 to 3 give the design, platform, numerator and verbatim definition for each source, the overlap map and the ascertainment domain.

Two denominators are narrower than the cohort reported elsewhere in the same source and are carried through as published. Baig Mirza 2026 reports 2 of 39 documented cases rather than all 50 consecutive patients, because checklist fields were introduced progressively [33]; Singh 2025 reports 2 of 56 drawn from 85 consecutive operations after excluding those

without 90-day follow-up [36]. If conversion can occur in the excluded cases, both rates are conditioned on documentation or follow-up rather than on exposure and are likely upper bounds at those units.

Any conversion per case

Seventy-six conversion events occurred in 2,728 robot-attempted cases. The pooled proportion was 3.6% (95% CI 2.1–6.3), with high heterogeneity (I^2 81%, τ^2 0.82, Q 63.2 on 12 degrees of freedom, $p < 0.001$) and a 95% prediction interval of 0.5–23.6%. The prediction interval rather than the point estimate describes what a unit should expect: in a future series drawn from the same population of studies, conversion of between roughly 1 in 200 and 1 in 4 cases is compatible with these data. Study-level Clopper–Pearson intervals ranged from 0 of 40 (Yan 2022) and 0.6% (PRoGRSS 2026) to 7 of 30 (Ringel 2012, 9.9–42.3%). In Figure 2 a dagger marks sources reporting Grade 3 only, for which the contribution is a lower bound; because most numerators are of that kind, the pooled 3.6% is itself a lower bound on the primary outcome and should be quoted as one. The penalised quasi-likelihood GLMM, which accommodates zero-event studies without a continuity correction, gave 3.4% (1.9–6.0) and Freeman–Tukey 3.3% (1.6–5.4); the point estimate was stable across transformations and the prediction interval remained wide under all of them.

Grade 3 total abandonment

Pooled Grade 3 incidence was 2.8% (95% CI 1.7–4.6; 95% PI 0.6–13.2; I^2 69%, τ^2 0.50). Ringel 2012 reported Grade 2 events only, in which a cannula entered soft tissue and the robot was retained for the remaining screws, and enters this model as 0 of 30. Jain 2019 reported five Grade 3 and two Grade 2 cases, so the Grade 3 model uses 5 of 106. Pojskić 2025 requires the same distinction: its nine abandonments among 109 robot-attempted surgeries are partial, the arm being replaced by a navigated drill guide for the affected

trajectories while the remainder of each construct was completed robotically, which meets the Grade 2 definition. Three further operations at that unit, planned for robotic placement but in which the arm was never used because of installation or restart problems arising before any screw was placed, lie outside the 109-surgery denominator and are counted in neither model. The prediction interval of 0.6–13.2% answers the procurement question in its narrowest form, namely how often the robot leaves the field altogether (Fig. 3).

Per-screw conversion

Six sources provided a matched screw denominator, ranging from Jain 2019 (5 of 636, 0.8%, excluding five cases aborted before any screw was placed) to Lonjon 2016 (4 of 40, 10.0%), with Ringel 2012, Avrumova 2021, Oh 2024 and Abel 2023 between them (Table 4). I^2 was 81%, below the pre-specified threshold of 85%, so the pooled estimate of 4.4% (2.7–7.1) is permissible. Figure 4 omits a pooled diamond because the observed range, approximately 1 in 125 to 1 in 10 screws, is the more useful statement and because the Jain denominator is biased downwards by the exclusion of aborted cases.

Platform family

Platform generation and operator experience are confounded in these data, and the platform subgroups follow the same pattern as the experience subgroups below. First-generation Mazor (SpineAssist) pooled at 13.1% (95% CI 3.9–35.8) across two sources, although I^2 with two studies is uninformative, and Renaissance at 6.5% with I^2 of 95%, which is not interpretable as a platform effect. ExcelsiusGPS pooled at 2.9% (1.1–7.5; I^2 58%) across four sources, third-generation Mazor at 1.5% (0.5–4.7), and Cirq, two European implementation series, at 7.2% (4.0–12.6; I^2 0%). The single TiRobot source reported no abandonment in 40 cases (0.0–8.8) [34]. In Figure 5 the matched platform arms of Lee 2021 and Lee 2023 replace the mixed-platform Lee 2022 cohort, so the subtotals are drawn from a partly

different analysis set and do not sum to the primary denominator of 2,728 cases; the figure legend and Supplementary Table S11 give every platform arm with its numerator, denominator and model, and reconcile the two analysis sets. Mounting type could not be separated from platform.

Small-study effects

Funnel inspection did not suggest small-study effects that would change the interpretation (Fig. 6). Funnel asymmetry is unreliable for proportions, and the dominant bias in this literature is selective reporting of conversion, which a funnel restricted to the series that did report it cannot detect.

Failure modes

Seventy-four events with a stated mechanism were coded (Fig. 7, Table 5). This is not a complete census and its denominator is stated mechanisms rather than converted cases. Two sources contribute at event rather than case level: Ringel 2012, ten skived cannulae in seven cases, and Wandvik 2024, 20 classified events in five abandoned attempts. Lee 2022 (18 events), Baig Mirza 2026 (two) and two of the nine Pojskić 2025 abandonments give no mechanism and are absent. The percentages below are therefore proportions of classified mechanisms, not of the 76 case-level conversion events, and are sensitive to Wandvik 2024 alone supplying 20 of the 74.

Category A, registration and tracking, accounted for 42 events (57%), the largest contributions being 12 registration failures and software crashes in Hu 2014, 14 tracking events in Wandvik 2024 and all four PRoGRSS abandonments. Category B, the instrument–bone interface, accounted for 17 (23%), chiefly ten skived cannulae in Ringel 2012 and six skive or soft-tissue events in Wandvik 2024. Category D, hardware and software, accounted for 14 (19%), including seven Cirq-arm failures in Pojskić 2025, five attributed to ergonomic

problems and two to technical problems. Category C accounted for one event, a conversion for severe scoliosis. Categories E and F were empty, which more likely reflects reporting practice than an absence of team or organisational failure. Table 5 gives the source-by-source breakdown.

Sensitivity analyses

Removing PRoGRSS moved the pooled any-conversion estimate from 3.6% to 4.3% (2.6–7.3), and removing both industry-related sources gave 4.0% (2.2–7.2). Leave-one-out estimates ranged from 3.1% when Ringel 2012 was omitted to 4.3% when PRoGRSS was omitted, so no single source accounts for the finding (Fig. 8). The prediction interval remained an order of magnitude wide in every sensitivity analysis.

Learning curve and operator experience

The pre-specified subgroup that shifts the estimate is experience. Eight sources including the first institutional series pooled at 7.1% (95% CI 4.5–11.0; 95% PI 2.0–22.0; I^2 52%). Five later or larger sources pooled at 1.4% (0.8–2.6; 95% PI 0.3–7.7; I^2 49%), the lowest heterogeneity of any subgroup examined. Hu 2014 provides the within-surgeon demonstration: conversion to manual placement occurred in 17% of cases 1 to 30 and in 4 to 8% thereafter, while malposition of the screws the robot did place remained between 0 and 1.4% across all five blocks of 30 [8]. Lee 2022 reported abandonment falling from 10.3% in 2016 to 1.1% in 2019 [16]. Figures 9 and 9b show both datasets and the pooled experience split.

Certainty of evidence

Certainty for the incidence of any conversion is very low (Table 6). Risk of bias is serious because most sources are retrospective and ascertainment often incidental; inconsistency is

serious; imprecision is serious once the prediction interval is taken as the estimand of interest; and publication and selective-reporting bias is very serious, since Strand A was not run, silent series were not imputed, and the sample consists of authors who chose to report the outcome. Indirectness is not serious. The Grade 3 estimate is also of very low certainty. The learning-curve split is low rather than very low, because inconsistency falls and two independent designs point in the same direction. No estimate reaches moderate or high certainty.

DISCUSSION

Principal findings

The pooled any-conversion incidence of 3.6% and the pooled Grade 3 incidence of 2.8% describe the studies analysed here. Neither describes the next unit to install a system, and neither should enter a business case as a planning figure: in a future series resembling those analysed, the proportion of robot-attempted cases losing at least one screw from the robotic workflow is compatible with values between 0.5% and 23.6%, and total abandonment with values between 0.6% and 13.2%.

Operator experience accounts for much of that spread: 7.1% in series that include the first institutional cases against 1.4% in later or larger series, where heterogeneity fell to 49%. The single-surgeon SpineAssist series of Hu shows the same pattern within one practice, while malposition among completed robotic screws did not change [8]. Conversion therefore behaves as a learning-curve and platform-generation phenomenon whereas malposition does not, and credentialling should be designed around that distinction.

Registration technique and reference-array management, which together account for more than half of classified mechanisms, are amenable to training, whereas skiving depends on platform and technique. The randomised trial of Ringel, using a bed-mounted first-generation hexapod, recorded at least one converted screw in 23% of cases together with robotic

accuracy inferior to freehand [12]; later floor- and table-mounted systems with integrated navigation have not reproduced that, which reflects platform generation rather than a reason to disregard the trial. The two Cirq implementation series [22,33] sit closer to early-generation rates, consistent with early institutional experience rather than a mature platform effect.

Implications for the accuracy literature

Every accuracy meta-analysis cited above shares the same hidden selection: screws the robot could not place are removed from the denominator and the remainder graded, and the resulting 97 to 98% Grade A or B is then set against freehand series that have no equivalent exclusion. Almost no included study reported Gertzbein–Robbins grades for converted screws separately, so only a directional statement is possible. Hu 2013 aborted 110 of 1,085 planned screws (10%) and reported 98.9% accuracy among the 960 placed [7]; Ringel 2012 converted 10 of 146 screws and reported 85% Grade A or B among those implanted, already lower than freehand in that trial [12]; van Dijk 2015 excluded aborted cases before calculating 97.9% [42]. If a converted-screw fraction of 5 to 10% is counted as not Grade A or B, a published accuracy of 98% corresponds to an intention-to-treat accuracy of 88 to 93%. That arithmetic illustrates direction and approximate size; it is not a pooled estimate and is not carried into the abstract or conclusion. Units comparing a robot with freehand technique on accuracy alone are comparing two different analysis sets.

Relation to device-surveillance data

Analyses of the FDA Manufacturer and User Facility Device Experience (MAUDE) database for Mazor X and ExcelsiusGPS describe precision issues as the majority of reported adverse events, with robot abandonment in about a third and mean delays of the order of an hour [45]. MAUDE provides a numerator without a denominator and was pre-specified as narrative

context only, but it is consistent with the dominance of Categories A and B in the clinical series and a reminder that series which never mention conversion are not a safe source of zeros.

Limitations

Strand A was not run, so reporting completeness could not be estimated and the included sources are those whose authors named the outcome; no literature-wide completeness percentage is offered, because none can be calculated from a sample assembled in this way. OpenAlex records were not retrieved and CNKI and Wanfang were not searched, so some Embase- and Scopus-only records may have been missed and TiRobot experience rests on a single 40-case source. Most any-conversion numerators capture Grade 3 only, so Grade 2 is under-captured and 3.6% is a lower bound. Body mass index, bone quality, revision status and surgeon volume were too sparsely reported to support the pre-specified meta-regression. PRoGRSS may overlap the Lee Stealth cohort, although removing it does not change the prediction interval. No corresponding author was contacted and no PRESS review was obtained.

Two further limitations concern what is being pooled. The included platforms are not equivalent devices: shared-control systems, supervised end-effector positioning systems and passive navigated alignment arms differ in what abandonment means and in what remains available once the robotic component is withdrawn. The failure-mode census is also not on the same footing as the incidence models, its denominator being stated mechanisms, with two sources contributing event-level counts and one supplying 20 of the 74 classified events, so the taxonomy percentages must not be applied to the 76 case-level conversion events. None of this argues for withholding the prediction interval. It argues for reading the point estimate

as a description of a reporting-selected sample and the prediction interval as the range available for planning.

Implications for practice and research

For procurement, the first year of use should be modelled at a conversion rate near 7% rather than at the mature rate of approximately 1%, and total abandonment at approximately 3% with an upper range in double figures on a first-generation platform or during the learning curve. The upper limits of the prediction intervals, and not the pooled point estimates, should set the contingency. For credentialling, these data support structured proctoring through approximately the first thirty cases, and do not support treating conversion as a random device fault thereafter, at least on current third-generation and floor-mounted platforms.

For trialists and registry designers, conversion should be reported as a pre-specified graded endpoint with a matched denominator, a stated mechanism, and an explicit statement of whether converted screws enter the accuracy analysis; where a denominator is restricted by documentation or follow-up, that restriction should be stated alongside the rate. Box 1 sets out a minimum reporting set. Series that remain silent on conversion should be treated as uninformative rather than as reporting zero, and an intention-to-treat accuracy re-analysis counting converted screws as failures is overdue.

CONCLUSION

Intraoperative conversion from robotic to freehand or navigated technique occurs often enough to belong in a capital business case and, in mature series, infrequently enough to be missed if it is not specifically sought. Across 13 independent sources the pooled any-conversion incidence was 3.6% (95% prediction interval 0.5 to 23.6) and total abandonment 2.8% (0.6 to 13.2). Those pooled percentages describe the studies analysed, rest on very low certainty evidence, and are lower bounds to the extent that most numerators capture total

abandonment only; the prediction intervals are the figures a unit should plan against, and most of the spread is attributable to operator experience and platform generation. Registration and tracking failures, followed by skiving at the instrument–bone interface, dominate among the series that state a mechanism. Accuracy figures that exclude the screws the robot could not place do not describe what the device delivers in practice.

REFERENCES

1. Ghasem A, Sharma A, Greif DN et al (2018) The arrival of robotics in spine surgery: a review of the literature. Spine (Phila Pa 1976) 43:1670–1677. https://doi.org/10.1097/BRS.0000000000002695
2. Fatima N, Massaad E, Hadzipasic M et al (2021) Safety and accuracy of robot-assisted placement of pedicle screws compared to conventional free-hand technique: a systematic review and meta-analysis. Spine J 21(2):181–192. https://doi.org/10.1016/j.spinee.2020.09.007
3. Peng YN, Tsai LC, Hsu HC, Kao CH (2020) Accuracy of robot-assisted versus conventional freehand pedicle screw placement in spine surgery: a systematic review and meta-analysis of randomized controlled trials. Ann Transl Med
4. Zhou LP, Zhang RJ, Sun YL et al (2023) Accuracy of pedicle screw placement and four other clinical outcomes of robotic guidance versus freehand in spine surgery: an updated systematic review and meta-analysis. Global Spine J
5. Matur AV, Palmisciano P, Duah HO et al (2023) Robotic and navigated pedicle screws are safer and more accurate than fluoroscopic freehand screws: a systematic review and meta-analysis. Spine J 23(2):197–208. https://doi.org/10.1016/j.spinee.2022.10.006

6. Li J, Fang Y, Jin Z et al (2024) The impact of robot-assisted spine surgery on pedicle screw accuracy, radiation, and outcomes: an updated systematic review and meta-analysis. Spine J

7. Hu X, Ohnmeiss DD, Lieberman IH (2013) Robotic-assisted pedicle screw placement: lessons learned from the first 102 patients. Eur Spine J 22(3):661–666. https://doi.org/10.1007/s00586-012-2499-1

8. Hu X, Lieberman IH (2014) What is the learning curve for robotic-assisted pedicle screw placement in spine surgery? Clin Orthop Relat Res 472(6):1839–1844. https://doi.org/10.1007/s11999-013-3291-1

9. Jain D, Manning J, Lord E et al (2019) Initial single-institution experience with a novel robotic-navigation system for thoracolumbar pedicle screw and pelvic screw placement with 643 screws. Int J Spine Surg 13:459–463. https://doi.org/10.14444/6062

10. Benech CA, Perez R, Benech F et al (2020) Navigated robotic assistance results in improved screw accuracy and positive clinical outcomes: an evaluation of the first 54 cases. J Robot Surg 14:431–437. https://doi.org/10.1007/s11701-019-01007-z

11. Keric N, Doenitz C, Haj A et al (2017) Evaluation of robot-guided minimally invasive implantation of 2067 pedicle screws. Neurosurg Focus 42:E11. https://doi.org/10.3171/2017.2.FOCUS16552

12. Ringel F, Stüer C, Reinke A et al (2012) Accuracy of robot-assisted placement of lumbar and sacral pedicle screws: a prospective randomized comparison to conventional freehand screw implantation. Spine (Phila Pa 1976) 37:E496–E501. https://doi.org/10.1097/BRS.0b013e31824b7767

13. Wandvik C, Greil ME, Colby S et al (2024) Limitations of current robot-assisted pedicle screw insertion systems. Neurosurg Focus 57:E14. https://doi.org/10.3171/2024.9.FOCUS24506

14. Lee NJ, Zuckerman SL, Chaudhary SB et al (2021) Is there a difference between navigated and non-navigated robot cohorts in robot-assisted spine surgery? A multicenter, propensity-matched analysis of 2,800 screws and 372 patients. Spine J 21:1504–1512. https://doi.org/10.1016/j.spinee.2021.05.015

15. Lee NJ, Zuckerman SL, Buchanan IA et al (2023) Is there a difference in screw accuracy, robot time per screw, robot abandonment, and radiation exposure between the Mazor X and the Renaissance? A propensity-matched analysis of 1179 robot-assisted screws. Global Spine J 13:1286–1292. https://doi.org/10.1177/21925682211029867

16. Lee NJ, Lombardi JM, Boddapati V et al (2022) A multicenter study of the 5-year trends in robot-assisted spine surgery outcomes and complications. J Spine Surg 8:9–20. https://doi.org/10.21037/jss-21-102

17. Page MJ, McKenzie JE, Bossuyt PM et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71

18. Rethlefsen ML, Kirtley S, Waffenschmidt S et al (2021) PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 10:39. https://doi.org/10.1186/s13643-020-01542-z

19. Campbell M, McKenzie JE, Sowden A et al (2020) Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ 368:l6890. https://doi.org/10.1136/bmj.l6890

20. Munn Z, Moola S, Lisy K et al (2015) Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. Int J Evid Based Healthc 13:147–153. https://doi.org/10.1097/XEB.0000000000000054

21. Lee NJ, Buchanan IA, Boddapati V et al (2021) Do robot-related complications influence 1 year reoperations and other clinical outcomes after robot-assisted lumbar arthrodesis? A multicenter assessment of 320 patients. J Orthop Surg Res 16:320. https://doi.org/10.1186/s13018-021-02452-z

22. Pojskić M, Bopp M, Alwakaa O et al (2025) Robotic-guided spine surgery: implementation of a system in routine clinical practice—an update. J Clin Med 14(13):4463. https://doi.org/10.3390/jcm14134463

23. Pojskić M, Bopp M, Nimsky C et al (2021) Initial intraoperative experience with robotic-assisted pedicle screw placement with Cirq® robotic alignment: an evaluation of the first 70 screws. J Clin Med 10(24):5725. https://doi.org/10.3390/jcm10245725

24. Orosz LD, Poulter GT, Haines CM et al (2026) Comprehensive outcomes following navigated robotics in thoracolumbar spine surgery: the PRoGRSS final analysis. Global Spine J. Epub ahead of print 30 January 2026. https://doi.org/10.1177/21925682261421190

25. Sterne JAC, Savović J, Page MJ et al (2019) RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 366:l4898. https://doi.org/10.1136/bmj.l4898

26. Sterne JA, Hernán MA, Reeves BC et al (2016) ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 355:i4919. https://doi.org/10.1136/bmj.i4919

27. Munn Z, Barker TH, Moola S et al (2020) Methodological quality of case series studies: an introduction to the JBI critical appraisal tool. JBI Evid Synth 18:2127–2133. https://doi.org/10.11124/JBISRIR-D-19-00099

28. DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Control Clin Trials 7:177–188. https://doi.org/10.1016/0197-2456(86)90046-2

449 29. Nyaga VN, Arbyn M, Aerts M (2014) Metaprop: a Stata command to perform meta-
450 analysis of binomial data. Arch Public Health 72:39. https://doi.org/10.1186/2049-
451 3258-72-39
452 30. Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the
453 case of the binomial. Biometrika 26:404–413. https://doi.org/10.1093/biomet/26.4.404
454 31. Barendregt JJ, Doi SA, Lee YY et al (2013) Meta-analysis of prevalence. J Epidemiol
455 Community Health 67:974–978. https://doi.org/10.1136/jech-2013-203104
456 32. Higgins JPT, Thompson SG, Spiegelhalter DJ (2009) A re-evaluation of random-effects
457 meta-analysis. J R Stat Soc Ser A 172:137–159. https://doi.org/10.1111/j.1467-
458 985X.2008.00552.x
459 33. Baig Mirza A, Rauf W, Muhammad IM et al (2026) Robotic-assisted spinal
460 instrumentation from C1 to S1: experience of a UK neurosurgical tertiary referral
461 centre. J Clin Med 15(15):5905. https://doi.org/10.3390/jcm15155905
462 34. Yan K, Zhang Q, Tian W (2022) Comparison of accuracy and safety between second-
463 generation TiRobot-assisted and free-hand thoracolumbar pedicle screw placement.
464 BMC Surg 22:214. https://doi.org/10.1186/s12893-022-01667-z
465 35. Shahi P, Maayan O, Shinn D et al (2023) Floor-mounted robotic pedicle screw placement
466 in lumbar spine surgery: an analysis of 1,050 screws. Neurospine 20:577–586.
467 https://doi.org/10.14245/ns.2346210.105
468 36. Singh E, Cardinal T, Stillman K et al (2025) Clinical outcomes and complications of
469 robotic-assisted spine surgery using the ExcelsiusGPS system: a single-surgeon case
470 series. J Robot Surg
471 37. Lonjon N, Chan-Seng E, Costalat V et al (2016) Robot-assisted spine surgery: feasibility
472 study through a prospective case-matched analysis. Eur Spine J 25:947–955.
473 https://doi.org/10.1007/s00586-015-3758-8

38. Avrumova F, Morse KW, Heath M et al (2021) Evaluation of K-wireless robotic and navigation assisted pedicle screw placement in adult degenerative spinal surgery: learning curve and technical notes. J Spine Surg 7:141–154. https://doi.org/10.21037/jss-20-687

39. Oh BK, Son DW, Lee SH et al (2024) A single-center experience of robotic-assisted spine surgery in Korea: analysis of screw accuracy, potential risk factor of screw malposition and learning curve. J Korean Neurosurg Soc 67:54–64. https://doi.org/10.3340/jkns.2023.0121

40. Abel F, Avrumova F, Goldman SN et al (2023) Robotic-navigated assistance in spine surgery. Bone Joint J 105-B:543–550. https://doi.org/10.1302/0301-620X.105B5.BJJ-2022-1321.R1

41. Devito DP, Kaplan L, Dietl R et al (2010) Clinical acceptance and accuracy assessment of spinal implants guided with SpineAssist surgical robot: retrospective study. Spine (Phila Pa 1976) 35:2109–2115. https://doi.org/10.1097/BRS.0b013e3181d323ab

42. van Dijk JD, van den Ende RP, Stramigioli S et al (2015) Clinical pedicle screw accuracy and deviation from planning in robot-guided spine surgery. Spine (Phila Pa 1976) 40:E986–E991. https://doi.org/10.1097/BRS.0000000000000960

43. Hyun SJ, Kim KJ, Jahng TA et al (2017) Minimally invasive robotic versus open fluoroscopic-guided spinal instrumented fusions: a randomized controlled trial. Spine (Phila Pa 1976) 42:353–358. https://doi.org/10.1097/BRS.0000000000001778

44. Han X, Tian W, Liu Y et al (2019) Safety and accuracy of robot-assisted versus fluoroscopy-assisted pedicle screw insertion in thoracolumbar spinal surgery: a prospective randomized controlled trial. J Neurosurg Spine 30:615–622. https://doi.org/10.3171/2018.10.SPINE18487

45. Schneider D, Brown EDL, Elsamadicy AA et al (2026) The hidden cost of robotic spine surgery: real-world adverse events cause 58-minute delays and undermine economic viability. Spine J 26:189–198. https://doi.org/10.1016/j.spinee.2025.06.011

Figure legends

Intraoperative failure of the robotic workflow in spine surgery: incidence, failure modes and predictors of conversion to freehand or navigated technique

PROSPERO CRD420261506933. Figures 1 to 9b. High-resolution PNG and PDF files accompany this submission.

Fig. 1 PRISMA 2020 flow diagram

Identification used protocol Strand B, an outcome-term overlay applied to a robot-and-spine strategy, together with a known-item seed set and citation chasing across MEDLINE via PubMed, Europe PMC and Cochrane CENTRAL on 18 September 2026. OpenAlex was attempted and returned HTTP 429, so those records were not retrieved. Strand A (robot AND spine, with no outcome terms) was not run. Series silent on conversion were not imputed as zero. Lonjon 2016, Avrumova 2021, Abel 2023 and Oh 2024 appear in the per-screw display only, because their per-case denominator is unmatched. Devito 2010 and van Dijk 2015 could not yield an unbiased rate and were not pooled; van Dijk 2015 excluded aborted cases from its published denominator.

Fig. 2 Forest plot of any reported conversion (Grade 2 or 3) per robot-attempted case

Random-effects DerSimonian–Laird model on the logit scale. Squares are study proportions, horizontal bars are Clopper–Pearson 95% confidence intervals and square size is proportional to inverse-variance weight. The crimson diamond is the pooled estimate with its 95% confidence interval. The gold bar is the 95% Higgins–Thompson prediction interval, which is the figure a unit should use for planning. A dagger marks sources that reported Grade 3 only, for which the contribution is a lower bound on the primary outcome. k = 13, 76 events in 2,728 cases; pooled 3.6% (95% CI 2.1–6.3); 95% prediction interval 0.5–23.6%; I^2 81%, τ^2 0.82.

Fig. 3 Forest plot of Grade 3 total abandonment per robot-attempted case

Model as for Figure 2, restricted to total abandonment of the robotic workflow for the remainder of the case. Ringel 2012 enters as 0 of 30, having reported Grade 2 events only. Jain 2019 enters as 5 of 106. Pooled 2.8% (95% CI 1.7–4.6); 95% prediction interval 0.6–13.2%; I^2 69%.

Fig. 4 Per-screw conversion, display only

Six independent sources with a matched screw denominator. I^2 of 81% sits just below the pre-specified threshold of 85%, so a pooled estimate is reported in Table 4, but the forest is drawn without a diamond because the observed range, from 0.8% in Jain 2019 to 10% in Lonjon 2016, is the more useful statement. Jain 2019 is biased downwards because the five aborted cases are absent from the 636-screw denominator. Avrumova 2021 counts freehand conversions only (21 of 364); a further 17 screws converted to K-wire technique are not included.

Fig. 5 Subgroup forest plot by platform family

Navy diamonds are family subtotals. The crimson diamond and gold bar are the overall logit DerSimonian–Laird estimate and prediction interval from the primary non-overlapping set. In this figure only, the matched Lee 2021 Mazor X and Stealth arms and the Lee 2023 Renaissance arm replace the mixed-platform Lee 2022 cohort, so the subtotals are drawn from a partly different analysis set and do not sum to the primary denominator of 2,728 cases. First-generation Mazor sits well above floor-mounted ExcelsiusGPS and third-generation Mazor. Cirq, represented by two European implementation series, pools at 7.2% with I^2 of 0%. Platform generation is confounded with operator experience. Every source and arm entering each subgroup is listed with its numerator and denominator in supplementary Table S11.

Fig. 6 Funnel plot on the logit scale

Funnel asymmetry is unreliable for proportions and cannot detect the dominant bias in this literature, which is selective reporting of conversion as an outcome. Series that never mention conversion do not appear in this plot.

Fig. 7 Failure-mode taxonomy among 74 classified mechanisms

Not a complete census, and the denominator is stated mechanisms rather than converted cases. Only events with a stated mechanism are coded. Registration and tracking account for 42 of 74 (57%), the instrument–bone interface for 17 of 74 (23%), hardware and software for 14 of 74 (19%) and anatomy for 1 of 74 (1%). Categories E and F were empty among classified events. Wandvik 2024 contributes 20 of the 74 because that paper set out to classify limitations. Lee 2022 and Baig Mirza 2026 reported abandonment without a mechanism and are absent.

Fig. 8 Leave-one-out sensitivity analysis for any conversion

Each row is the logit DerSimonian–Laird pooled proportion after omission of the named source. Estimates run from 3.1% when Ringel 2012 is omitted to 4.3% when PRoGRSS is omitted. No single source accounts for the finding, and the prediction interval remains an order of magnitude wide under every omission.

Fig. 9 Learning curve within a single surgeon and across a multicentre series

Left panel: Hu 2014, single surgeon, SpineAssist. Conversion to manual placement occurred in 17% of cases 1 to 30 and in 4 to 8% thereafter, while malposition of the screws the robot did place remained between 0 and 1.4% across all five blocks of 30. Conversion behaves as a learning-curve phenomenon; malposition among completed robotic screws does not. Right panel: Lee 2022 (n = 722) reported abandonment falling from 10.3% in 2016 to 1.1% in 2019.

Fig. 9b Pooled any conversion by inclusion of the learning curve

Random-effects DerSimonian–Laird estimates for the two pre-specified experience subgroups. Eight sources including the first institutional series pool at 7.1% (95% CI 4.5–11.0; 95% prediction interval 2.0–22.0; I^2 52%). Five later or larger sources pool at 1.4% (95% CI 0.8–2.6; 95% prediction interval 0.3–7.7; I^2 49%), the lowest heterogeneity of any subgroup examined.

FIGURES

Intraoperative failure of the robotic workflow in spine surgery

Figure 1. PRISMA 2020 flow diagram.

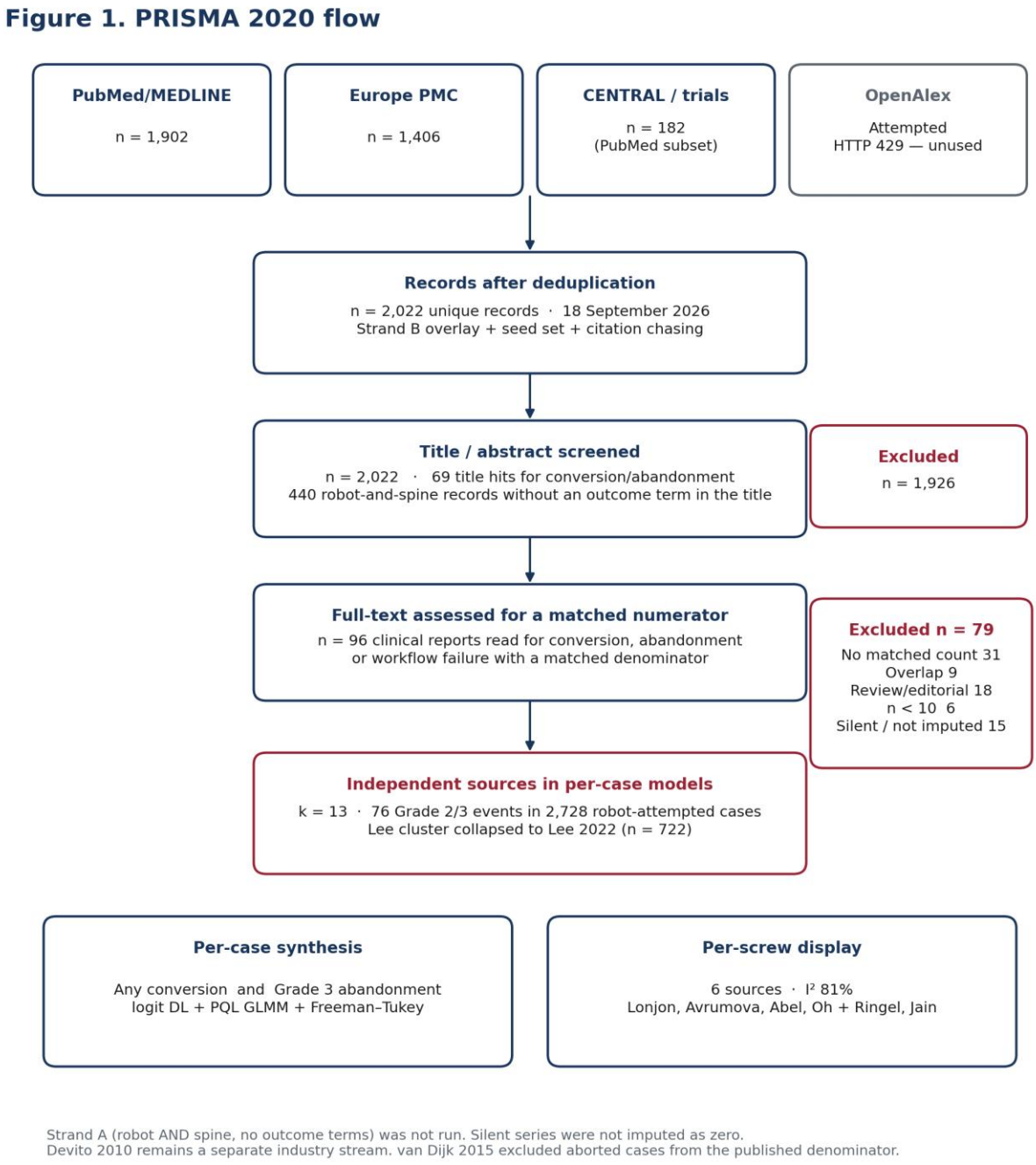


Figure 2. Any conversion (Grade 2 or 3) per robot-attempted case.

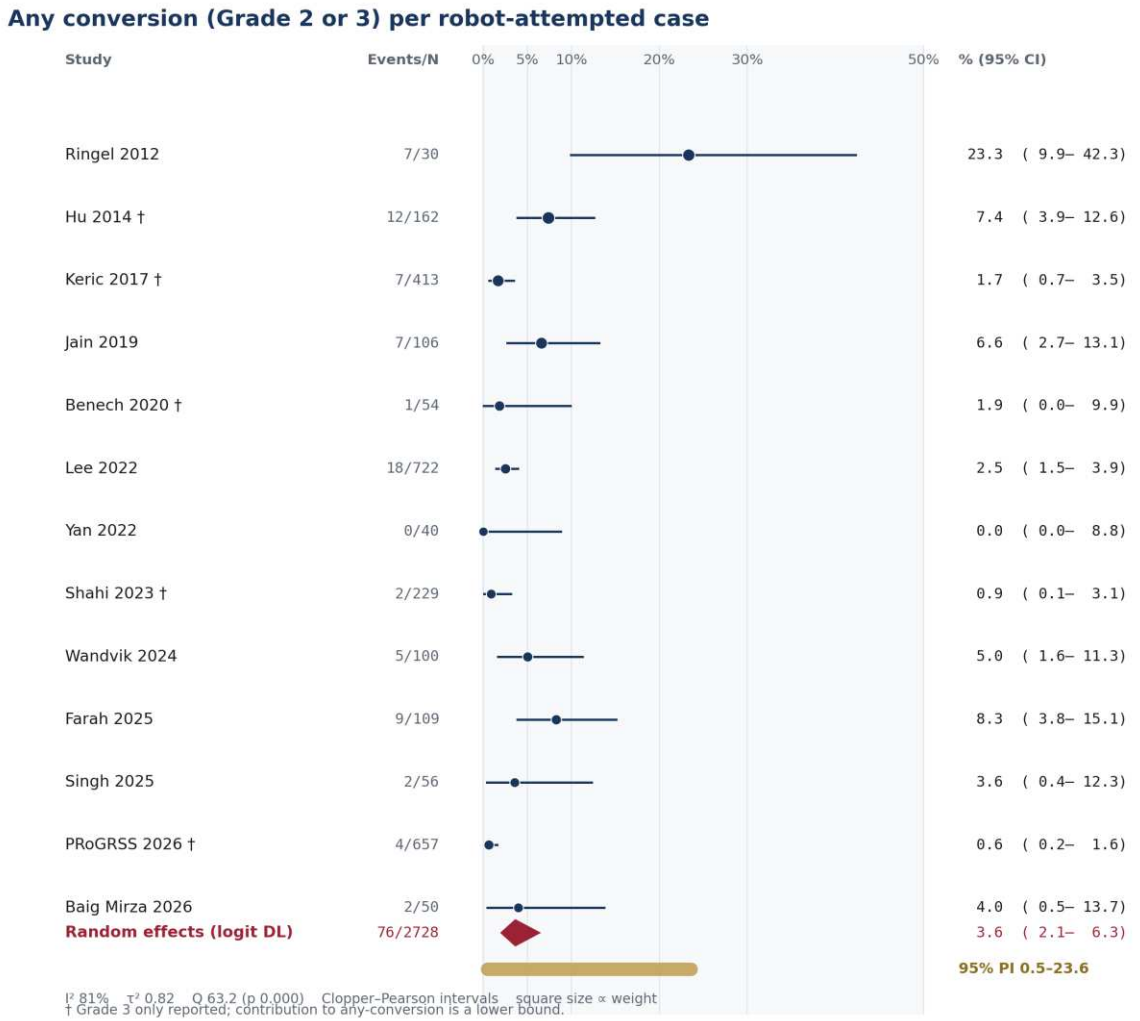


Figure 3. Grade 3 total abandonment per robot-attempted case.

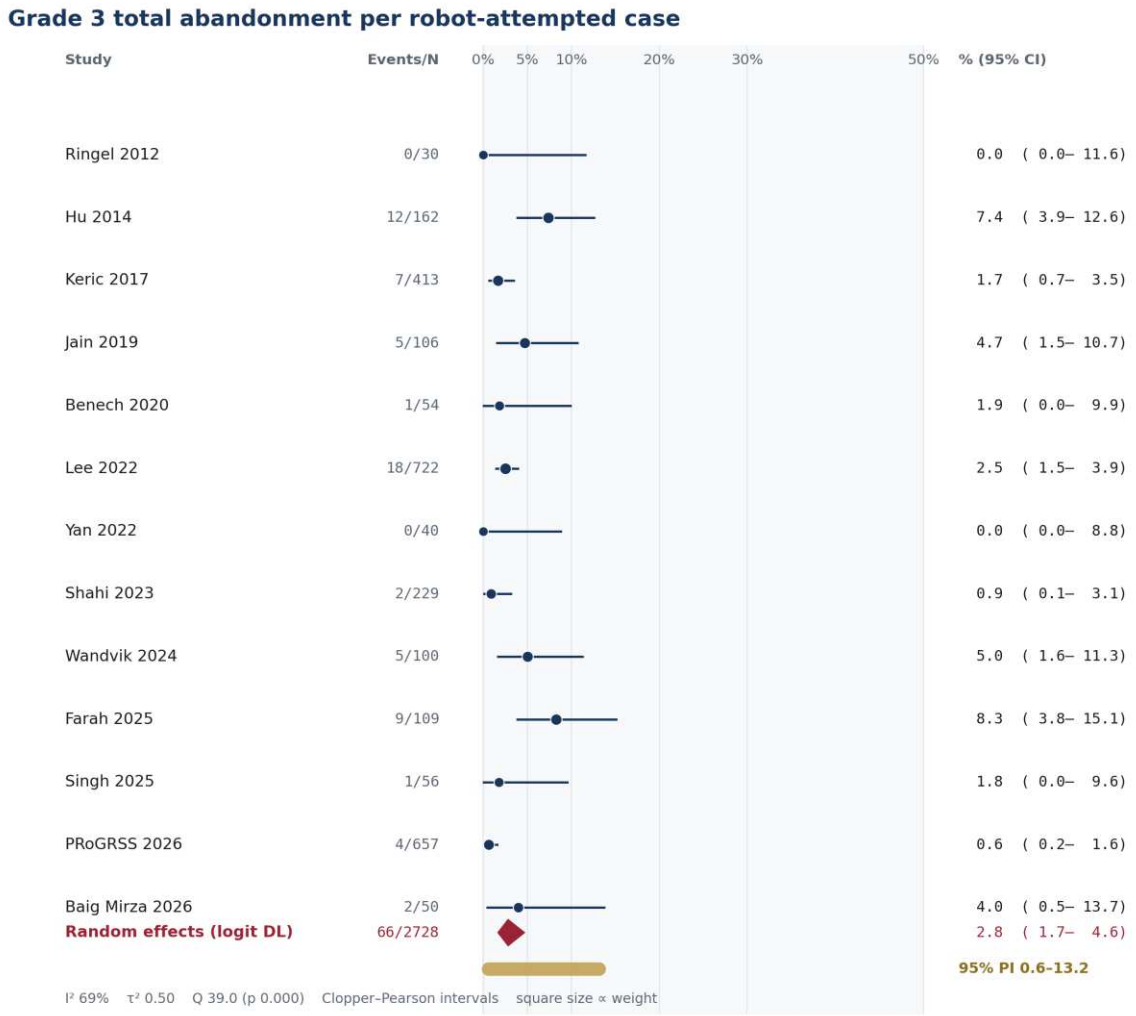


Figure 4. Per-screw conversion, display only.

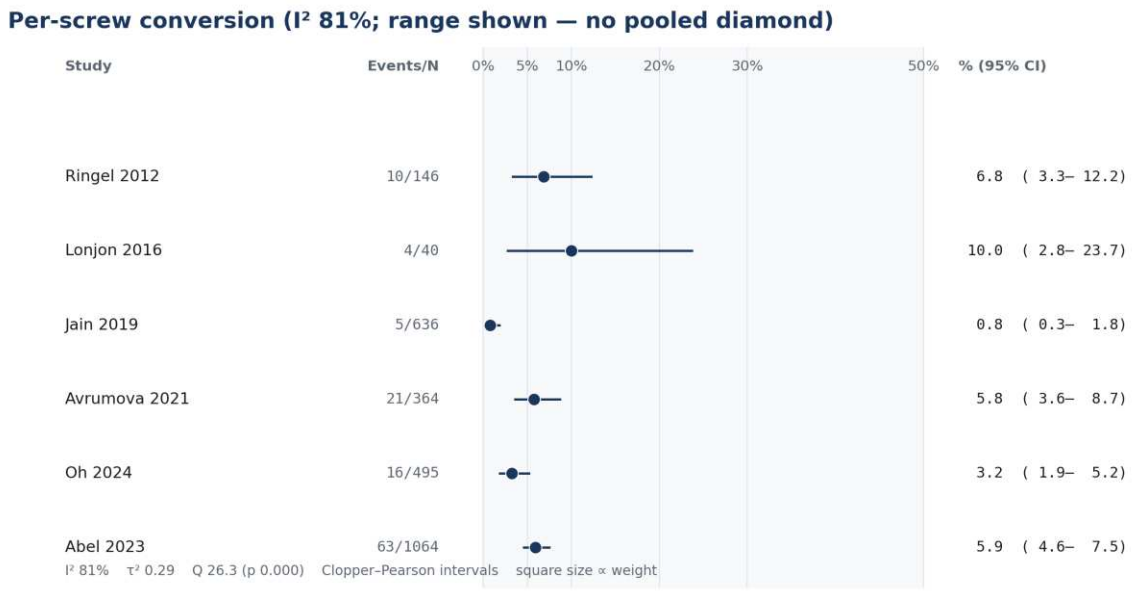


Figure 5. Subgroup forest by platform family.

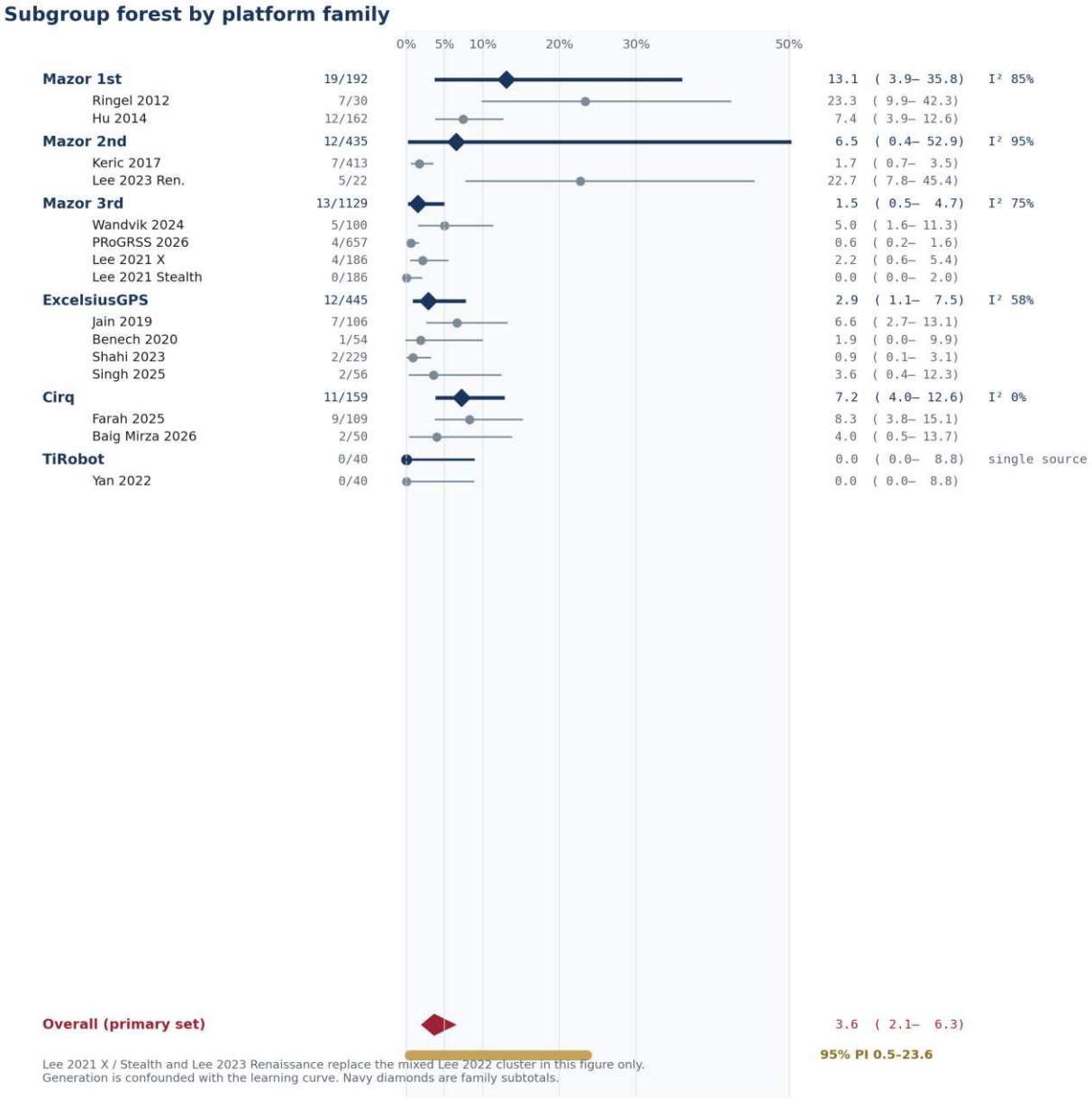


Figure 6. Funnel plot on the logit scale.

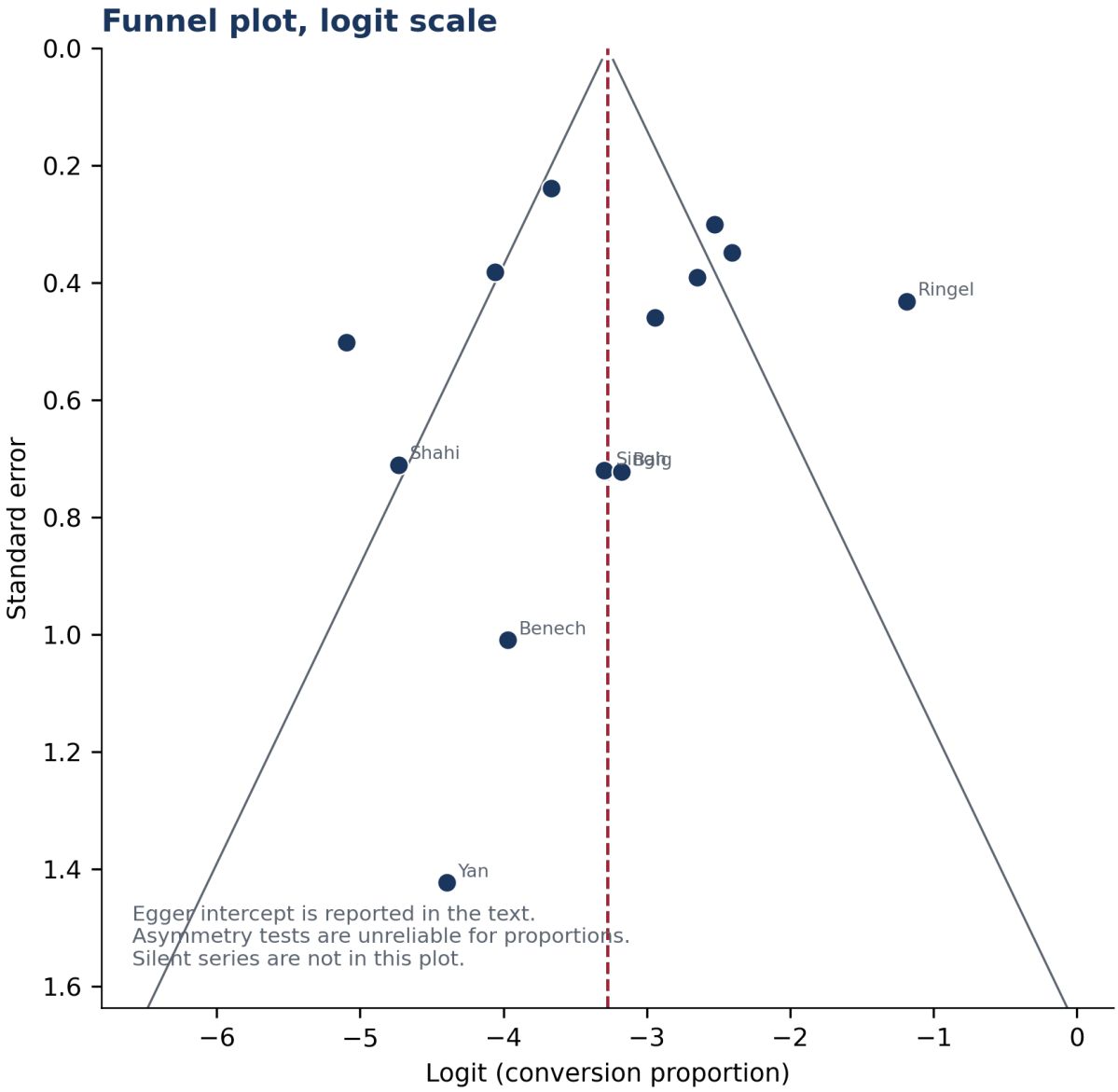


Figure 7. Failure-mode taxonomy among classified events.

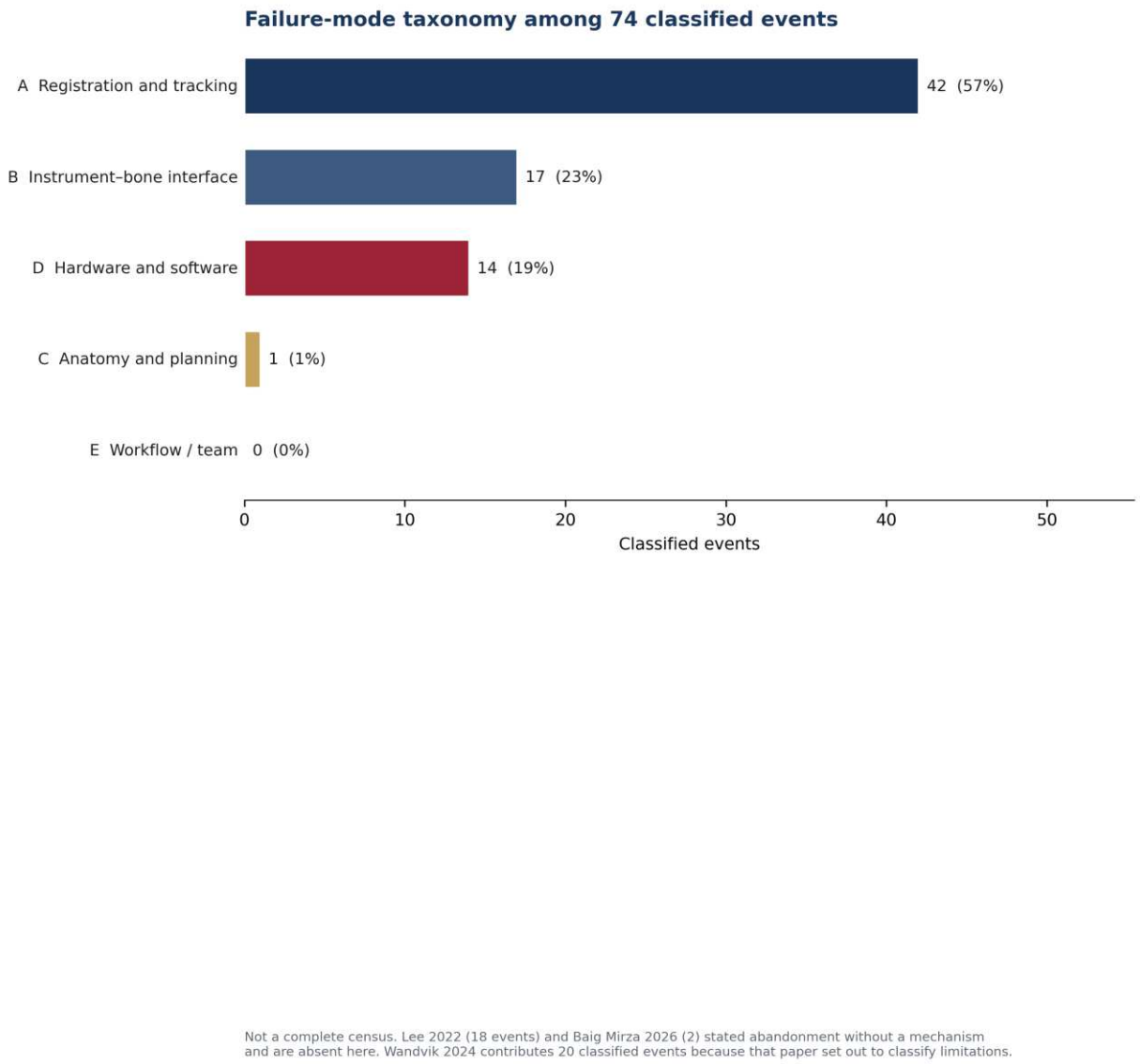


Figure 8. Leave-one-out sensitivity.

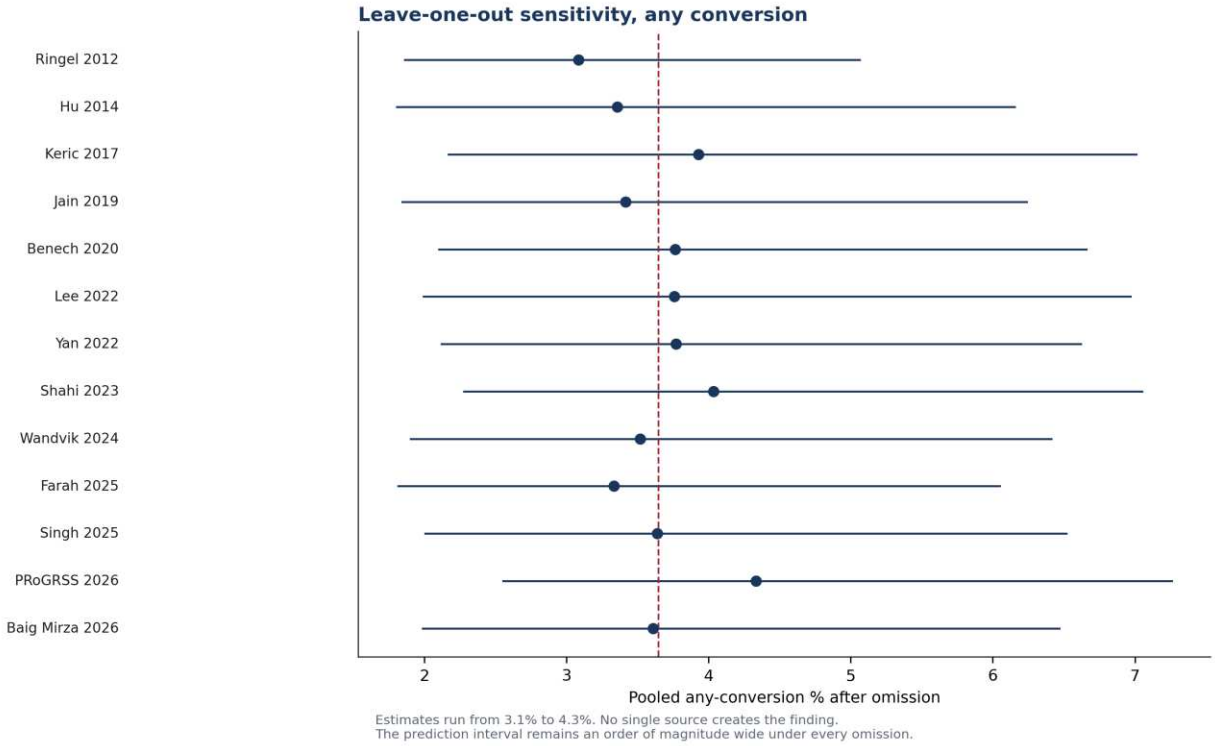


Figure 9. Learning curve (Hu 2014 and Lee 2022).

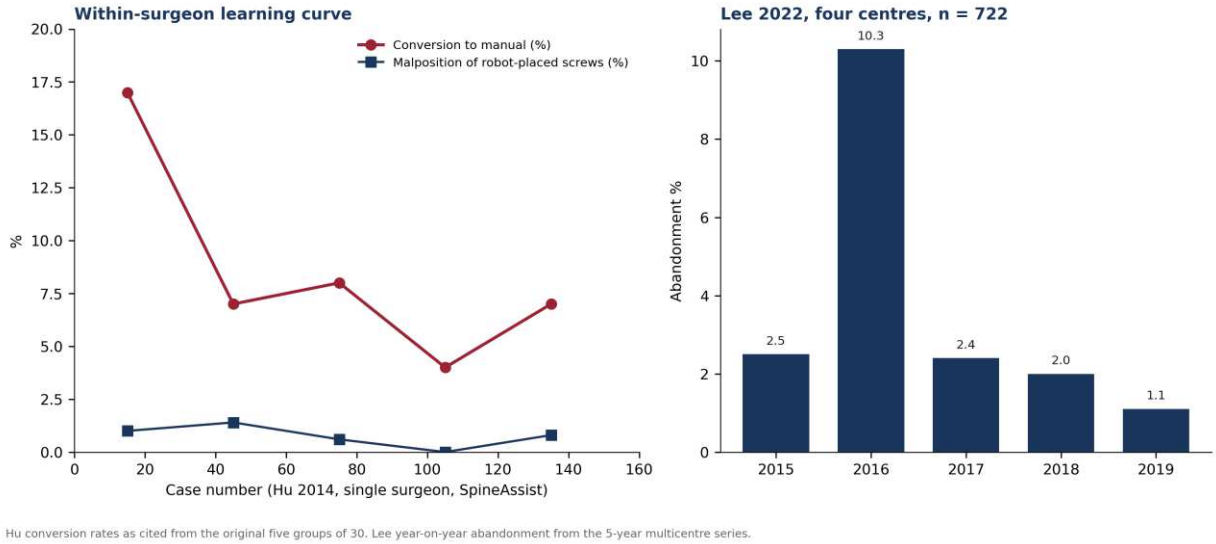
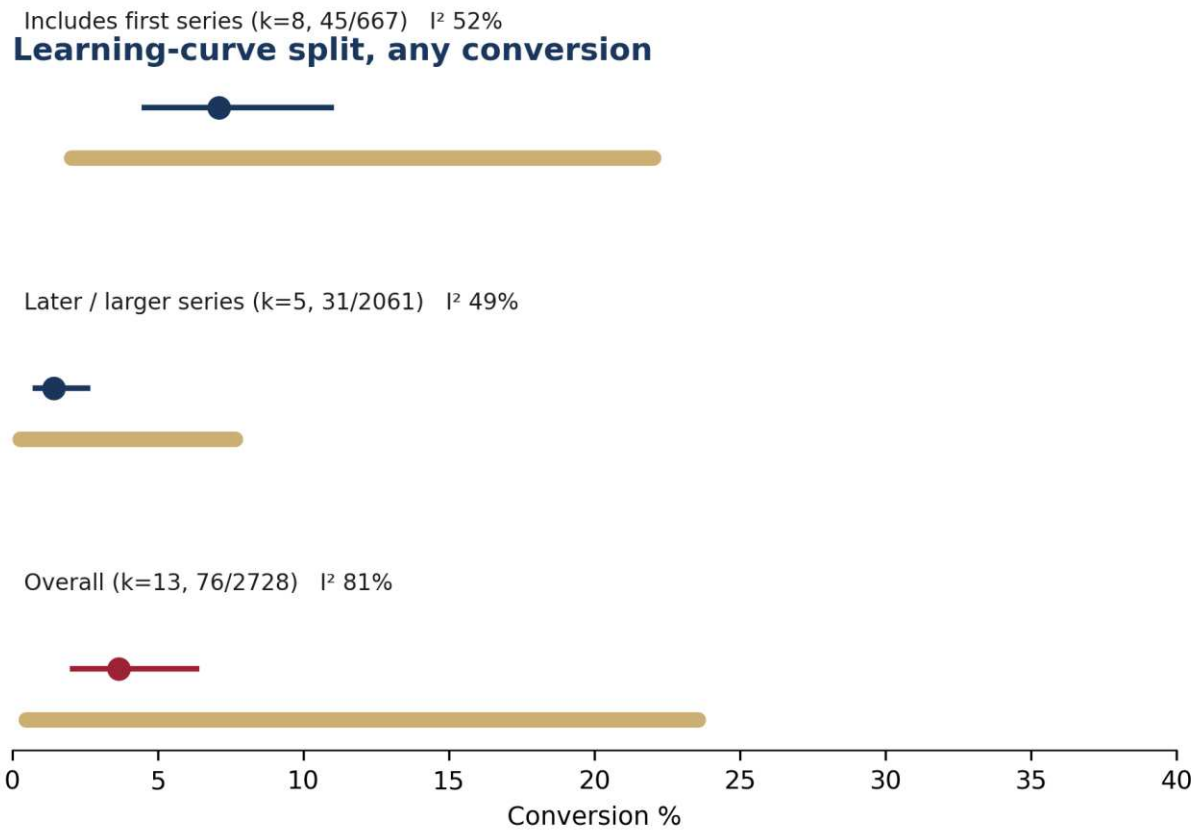


Figure 9b. Learning-curve split, pooled any conversion.



Gold bars: 95% prediction intervals. First-series sources convert higher, with more scatter.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 02Tables.docx
- 05Supplementarysearchandextraction.docx