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.

Effectiveness of Formal Mentorship in Medical Education: A Systematic Review and Meta-Analysis

Arbind Kumar Choudhary
arbindkch@gmail.com

Government Erode Medical College https://orcid.org/0000-0001-8910-1745

Ambili Remesh
Sree Uthradom Thirunal Academy of Medical Sciences (SUTAMS) https://orcid.org/0000-0002-4321-7002

Aiswarya R
Amrita Institute of Medical Sciences and Research Centre

Systematic Review

Keywords: Mentorship, mentoring, medical education, undergraduate medical education, graduate medical education, faculty development, residency match, systematic review, meta-analysis, risk of bias

Posted Date: September 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10965398/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Background. Formal mentorship — the institutionally sponsored pairing of a learner with a more experienced mentor according to a defined structure — is widely adopted in medical education, and advocacy has outpaced outcome evidence. No previous synthesis has restricted attention to formally structured programmes reporting effectiveness outcomes, or attempted quantitative pooling.

Methods. Reported in accordance with PRISMA 2020. A protocol was written before screening, but the review was not prospectively registered. Six databases and Google Scholar were searched from inception to 31 March 2026, with hand-searching, citation tracking and a supplementary reference-list search conducted during search-strategy peer review. Eligible studies evaluated a formal, structured mentorship programme — one in which pairing is assigned or systematically brokered — for medical students, residents or academic faculty, and reported an effectiveness outcome. Two reviewers worked independently. RoB 2 was applied to the randomised study, ROBINS-I to the sixteen non-randomised studies with a comparator, and AXIS with the Mixed Methods Appraisal Tool to the three single-arm studies. Random-effects meta-analysis (DerSimonian–Laird, logit) was performed where three or more comparable estimates were available, with Hartung–Knapp–Sidik–Jonkman adjustment, restricted maximum likelihood, a binomial–normal generalised linear mixed model and prediction intervals as robustness analyses.

Results. Twenty studies were included (14 United States, 2 Canada, and one each from Nigeria, India, Pakistan and Saudi Arabia; approximately 3,900 participants), with a further five eligible on title and abstract but awaiting classification because the full text could not be obtained. One was randomised. The exploratory pooled match proportion among applicants (five cohorts, four studies) was 80.0% (95% CI 66.2–89.1; I^2 = 22.0%; τ^2 = 0.148; prediction interval 42.6–95.6), widening to 59.2–91.7 under Hartung–Knapp–Sidik–Jonkman adjustment and rising to 84.8% (70.8–92.8) under the mixed model; it pools four specialties and lies below the contemporaneous United States allopathic senior match rate of approximately 93.9%. Three studies reported null effects: on residency match, on examination performance, and on vitality and engagement. Formal faculty mentoring was associated with promotion 1.0 year sooner (95% CI 0.23–1.77) and with senior rank at long-term follow-up (47% versus 13%). Mentor prevalence rose from 50% to 96% in the one study that measured it. Two studies measured validated wellbeing instruments. Fifteen of the sixteen non-randomised comparative studies were at serious or critical risk of bias, and the randomised study was at high risk.

Conclusions. Formal mentorship is predominantly but not uniformly associated with favourable outcomes; the addition of null results materially qualifies the previously reported consistency of direction. The evidence remains dominated by small, single-institution, uncontrolled evaluations at serious risk of bias, and no pooled estimate is robust to plausible analytical choices. Structured mentorship is justified on grounds of equity of access, acceptability and low cost; cluster-randomised or stepped-wedge designs, a core outcome set and cost-effectiveness analysis are required before magnitude of benefit can be estimated.

Registration. This review was not registered.

Introduction

Mentorship occupies an unusual position in medical education. Few interventions command such broad normative endorsement, and few rest on such a limited quantitative foundation. The systematic review by Sambunjak and colleagues, published in 2006, examined 39 studies and concluded that although mentorship was widely believed to influence personal development, career guidance, career choice and research productivity, fewer than half of medical students and in some fields fewer than a third of faculty reported having a mentor at all, and the evidence describing what mentorship achieved was limited [1]. The subsequent qualitative synthesis by the same group characterised mentoring as a complex, reciprocal, developmentally oriented relationship that could not be reduced to a discrete intervention [2]. Straus and colleagues, in an interview study across two academic health centres, identified reciprocity, mutual respect, clear expectations, personal connection and shared values as the constituents of successful relationships, and poor communication, lack of commitment, personality mismatch, perceived competition and conflicts of interest as the constituents of failed ones [3]. Jackson and colleagues framed the same phenomenon in the vocabulary of interpersonal chemistry, noting that the effective mentoring dyad is characterised by an affective bond that institutions can facilitate but cannot manufacture [4].

This observation frames the central question of the present review. If mentoring operates through a relationship whose quality depends on interpersonal fit, can a *formal* programme — one in which the institution assigns or systematically brokers the pairing, specifies the cadence of meetings and holds the dyad accountable — reliably produce the benefits attributed to organically formed mentorship? Buddeberg-Fischer and Herta, reviewing formal mentoring programmes for medical students and doctors, found programmes to be heterogeneous in aim, structure and evaluation, with outcome data largely confined to satisfaction [5]. Frei, Stamm and Buddeberg-Fischer reached a similar conclusion for the period 2000–2008 and noted that formal programmes were absent in most countries [6]. Kashiwagi, Varkey and Cook, synthesising mentoring programmes for physicians in academic medicine, found that programmes commonly reported improved perceptions of mentoring but that objective outcomes were reported inconsistently and rarely with adequate comparison [7].

The most methodologically detailed recent contribution is the systematic review of mentorship of US medical students by Farkas and colleagues, which screened 3,743 unique citations and included 30 articles [8]. Its central empirical finding is instructive: of 30 programmes, only seven reported objective outcomes — clerkship grades, research productivity or residency match — while nineteen reported only subjective outcomes, predominantly satisfaction ratings. Atlas and colleagues, examining 36 studies published between 2014 and 2019 from a global perspective, found the same asymmetry internationally and reported that students consistently prefer highly invested mentors with strong interpersonal skills, irrespective of national context [9]. Nimmons, Giny and Rosenthal similarly concluded that the evidence base is descriptive rather than evaluative [10].

The rationale for institutional investment nonetheless remains strong on several independent grounds. First, the unmet need is large and persistent. Kalet, Krackov and Rey argued two decades ago that the traditional apprenticeship model had been eroded by clinical productivity pressures and required

deliberate institutional replacement [11]. Aagaard and Hauer found that only a minority of medical students formed a mentoring relationship spontaneously, and that those who did were disproportionately students already advantaged in social capital and research exposure [12]. Focus-group work by Hauer and colleagues showed that students who lacked mentors often did not know how to seek one [13]. Dimitriadis and colleagues, characterising mentoring relationships formed by German medical students and faculty, and Meinel and colleagues, surveying mentoring programmes across German medical schools, both documented substantial unmet demand [14, 15]; von der Borch and colleagues subsequently designed a large-scale programme on the basis of a quantitative and qualitative needs analysis [16].

Second, the outcomes plausibly affected by mentorship are of high institutional consequence. Ramanan and colleagues reported in two separate studies that mentored medical students and residents were more satisfied with their training and more likely to report adequate career preparation [17, 18].

Third, learner distress is prevalent and consequential. Dyrbye, Thomas and Shanafelt documented high rates of psychological distress among North American medical students [19]; Rotenstein and colleagues, in a meta-analysis of 195 studies, estimated the prevalence of depression or depressive symptoms among medical students at 27.2% and of suicidal ideation at 11.1% [20]; Dyrbye and colleagues found burnout to be more common among medical students, residents and early-career physicians than in the age-matched general population [21]; and Hu and colleagues documented widespread discrimination, abuse and harassment in surgical residency training, with strong associations with burnout [22]. Mentorship is one of the few structural interventions plausibly capable of interrupting these trajectories at scale.

Fourth, mentorship has been advanced as a lever for equity. The Association of American Medical Colleges has documented persistent underrepresentation of Black, Hispanic and American Indian/Alaska Native physicians relative to the US population [23]. Rodríguez, Campbell and Pololi described the disproportionate uncompensated service burden borne by faculty underrepresented in medicine, which itself impedes advancement [24]. Beech and colleagues systematically reviewed mentoring programmes for such faculty in academic medical centres and found promising but methodologically weak evidence [25]. Yehia and colleagues found that mentorship was among the strongest correlates of pursuing an academic career among residents from diverse backgrounds [26]. In the faculty literature, Palepu and colleagues found that junior faculty with mentors reported greater career satisfaction and research productivity [27]; Wasserstein, Quistberg and Shea reached similar conclusions in a single-institution faculty survey [28]; and DeCastro and colleagues argued for moving beyond the dyadic conception towards mentor networks [29].

Fifth, the architecture of mentorship has diversified in ways that alter both cost and plausible mechanism. Skjevik and colleagues reviewed group mentorship for undergraduate medical students [30]; Akinla, Hagan and Atiomo reviewed near-peer mentoring for first-year medical students [31]; Pethrick and colleagues reviewed peer mentoring in residency education [32]; and Burgess, van Diggele and Mellis reviewed mentorship across the health professions [33]. Sng and colleagues, and subsequently Tan and colleagues, produced thematic analyses that established a vocabulary — matching, training, duration, monitoring and host organisation support — for describing programme architecture systematically [34, 35].

Against this background, three gaps motivate the present review. The first is definitional. Previous syntheses have generally included any programme in which mentorship featured, conflating formal assigned dyads with advisory colleges, career-advising offices, online mentor databases and informal relationships. If the question of interest is whether institutions can produce benefit by formalising mentorship, the review must be restricted to programmes in which pairing is assigned or systematically brokered. The second is evaluative. Most previous reviews have stopped at narrative description and have not appraised risk of bias with a validated instrument. The third is quantitative. No previous review has attempted meta-analysis in this field; where pooling is not supportable, the specific reasons should be documented rather than assumed.

A fourth consideration concerns the level at which outcomes are measured. Kirkpatrick's four-level hierarchy — reaction, learning, behaviour, results — remains the dominant organising framework in health professions education evaluation [36], and the mentorship literature sits predominantly at the level of reaction. Cook and West have cautioned against the uncritical elevation of outcomes research in medical education, noting that distal outcomes are frequently confounded and that mechanistic clarification may be more valuable than effect estimation [37]; Norman made the related argument that randomised trials in education often produce results that are simultaneously confounded and trivial [38]. These positions bear directly on whether the deficiencies documented here represent remediable failures or intrinsic limits of the field, and are considered in the Discussion.

Finally, generalisability beyond high-income settings has been almost entirely unexamined. Ssemata and colleagues described a student and faculty mentorship programme at Makerere University, and Nakanjako and colleagues surveyed mentorship needs at the same institution, both documenting acute mentor shortages in resource-limited contexts [39, 40].

We therefore addressed the following question: among medical students, residents and faculty in academic medicine, does participation in a formal, structured mentorship programme, compared with no such programme or with informal mentorship, improve career and residency match outcomes, research productivity, academic performance, satisfaction and mentoring-relationship outcomes, or academic advancement?

Methods

Protocol, registration and reporting standards

This systematic review and meta-analysis was designed, conducted and reported in accordance with the PRISMA 2020 statement [41] and its explanation and elaboration document [42]. The search is reported in accordance with PRISMA-S [43]. Because the review concerns health professions education, the STORIES statement was additionally consulted for domain-specific reporting expectations [44]. A protocol specifying the research question, eligibility criteria, search strategy, outcome domains, risk-of-bias instruments and analysis plan was written before screening commenced and is available from the corresponding author on request. The review was not registered in PROSPERO or any other prospective register; although prospective registration is

recommended to reduce duplication and selective reporting [45], it was not undertaken, and the absence of a publicly archived protocol should be weighed when appraising the analyses reported here, several of which are explicitly identified as post hoc. Completed PRISMA 2020 and PRISMA-S checklists are provided as Supplementary Table S9. Deviations from the protocol are reported in the Results and Discussion.

Eligibility criteria

Eligibility was defined using a PICOS framework.

Population. Medical students enrolled in an undergraduate medical degree programme (MD, MBBS or equivalent); resident physicians or postgraduate trainees; or faculty members in academic medicine. Studies of nursing, pharmacy, dentistry, allied health or non-clinical graduate students were excluded unless medical trainees were reported separately. Academic faculty were included because formal mentorship is deployed along a continuum from matriculation to promotion, the theoretical mechanism is shared across levels, and a review restricted to trainees would exclude the studies reporting objective, institutionally verified outcomes with matched comparators. The faculty stratum was analysed as a distinct outcome domain and was not pooled with trainee outcomes.

Intervention. A formal mentorship programme, defined a priori as an institutionally sponsored programme in which (i) mentor–mentee pairings are assigned by the programme or systematically brokered on the basis of declared preferences or matching criteria, rather than left to spontaneous formation; (ii) the programme specifies at least one structural element — expected meeting frequency, minimum duration, defined mentor role, or mentor training; and (iii) the programme is described as mentorship rather than solely as academic advising, tutoring, remediation coaching or clinical supervision. Peer and near-peer mentoring programmes were eligible where the pairing was programme-brokered, and group mentoring formats were eligible where the groups were constituted by the programme. Where mentorship formed one component of a multi-component intervention, such as a pipeline or research-brokerage programme, the study was eligible only if the mentorship component was separately described and met criteria (i) to (iii); because such programmes confound mentorship with co-interventions, a post-hoc sensitivity analysis restricted to programmes whose principal component was mentorship was added.

Programmes consisting only of an online mentor database or directory, with no brokering of pairs, were excluded, as were advisory-college structures whose primary function was administrative or assessment-related. Interventions whose randomised or assigned element was the training of mentors, or the addition of a component to relationships that had already formed spontaneously, were excluded because the pairing itself was not programme-determined.

Comparator. Any comparator, including randomised concurrent controls, non-participants from the same institution, historical cohorts preceding programme implementation, national benchmark rates, or an internal pre–post comparison. Studies without any comparator were eligible for narrative synthesis and single-arm proportional meta-analysis but not for comparative pooling.

Outcomes. At least one reported effectiveness outcome falling within one of five pre-specified domains: (1) career and residency match outcomes, including specialty match rate, specialty choice, application volume and pipeline representation; (2) research productivity, including publications, presentations, funded projects and awards; (3) academic performance, including clerkship honours, examination performance and academic standing; (4) satisfaction, mentoring-relationship and wellbeing outcomes, including satisfaction ratings, perceived benefit, relationship continuation, mentor prevalence, perceived support and validated wellbeing instruments; and (5) faculty advancement, including time to promotion, rank attained and retention. Studies reporting only programme description, uptake figures or mentor-side outcomes were excluded.

Study design. Any design: randomised or non-randomised trials, prospective or retrospective cohort studies, controlled or uncontrolled before–after studies, cross-sectional surveys of programme participants, and mixed-methods programme evaluations. Systematic, narrative and scoping reviews, editorials, commentaries, conference abstracts without full text, and letters were excluded as primary sources but were retained for reference-list hand-searching and contextual citation.

Language and date. English language; database inception to 31 March 2026. No geographic restriction was applied.

Information sources and search strategy

We searched PubMed/MEDLINE, Embase (Elsevier), ERIC (EBSCO), the Cochrane Central Register of Controlled Trials and the Cochrane Database of Systematic Reviews, Scopus, and the first 300 records of Google Scholar sorted by relevance. Searches combined controlled vocabulary and free-text terms across three concept blocks — mentorship, medical education, and evaluation or effectiveness — joined with the Boolean AND operator. The full MEDLINE strategy is reproduced in Supplementary Table S1; representative elements were:

("Mentors"[Mesh] OR "Mentoring"[Mesh] OR mentor*[tiab] OR mentee*[tiab] OR preceptor*[tiab]) AND ("Education, Medical"[Mesh] OR "Students, Medical"[Mesh] OR "Internship and Residency"[Mesh] OR "Faculty, Medical"[Mesh] OR "medical student*"[tiab] OR resident*[tiab] OR "medical school*"[tiab] OR "academic medicine"[tiab]) AND (formal*[tiab] OR structured[tiab] OR assign*[tiab] OR program*[tiab] OR "Program Evaluation"[Mesh] OR outcome*[tiab] OR effectiveness[tiab] OR impact[tiab] OR evaluat*[tiab])

The strategy was drafted by the review team, peer-reviewed against the PRESS 2015 checklist by an information specialist not otherwise involved in the review [46], and revised accordingly. Search strings were translated for each platform with attention to database-specific syntax and thesauri. In addition, we hand-searched the reference lists of all previous systematic, scoping and narrative reviews identified; hand-searched the tables of contents of *Medical Education*, *Academic Medicine*, *Medical Teacher*, *BMC Medical Education* and *Journal of Graduate Medical Education* for the preceding five years; performed

forward citation tracking in Scopus on all included studies and on the reviews by Farkas and by Sambunjak and colleagues; and examined MedEdPORTAL and the AAMC website for programme reports. The final search was executed on 31 March 2026 and re-run immediately before analysis, with no further eligible records identified.

A second, supplementary search was undertaken during peer review of the search strategy. Reviewing the yield of the primary search — 24 included studies from 4,710 unique records — indicated that the third concept block, which required a term such as *formal*, *structured*, *assign*, *program*, *outcome*, *effectiveness*, *impact* or *evaluat* in the title or abstract, would fail to retrieve reports describing a mentoring *scheme*, *dyad* or *matched pair* and reporting *attitudes* or *perceptions*. We therefore performed a targeted citation and reference-list search of all previous reviews in this field and of the reference list of the present manuscript, retrieving 16 additional reports for full-text assessment. This supplementary stream is reported separately in Figure 1, and its consequences for the completeness of the primary search are considered as a limitation.

Study selection

Records were deduplicated in EndNote and uploaded to Rayyan for screening [47]. Two reviewers independently screened titles and abstracts against the eligibility criteria, blinded to each other's decisions. Full texts of all records passing the first stage were retrieved and independently assessed by the same two reviewers, with reasons recorded for every exclusion at the full-text stage. Disagreements were resolved by discussion, with a third reviewer adjudicating where consensus was not reached. Inter-rater agreement was quantified using Cohen's kappa [48] and interpreted using the benchmarks of Landis and Koch [49].

Data extraction and data verification

A structured extraction form was developed, piloted on three studies and revised before full deployment. Extracted fields comprised bibliographic details; country and setting; institution type; participant level, number and any special targeting; study design; programme architecture, including mentor type, matching method, meeting frequency, duration, mentor training, incentives and additional components; comparator; outcomes assessed; and all reported quantitative results with numerators, denominators, measures of dispersion and P values where available. Two reviewers extracted data independently into separate copies of the form and reconciled discrepancies against the source.

Every extracted datum was assigned an explicit verification level, and the level is reported alongside the datum rather than only in the supplement:

- **Level A** — confirmed against the primary full text through two independent retrieval routes, with an identity check confirming that the retrieved record matched the intended citation before any value was extracted.
- **Level B** — confirmed against the indexed abstract of the primary article and independently corroborated against a detailed extraction table in a previous high-quality systematic review [8].
- **Level C** — carried forward unchanged from a previous version of this synthesis.

Every included study carries data at one of these three levels. Records that were judged eligible on the information available but whose full text could not be obtained within the review period were not admitted to the included set at any verification level; in accordance with PRISMA 2020 they are reported separately as studies awaiting classification, with their citations and the reason for their status given in Supplementary Table S8, and they contribute to no table, figure or analysis.

The identity check at level A was introduced after an automated retrieval of one article's digital object identifier returned the full text of a different article, together with plausible but incorrect sample sizes and effect estimates. No value from that retrieval was used. We report this because the failure mode is silent: the returned content was internally coherent and would not have been detected by inspection of the extracted numbers alone.

Risk of bias assessment

Risk of bias was assessed independently by two reviewers, using the instrument appropriate to each study design.

The Cochrane risk-of-bias tool for randomised trials, RoB 2 [50], was applied to the single randomised study, generating domain-level and overall judgements of low risk, some concerns, or high risk across five domains: the randomisation process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result.

ROBINS-I [51] was applied to the sixteen non-randomised studies with a comparator (concurrent, historical, external or internal pre–post), generating domain-level and overall judgements of low, moderate, serious or critical risk of bias across seven domains: confounding, selection of participants, classification of interventions, deviations from intended interventions, missing data, measurement of outcomes, and selection of the reported result. We pre-specified that a study would be judged at serious risk under ROBINS-I if participants self-selected into the programme and no adjustment was made for factors plausibly associated with both participation and outcome, and at critical risk if, in addition, the comparator was drawn from a different time period or a different population without adjustment, or if the design permitted no separation of the programme effect from maturation or regression to the mean.

ROBINS-I was not applied to the single-arm cross-sectional surveys, because the tool is designed for studies estimating the effect of an intervention against a comparator and its domains are not interpretable in the absence of one. These studies were appraised with the AXIS critical appraisal tool for cross-

sectional studies [52] and, for mixed-methods components, the Mixed Methods Appraisal Tool version 2018 [53]. Single-arm studies are excluded from Figure 2 and from all risk-of-bias proportions, which are computed over the sixteen ROBINS-I-appraised studies only.

The Newcastle–Ottawa Scale for cohort studies [54] was additionally recorded for all studies with a comparator, scoring selection (four items), comparability (two items) and outcome (three items) for a maximum of nine stars; 7–9 stars were pre-specified as low, 4–6 as moderate and 0–3 as high risk of bias. Because the Newcastle–Ottawa Scale and ROBINS-I can yield discordant judgements, the Newcastle–Ottawa Scale was treated as a descriptive supplement rather than a co-primary instrument, and discordant ratings are reported in the Results. Four elements of the Medical Education Research Study Quality Instrument framework described by Reed and colleagues [55] were also recorded for descriptive purposes: multi-institutional design, response rate above 75%, validity evidence for the outcome instrument, and reporting of an outcome beyond satisfaction.

Data synthesis and statistical analysis

Synthesis was organised by the five pre-specified outcome domains. Where meta-analysis was not appropriate, a structured narrative synthesis was conducted following the guidance of Popay and colleagues [56] and reported in accordance with the Synthesis Without Meta-analysis (SWiM) reporting guideline [57], specifying for each domain the grouping rationale, the standardised metric used for comparison, the method of synthesis and the criteria for prioritising results.

Random-effects meta-analysis was pre-specified where three or more estimates addressing a sufficiently homogeneous outcome construct were available, using the DerSimonian–Laird estimator of between-study variance [58]. Because the DerSimonian–Laird estimator under-covers when few estimates contribute, the primary pooled estimate was additionally computed with restricted maximum likelihood estimation of τ^2 and with the Hartung–Knapp–Sidik–Jonkman variance adjustment, which is recommended in this situation [59,60]. We used the ad hoc variant of the Hartung–Knapp–Sidik–Jonkman adjustment, in which the adjusted standard error is not permitted to fall below the unadjusted value; both variants are reported, and they coincide for the primary analysis.

Three classes of analysis were planned. For comparative dichotomous outcomes, risk ratios and risk differences with 95% confidence intervals were pooled on the log scale. For single-arm proportions, two transformations were pre-specified: the Freeman–Tukey double arcsine transformation [61], back-transformed using the harmonic mean of the sample sizes as described by Miller [62] and implemented as described by Barendregt and colleagues [63], was performed as a sensitivity analysis, while the logit transformation was pre-specified as the primary metric following the demonstration by Schwarzer and colleagues that the inverse Freeman–Tukey transformation can yield misleading results when study sizes are markedly unequal [64]. A continuity correction of 0.5 was applied to cells in which no participant, or every participant, experienced the event; where a continuity correction was applied, the study-level point estimate displayed in the forest plot is the corrected proportion, so that the plotted estimate and its interval refer to the same quantity. Because correction of a boundary proportion can itself determine the pooled estimate, a random-effects binomial–normal generalised linear mixed model, which requires no continuity correction, was fitted as a pre-specified robustness analysis [65]. For continuous outcomes, mean differences were used where the measurement scale was common across studies, and standardised mean differences (Hedges' g) otherwise [66].

Heterogeneity was assessed using the Cochran Q statistic, the I^2 statistic with the thresholds proposed by Higgins and colleagues [67], and the between-study variance τ^2. Because I^2 is a relative rather than absolute measure and is strongly influenced by within-study precision, τ^2 was reported alongside it and, following Higgins, Thompson and Spiegelhalter [68] and IntHout and colleagues [69], a 95% prediction interval was calculated for every pooled estimate derived from three or more studies, on the logit scale using the t distribution with $k - 2$ degrees of freedom and back-transformed.

Small-study effects and publication bias were to be assessed by visual inspection of funnel plots and Egger's regression test [70]; in accordance with the recommendations of Sterne and colleagues these were not performed where fewer than ten studies contributed to a pooled estimate [71], which was the case for every analysis in this review.

Analyses were conducted in Python 3.11. Random-effects pooling, heterogeneity statistics, prediction intervals and the Hartung–Knapp–Sidik–Jonkman and restricted maximum likelihood variants were implemented directly using NumPy and SciPy, following the formulations described by Balduzzi, Rücker and Schwarzer [72]. The binomial–normal generalised linear mixed model was fitted by maximum likelihood using adaptive Gauss–Hermite quadrature with 80 nodes, with the standard error of the pooled logit obtained from the numerically differentiated observed information [73]. The complete analysis code is provided as Supplementary Table S10 and is deposited with the dataset.

Pre-specified sensitivity analyses comprised leave-one-out analysis for every pooled estimate; re-analysis using the alternative transformation; re-analysis substituting discrepant values identified during data verification; restriction to studies at moderate or lower risk of bias; combination into a single estimate of the two cohorts contributed by one study; and restriction to programmes whose principal component was mentorship. Pre-specified subgroup analyses by learner level, mentor type (faculty versus peer or resident), programme duration, and specialty-specific versus general programme were planned.

Certainty of evidence for each outcome domain was rated using the GRADE approach [74], with explicit consideration of inconsistency [75] and indirectness [76]. Non-randomised evidence began at low certainty, randomised evidence at high certainty; each was rated down for risk of bias, imprecision, inconsistency or indirectness as applicable, and rated up only where a large magnitude of effect was demonstrated with a credible comparator.

Results

Study selection

The study selection process is summarised in Fig. 1.

The primary search identified 7,329 records from databases and 483 from other sources. After removal of 2,981 duplicate database records and 121 records duplicated between sources, 4,710 unique records were screened at title and abstract, of which 4,420 were excluded. Inter-rater agreement at this stage was substantial (κ = 0.81, 95% CI 0.77–0.85). Of 290 reports sought for retrieval, two were not retrieved despite interlibrary loan request. The supplementary citation and reference-list search retrieved a further 16 reports. A total of 304 reports were assessed for eligibility. Two hundred and seventy-nine reports were excluded at full text, with reasons detailed in Fig. 1 and tabulated by exclusion category in Supplementary Table S2. Twenty studies met the inclusion criteria. A further five records were judged eligible on the basis of their titles and abstracts but could not be obtained in full text within the review period; in accordance with PRISMA 2020 these are reported as studies awaiting classification, are listed with their citations and the reason for their status in Supplementary Table S8, and contribute to no analysis.

Seven conference abstracts with no corresponding full paper were excluded under the category *secondary publication type* and are counted within the 279 full-text exclusions. In the previous version of this synthesis they were reported among reports not retrieved; the present accounting is the correct one, since their exclusion was a decision about publication type rather than a retrieval failure.

The yield ratio is itself informative, and its interpretation has changed. Of 4,710 unique records screened in the primary search, 288 warranted full-text assessment and 11 met the inclusion criteria — 0.23% of unique records. The supplementary search of 16 targeted reports yielded a further 9 inclusions and 5 studies awaiting classification, a yield of 56% even before the awaiting-classification records are counted. That a 16-report targeted search should identify almost as many eligible studies as a 4,710-record database search is not a comment on the value of citation searching; it is evidence that the primary search under-retrieved. The largest single exclusion category at full text remained mentorship that was not programme-brokered (n = 71), confirming that a substantial proportion of what the literature labels mentorship programmes consists of infrastructure such as databases, directories, matching websites and advisory structures rather than assigned or brokered relationships. This is consistent with the observation of Farkas and colleagues that only seven of thirty included programmes reported objective outcomes [8].

Agreement at the full-text stage was recomputed for this revision from the screening dataset rather than carried forward. Fourteen disagreements among the 304 reports assessed corresponds to raw agreement of 95.4%; with the marginal distribution implied by 20 consensus inclusions this gives Cohen’s kappa 0.72 (95% CI 0.57–0.86), substantial agreement on the benchmarks of Landis and Koch. The value is insensitive to how the fourteen disagreements are apportioned between reviewers, varying by less than 0.01 across the plausible range. Eleven of the fourteen disagreements were resolved by discussion without recourse to the third reviewer.

Characteristics of included studies

The characteristics of the twenty included studies are presented in Table 1, and the studies added in the present revision, together with the reasons for the principal documented exclusions, are set out in Table 6.

Table 1

Characteristics of the 20 included studies. Studies marked were added in the present revision. Verification level: A, primary full text confirmed through two independent retrieval routes with an identity check; B, indexed abstract with corroboration from a previous high-quality review; C, carried forward unchanged from the previous version of this synthesis.

#	Study	Country and setting	Population and sample	Learner level	Study design	Comparator	Duration	Mentor type	Ver.
1	Indyk 2011	USA; Albert Einstein College of Medicine	Year 1 cohort n = 29 (16 completers); year 2 cohort n = 33 (25 completers, 19 with match data)	Medical students, entry to match	Prospective programme evaluation	Concurrent classmates and within-programme non-completers	4 years	Faculty (primary care)	C
2	Day 2016	USA; single academic surgical department	Junior medical students versus non-participants; denominators not reported	Medical students, clerkship	Survey and retrospective grade review	Concurrent non-participants	Clerkship phase	Surgical residents	C
3	Zuzuarregui 2015	USA; Boston Medical Center, neurology	CORTEX participants; n not separately reported	Medical students	Retrospective pre–post	Historical institutional averages	Multi-year	Neurology faculty	C
4	Hirsch 2015	USA; single institution, radiation oncology	58 participants over 10 years; 17 applied	Medical students	Retrospective 10-year review	External national benchmark	10-year programme	Radiation oncology faculty	C
5	Nellis 2016	USA; Johns Hopkins, otolaryngology	15 participants underrepresented in medicine; 7 applied	Medical students, targeted	Survey and retrospective review	Historical institutional representation	Weeks to months	Faculty and residents	C
6	Sobbing 2015	USA; single obstetrics and gynaecology department	Clerkship students with assigned resident mentors	Medical students, clerkship	Survey and retrospective review	Historical application counts	Clerkship block	Residents	C
7	Johnson 1998	USA; University of Pennsylvania	Students and junior faculty underrepresented in medicine	Students and faculty	Retrospective pre–post	Historical pre-programme	Multi-year	Faculty	C
8	Cohee 2015	USA; Walter Reed, internal medicine	26 survey respondents	Residents	Cross-sectional with retrospective pre-programme comparison	Internal pre-programme	Up to 3 years	Faculty	C
9	Morrison 2014	Canada; University of Toronto	611 full-time faculty (382 pre-2003, 229 from 2003)	Faculty	Comparative administrative data 1988–2010	Concurrent covariate-matched and historical	Continuous from appointment	Senior faculty	C
10	Altonji 2019	USA; University of Alabama	First-year students; response rate 49.5%	Medical students	Cross-sectional survey	None (single-arm)	One academic year	Peer (second-year students)	C
11	Wasim 2025	Pakistan; Services Institute of Medical Sciences, Lahore	362 MBBS students providing feedback	Medical students	Cross-sectional survey	None (single-arm)	11 months	Trained faculty	C
12	Pololi 2023	USA; 27 medical schools across 16 states	40 early-midcareer research faculty holding first R01 awards	Faculty	Stratified randomised controlled trial	Waitlist control	1 year	Facilitated peer group	A
13	Efstathiou 2018	USA; Massachusetts General Hospital	23 mentees and 91 non-enrolled controls	Faculty	Prospective longitudinal cohort 2009–2016	Concurrent non-enrolled controls	9 months plus follow-up to 7 years	Senior faculty	A
14	Witten 2021	USA; John A. Burns School of Medicine	288 students; 105 (36%) applied; 20 mentees, 17 surveyed	Medical students, first year	Prospective programme evaluation	Concurrent non-participating classmates	6 months	Primary care faculty	A
15	Zhang 2017	Canada; University of Alberta, otolaryngology	8 residents, PGY-1 to PGY-4	Residents	Prospective interventional	Internal pre–post at 12 months	12 months	Faculty (tiered mosaic model)	A

#	Study	Country and setting	Population and sample	Learner level	Study design	Comparator	Duration	Mentor type	Ver.
16	Eleje 2025	Nigeria; Nnamdi Azikiwe University	7 mentees, 12 peer mentors, 30 controls	Medical students, preclinical	Quasi-experimental with concurrent control	Concurrent control group	6 months	Peer (high-performing classmates)	A
17	Boedeker 2024	USA; military Health Professions Scholarship Program	31 mentors and 25 mentees; 21 applied	Medical students	Quasi-experimental pre–post	Internal pre–post	~ 4 months	Military obstetrics and gynaecology faculty	A
18	Fallatah 2018	Saudi Arabia; King Abdulaziz University	185 fourth-year students; 184 completed the rotation	Medical students, clinical years	Cross-sectional with internal comparison	Attenders versus non-attenders	One clinical rotation	Faculty, one per ten students	A
19	Guhan 2020	India; Al-Azhar Medical College, Kerala	148 first-year MBBS students	Medical students, first year	Single-group pre–post	Internal pre–post	6 months	Faculty (six mentors)	A
20	Andre 2017	USA; UT Health San Antonio	670 all-student survey responses across four annual waves; 132 mentor responses	Medical students	Longitudinal programme evaluation, 2011–2014	None (single-arm; serial annual surveys)	Longitudinal, by learning community	Peer (rising third-year students)	A

CORTEX = Comprehensive Opportunities for Research and Teaching Experience; PGY = postgraduate year.

Fourteen of the twenty studies were conducted in the United States, two in Canada, and one each in Nigeria, India, Pakistan and Saudi Arabia. Nineteen of twenty were single-institution evaluations; the sole exception is the randomised trial, which enrolled faculty from 27 medical schools across 16 states under a common protocol [77]. One study was randomised. Seventeen studies incorporated a comparator of some kind: one randomised concurrent control [77]; five concurrent non-randomised comparators [78–82]; one concurrent control group defined by baseline attainment [83]; four historical or pre-implementation comparisons [84–87]; one external national benchmark [88]; one historical institutional benchmark for a secondary outcome [89]; and four internal pre–post or within-programme comparisons [90–93]. Three studies were single-arm [94–96].

Programme architecture varied along every axis identified in the thematic frameworks of Sng and of Tan and colleagues [34, 35]. Mentors were faculty in eleven programmes, residents in two, mixed in one, and peers in four; one programme drew on faculty, residents and senior students in a tiered structure. Matching was preference-based in the most structured programmes [78, 81], algorithmic on personality and occupational compatibility for junior residents in one [90], explicitly random in two [91, 92], by administrative identifier in one [93], and unspecified elsewhere. Mentor training was reported in six of the twenty programmes [78, 95, 81, 83, 92, 93], and only one reported financial compensation to mentors, a stipend of US$30 per documented meeting [78]. Duration ranged from a single clerkship block to four years to indefinite. This variation is not incidental: it constitutes the intervention, and any pooled estimate is therefore an average across programmes that differ in dose, mechanism and plausibly in effect.

Approximately 3,900 participants were enrolled across the sixteen studies with extractable denominators. Excluding the three largest surveys [80, 95, 96], the remaining thirteen studies together contributed fewer than 1,300 participants, and the median number of participants contributing to any single outcome comparison was fewer than 30.

Detailed outcomes of included studies

Detailed extracted outcomes for the studies added in this revision are presented in Table 2; the corresponding data for the eleven studies carried forward are unchanged and are reproduced in Supplementary Table S3.

Table 2
Programme model, outcomes assessed and verified quantitative results of the 9 studies added in the present revision.

#	Study	Formal mentorship model	Outcomes assessed	Verified quantitative results	Ver.
12	Pololi 2023	C-Change Mentoring and Leadership Institute: year-long facilitated peer-group mentoring; groups constituted by the programme; quarterly structured two-to-three-day sessions; stratified randomisation with waitlist control	Vitality; research and career self-efficacy; valuing diversity; empathy; self-efficacy in mentoring others	Perceived vitality *D* = 1.47, P = 0.03; research self-efficacy *D* = 0.57, P = 0.02; career-advancement self-efficacy *D* = 0.41, P = 0.05; valuing diversity *D* = 0.46, P = 0.02; cognitive empathy *D* = 0.85, P = 0.03; skills against sexism and racism *D* = 0.71, P = 0.01; self-efficacy in mentoring others *D* = 1.14, P = 0.007; objective vitality *D* = 0.31, P = 0.20 (not significant)	A
13	Efstathiou 2018	Structured faculty mentoring: mentees ranked five development priorities, mentors listed expertise areas, senior faculty facilitated the final pairing; three two-hour training sessions over nine months plus regular meetings	Attainment of senior faculty rank; retention; external funding	Senior faculty rank at follow-up 47% versus 13%, P = 0.030; retention 15/20 (75%) versus 23/32 (72%); median external funding US$1.78M versus US$0.50M	A
14	Witten 2021	Six-month Primary Care Mentorship Program for first-year students; applicants assigned to primary care faculty mentors	Match into a primary care residency; relationship continuation	Participation "does not correlate with an increased likelihood of matching into a primary care GME residency specialty" — **null**. 7/17 (41%) continued the mentoring relationship after the programme ended; 105/288 (36%) of the eligible classes applied to participate	A
15	Zhang 2017	Tiered mosaic model: one primary mentor meeting monthly with supplemental mentors available; PGY-1 residents algorithm-matched on personality and occupational compatibility, PGY-2 to PGY-4 selecting from programme-supplied lists	Perceived stress; burnout; quality of life	Perceived Stress Scale 18.5 → 7.9, P = 0.001. Maslach Burnout Inventory: emotional exhaustion 47.6 → 14.9; depersonalisation 50.6 → 20.1; personal achievement 16.5 → 42.5; all P < 0.0001	A
16	Eleje 2025	Intra-class peer mentorship: mentees (continuous assessment below 45%) paired by the programme with high-performing classmates (≥ 70%); six months, weekly meetings, six hours of structured mentor training; controls scored 50–70%	Continuous assessment score; subjective vitality; student engagement	Continuous Assessment Test, median: mentees 40.0% → 70.0% versus controls 60.0% → 57.0%, P = 0.003. Subjective vitality P = 0.722 (**null**); student engagement P = 0.388 (**null**); classroom participation P = 0.001	A
17	Boedeker 2024	Military residency match preparation: two mentors assigned per mentee, paired within service branch where possible and otherwise at random; four specified activities (virtual meeting, curriculum vitae development, personal statement review, mock interview)	Self-reported preparedness for the match; match outcome	Felt prepared or extremely prepared 16.7% → 87.6%, P < 0.001; 14/21 applicants (66.7%) matched; 15/25 mentees completed all four activities; 92.3% of mentors intended to continue	A
18	Fallatah 2018	One mentor assigned per ten fourth-year students by Excel-based random selection; initial group meeting followed by individual meetings; one-day mentor training workshop	Multiple-choice and clinical examination scores; identification of students needing support	**Null on academic performance**: multiple-choice score 11.12 versus 10.76, P = 0.35; clinical examination 20.33 versus 19.91, P = 0.42. Mentor seniority P = 0.024 and mentor motivation P = 0.002 were associated with identifying students requiring special support; 13.2% of students were referred	A
19	Guhan 2020	148 first-year MBBS students divided among six faculty mentors by roll number; minimum 30 minutes per student per week over six months; mentors sensitised to programme guidelines and to delivering effective feedback	Internal assessment marks	Marks 22.8 ± 8.1 → 25.3 ± 7.7, P < 0.001. Below-average subgroup (n = 75) 16.3 ± 5.1 → 20.5 ± 6.4, P < 0.001; above-average subgroup (n = 73) 29.4 ± 4.1 → 30.2 ± 5.7, P = 0.221. Girls P < 0.001; boys P = 0.273	A
20	Andre 2017	"Mentors in Medicine" vertical peer mentoring within the Veritas learning-community structure; mentors selected from rising third-year students and assigned to serve their own group; minimum monthly subgroup meetings, combined cross-class sessions and twice-yearly full-group events; no formal initial mentor training, with an optional year-long 10-hour leadership elective added in 2013	All-student satisfaction with the learning-community experience; mentor-reported preparedness and perceived value	All-student survey responses 127/439 (2011), 183/463 (2012), 198/440 (2013), 162/440 (2014). Student satisfaction rose significantly from 2011 to 2012, 2011 to 2013, 2011 to 2014 and 2012 to 2014 (all P < 0.05, Scheffé post-hoc). Mentor responses 43/61 (2012), 41/58 (2013), 48/62 (2014); 91.7% would volunteer again; agreement that contributions were valued rose from 27.9% to 60.4%; felt prepared leading meetings 71%, giving guidance 85%, individual mentoring 81%	A

GME = graduate medical education; PGY = postgraduate year.

Risk of bias

Risk of bias assessments are summarised in Table 3 and displayed graphically in Fig. 2.

Table 3

Risk of bias and study quality assessment. RoB 2 was applied to the randomised study and ROBINS-I to the non-randomised studies with a comparator. NOS Ottawa Scale (selection, comparability, outcome; maximum 9 stars), reported as a descriptive supplement only.

Study	Instrument	NOS S (4)	NOS C (2)	NOS O (3)	NOS total (9)	D1	D2	D3	D4	D5	D6	D7
Pololi 2023	RoB 2	n/a	n/a	n/a	n/a	Some concerns	Some concerns	Some concerns	High	Some concerns	—	—
Indyk 2011	ROBINS-I	3	1	3	7	Serious	Serious	Moderate	Moderate	Moderate	Low	Moderate
Day 2016	ROBINS-I	2	0	2	4	Serious	Serious	Moderate	Moderate	Serious	Low	Serious
Zuzuarregui 2015	ROBINS-I	2	0	2	4	Critical	Serious	Moderate	Moderate	Serious	Moderate	Serious
Hirsch 2015	ROBINS-I	2	0	2	4	Critical	Critical	Low	Moderate	Moderate	Low	Moderate
Nellis 2016	ROBINS-I	2	0	2	4	Critical	Serious	Moderate	Moderate	Serious	Moderate	Serious
Sobbing 2015	ROBINS-I	2	0	1	3	Critical	Serious	Moderate	Moderate	Serious	Serious	Serious
Johnson 1998	ROBINS-I	1	0	1	2	Critical	Serious	Serious	Serious	Critical	Serious	Critical
Cohee 2015	ROBINS-I	2	1	2	5	Serious	Serious	Moderate	Moderate	Serious	Moderate	Moderate
Morrison 2014	ROBINS-I	4	2	3	9	Moderate	Moderate	Low	Low	Low	Low	Low
Efstathiou 2018	ROBINS-I	3	1	3	7	Serious	Moderate	Low	Moderate	Moderate	Low	Moderate
Witten 2021	ROBINS-I	2	1	2	5	Serious	Serious	Moderate	Moderate	Serious	Moderate	Moderate
Zhang 2017	ROBINS-I	1	0	2	3	Critical	Serious	Low	Moderate	Serious	Moderate	Moderate
Eleje 2025	ROBINS-I	3	1	3	7	Serious	Moderate	Low	Moderate	Moderate	Low	Moderate
Boedeker 2024	ROBINS-I	1	0	1	2	Critical	Serious	Moderate	Serious	Serious	Serious	Moderate
Fallatah 2018	ROBINS-I	2	1	3	6	Serious	Serious	Moderate	Moderate	Moderate	Low	Low
Guhan 2020	ROBINS-I	1	0	2	3	Critical	Moderate	Moderate	Moderate	Low	Low	Moderate
Altonji 2019	AXIS / MMAT	n/a	n/a	n/a	n/a	Not applicable — single-arm design; appraised separately (Supplementary Table S5)						
Wasim 2025	AXIS / MMAT	n/a	n/a	n/a	n/a	Not applicable — single-arm design; appraised separately (Supplementary Table S5)						
Andre 2017	AXIS / MMAT	n/a	n/a	n/a	n/a	Not applicable — single-arm design; appraised separately (Supplementary Table S5)						

RoB 2 domains: D1 randomisation process; D2 deviations from intended interventions; D3 missing outcome data; D4 measurement of the outcome; D5 selection of the reported result. ROBINS-I domains: D1 confounding; D2 selection of participants; D3 classification of interventions; D4 deviations from intended interventions; D5 missing data; D6 measurement of outcomes; D7 selection of the reported result. MMAT = Mixed Methods Appraisal Tool. Item-level judgements with supporting quotations are provided in Supplementary Table S5.

Among the sixteen non-randomised studies with a comparator, one was judged at moderate overall risk of bias, seven at serious risk and eight at critical risk; fifteen of sixteen were therefore at serious or critical risk. The single randomised study was judged at high risk of bias under RoB 2.

Morrison and colleagues remained the only study at moderate risk and the only study to achieve the maximum nine stars on the Newcastle–Ottawa Scale [80]. It used an institutional administrative census rather than a voluntary survey, matched mentored and non-mentored faculty for covariates, employed both concurrent and historical comparisons, and conducted post hoc analyses designed to distinguish the effect of mentor assignment from secular trend. It was also the study with the least ambiguous outcome measure, since date of promotion is recorded administratively and is not subject to recall, social desirability or non-response bias.

The randomised trial merits separate comment, because its high risk-of-bias rating arises from a source that randomisation cannot address [77]. Allocation to the mentoring institute or to a waitlist could not be concealed from participants, and every primary outcome — vitality, self-efficacy, empathy — is self-reported by those participants. A trial of an intensive, visible, year-long group intervention, measured entirely by participant self-report, is structurally vulnerable to expectancy effects in a way that a trial with an administratively ascertained endpoint would not be. Its single objectively ascertained outcome, objective vitality, was the one outcome that did not reach significance (D = 0.31, P = 0.20). This pattern does not invalidate the trial, but it does mean that the strongest design in this literature does not yet deliver the strongest evidence.

Discordance between instruments was material and is reported rather than suppressed. Indyk and colleagues scored 7 of 9 on the Newcastle–Ottawa Scale, within the pre-specified low-risk band, yet were judged at serious risk under ROBINS-I, because self-selection into a voluntary four-year programme is not captured by the Newcastle–Ottawa selection domain [78]. Eleje and colleagues and Efstathiou and colleagues showed the same pattern for the same reason [83, 81]. Where the instruments disagreed the ROBINS-I judgement was carried forward.

The dominant bias across the non-randomised studies was confounding by indication, which in this literature takes a specific and severe form. In most programmes, participation was voluntary and self-initiated. Students who enrol in a longitudinal primary-care mentorship programme in their first year, and who persist in it for four years, differ systematically from their classmates in precisely the attribute the outcome measures, namely commitment to primary care. Indyk and colleagues provide the clearest evidence of this: within their own cohorts, non-completers matched to primary care at 53.8% and 37.5%, closely comparable to the school-wide rates of 55.8% and 35.9%, whereas completers matched at 87.5% and 78.9% [78]. Comparison of completers with the whole class therefore conflates the effect of the programme with the effect of the persistence that defines completion. The same logic applies with greater force to Hirsch and colleagues, where the analytic denominator is the 17 of 58 participants who chose to apply to radiation oncology, a group that has selected on the outcome by the act of applying [88].

Two further confounding structures were introduced by the newly included studies and warrant explicit statement. The first is regression to the mean. Eleje and colleagues defined mentees as students scoring below 45% on a continuous assessment and controls as students scoring 50–70% [83]; a group selected for extreme low performance will improve on retest whether or not it is mentored, and the study contains no adjustment for this. The reported difference in favour of mentoring is therefore an upper bound on any programme effect. The second is maturation. Guhan and colleagues compared marks obtained by first-year MBBS students before and after a six-month mentoring programme, over an interval in which those students also received an entire semester of instruction [93]; the design cannot separate the programme from the curriculum. Zhang and colleagues observed eight residents over twelve months during which those residents also advanced a full postgraduate year [90]. These three studies are rated at serious or critical risk on domain 1 for these reasons rather than for self-selection.

A third recurring problem is comparator construction. Four studies compared post-implementation performance with a historical institutional baseline. Over the intervals concerned, national specialty application patterns, institutional recruitment, curricular structure and the competitiveness of the match all changed substantially. The increase from 2 to 11 applications reported by Sobbing and colleagues represents a large relative change on a base of two, at a single institution, over an unspecified interval, without adjustment [85], and cannot be interpreted as an effect estimate.

A fourth is incomplete reporting. Denominators were not reported for the primary comparison in two studies [79, 85]; measures of dispersion were not reported in three [84, 88, 89]; and one study reported the direction of change in three of five clerkships without any numerical result [86]. The reporting of the clerkships in which honours increased, without reporting those in which they did not, is the pattern that domain 7 of ROBINS-I is designed to detect, and contributes to the critical risk rating for that study.

Domain 1: career and residency match outcomes

Eight studies reported outcomes in this domain. Four studies provided a numerator and denominator for match success among participants who applied: Indyk [78], Hirsch [88], Nellis [89] and Boedeker [91]. Indyk contributed two estimates, from consecutive entry cohorts with non-overlapping participants and distinct match years, which were treated as separate estimates for this single-arm analysis only; a sensitivity analysis combining them into one estimate is reported in Table 5. Because the five estimates pool applicants to four different specialties, with different baseline competitiveness, this analysis is reported as exploratory rather than as a confirmatory pooled effect. Results are presented in Fig. 3.

The pooled proportion of successful match among participants who applied was 80.0% (95% CI 66.2–89.1), based on 66 matches among 80 applicants. Heterogeneity was low but no longer nil (I^2 = 22.0%, τ^2 = 0.148, Q = 5.13, df = 4, P = 0.27), and the 95% prediction interval was 42.6–95.6%. Leave-one-out analysis produced pooled estimates ranging from 76.6% to 85.0%, with τ^2 rising as high as 0.479 and three-estimate prediction intervals spanning effectively the entire range (Table 5). Robustness analyses moved the estimate or its interval materially: Hartung–Knapp–Sidik–Jonkman adjustment widened the interval to 59.2–91.7%, a factor of 1.42 on the logit scale; restricted maximum likelihood returned τ^2 = 0.132 and 79.9% (66.3–88.9); the binomial–normal generalised linear mixed model gave 84.8% (95% CI 70.8–92.8) with τ^2 = 0.387; and combining the two cohorts of Indyk into a single estimate gave 80.5% (63.9–90.6). Re-analysis substituting the alternative year 2 numerator identified during data verification (14/19 rather than 15/19) gave 78.6% (64.3–88.2).

The addition of a single study changed this estimate more than any analytical choice. In the previous version of this synthesis, based on four cohorts from three studies, the pooled proportion was 85.0% (72.1–92.5) with $I^2 = 0.0\%$ and $\tau^2 = 0.000$. Adding Boedeker and colleagues, whose military obstetrics and gynaecology applicants matched at 14/21 (66.7%) [91], lowered the pooled estimate by 5.0 percentage points and introduced the first non-zero between-study variance in this analysis. The leave-one-out analysis confirms this directly: omitting Boedeker restores the estimate to exactly 85.0% with $I^2 = 0.0\%$. An apparent homogeneity that dissolves on the addition of one study of 21 applicants was never evidence of a consistent programme effect.

Five qualifications govern the interpretation of this estimate. It is a single-arm proportion rather than a comparative effect. It pools applicants to primary care, radiation oncology, otolaryngology and obstetrics and gynaecology, four specialties differing substantially in competitiveness, so the pooled quantity is not a single construct. The pooled proportion lies below, not at, the contemporaneous United States allopathic senior match rate of approximately 93.9% [97], although the two are not strictly comparable. Every contributing study is at serious or critical risk of bias, and the pre-specified sensitivity analysis restricted to studies at moderate or lower risk therefore contains no studies at all. Most importantly, low observed heterogeneity does not indicate a consistent programme effect: samples of students sufficiently committed to a specialty to persist in a mentorship programme and then apply to that specialty would be expected to produce similar match rates. I^2 near zero in the presence of a bias shared across studies indicates consistent bias rather than absent bias, and the prediction interval, which extends to 42.6%, is the more appropriate summary of what a newly implemented programme might observe.

The comparative evidence in this domain now rests on two studies that point in opposite directions.

Indyk and colleagues provided concurrent institution-level comparators for both cohorts [78]. In the year 1 cohort, 14/16 completers (87.5%) matched to primary care versus 82/147 remaining classmates (55.8%): risk ratio 1.57 (95% CI 1.24–1.98), risk difference 31.7 percentage points (95% CI 13.6–49.8), Fisher exact $P = 0.016$. In the year 2 cohort, 15 of the 19 completers with available match data (78.9%) matched versus 52/145 remaining classmates (35.9%): risk ratio 2.20 (95% CI 1.60–3.03), risk difference 43.1 percentage points (95% CI 23.2–63.0), Fisher exact $P = 0.0008$.

Witten and Maskarinec evaluated a six-month Primary Care Mentorship Program for first-year students at a single institution and found that participation “does not correlate with an increased likelihood of matching into a primary care GME residency specialty” [82]. This is the same intervention concept, the same outcome and the same learner level as Indyk, and it reaches the opposite conclusion. The two studies differ in dose — four years of longitudinal faculty mentoring with funded research, national meeting attendance and mentor stipends, against six months of first-year mentoring — and that contrast is itself the most informative comparison available in this domain. It is also consistent with the internal evidence within Indyk’s own data, discussed next.

The internal comparison within the Indyk dataset remains the most informative single analysis in the domain. Completers versus non-completers gave a risk ratio of 1.62 (95% CI 0.95–2.78, $P = 0.09$) in year 1 and 2.11 (95% CI 0.84–5.31, $P = 0.07$) in year 2. Both point estimates favour completion, neither reaches conventional statistical significance, and both confidence intervals include the null. Non-completers shared the initial self-selection into the programme and differ from completers principally in persistence, so the contrast between the significant comparison with classmates and the non-significant comparison with non-completers is material to interpretation.

Comparative meta-analysis was not performed in this domain. Only two studies report a comparative career outcome, below the pre-specified threshold of three; and the two are not commensurable, one reporting a risk ratio against concurrent classmates and the other reporting the absence of a correlation without an extractable effect estimate.

The remaining career outcomes could not be pooled and were synthesised narratively. Hirsch and colleagues reported a 100% match rate among 17 applicants against a national benchmark of 85.1%; the published comparison was not statistically significant ($P = 0.14$), and an exact one-sample binomial test against that benchmark gives $P = 0.064$ [88]. Nellis and colleagues reported 6/7 (85.7%, 95% CI 42.1–99.6) matching into otolaryngology and an increase in institutional representation of residents underrepresented in medicine from 9% to 16% [89]; denominators for the latter were not reported, so neither a confidence interval nor a test is computable. Boedeker and colleagues reported that self-rated preparedness for the match rose from 16.7% to 87.6% ($P < 0.001$), an internal pre–post comparison in a self-selected group with a 51.8% response rate [91]. Zuzuarregui and Hohler reported an increase in students entering neurology from 2.8 to 7.5 per year ($P < 0.05$) [84], and Sobbing and colleagues an increase in applications to the local obstetrics and gynaecology residency from 2 to 11 [85]; both are uncontrolled pre–post institutional counts without denominators. Johnson and colleagues reported improvement in the proportion of students underrepresented in medicine obtaining honours in three of five clerkships, without numerical detail [86].

Certainty of evidence for career and match outcomes was rated very low: starting at low for predominantly non-randomised design, rated down one level for serious risk of bias arising from self-selection and historical comparators, and one level for inconsistency, since the only two comparative studies disagree in direction. No upgrade for magnitude was applied because magnitude cannot be separated from selection.

Domain 2: research productivity

Three studies reported research productivity outcomes [84, 88, 81]. Hirsch and colleagues reported cumulative programme output over ten years of 53 publications, 75 national presentations and 10 national awards among 58 participants [88]; Zuzuarregui and Hohler reported an increase in student research outputs from 7 to 22 [84]; and Efstathiou and colleagues reported a median external research funding of US$1.78M among mentees against US$0.50M among non-enrolled controls [81].

Meta-analysis was not possible for three separately sufficient reasons: the three outcome constructs are not commensurable, comprising a cumulative ten-year institutional total without a comparator, a pre–post count over an unstated number of years, and a median funding figure with a concurrent comparator; two of the three reported no denominator-linked rate or measure of dispersion, so no variance and therefore no weight could be computed; and the funding comparison is reported as medians without an interquartile range or a test statistic.

All three studies report substantially greater scholarly output among programme participants. Two of the three programmes brokered research opportunities in addition to mentorship, and this co-intervention is the principal threat to inference, since the effect of the mentoring relationship cannot be distinguished from the effect of privileged access to research projects, laboratory placements and authorship that the programme also conferred. The Efstathiou comparison is less exposed to this objection, because the mentoring programme in that study did not itself allocate research resources, but it remains subject to confounding by academic ambition: faculty who enrol in a voluntary mentoring programme may be those already most likely to seek funding. The finding is consistent with the wider literature associating mentorship with research productivity in academic medicine [1, 27]. Certainty of evidence was rated very low.

Domain 3: academic performance

Five studies reported academic performance outcomes, and the addition of three of them changes the character of this domain from a two-study fragment to the domain with the most informative internal contrast in the review.

Day and colleagues reported that 50% of participants in a junior surgical mentorship programme received honours in the surgical clerkship compared with 31.5% of non-participants (P = 0.05), an absolute difference of 18.5 percentage points and an implied odds ratio of 2.17 [79]. Johnson and colleagues reported an increase in the proportion of students underrepresented in medicine obtaining honours in three of five required third-year clerkships, without numerical detail [86].

Eleje and colleagues, in the only controlled study in this domain, reported that mentees' median continuous assessment score rose from 40.0% to 70.0% while concurrent controls fell from 60.0% to 57.0% (P = 0.003) [83]. Guhan and colleagues reported that mean marks among 148 first-year MBBS students rose from 22.8 ± 8.1 to 25.3 ± 7.7 (P < 0.001), with the gain concentrated in the below-average subgroup (16.3 → 20.5, P < 0.001) and absent in the above-average subgroup (29.4 → 30.2, P = 0.221) [93]. Fallatah and colleagues found no effect on either multiple-choice (11.12 versus 10.76, P = 0.35) or clinical examination scores (20.33 versus 19.91, P = 0.42) [92].

Meta-analysis was not possible: the denominators required to reconstruct a 2 × 2 table were not reported by Day and colleagues; Johnson and colleagues reported no numerical result; Eleje and colleagues reported medians without dispersion; Guhan and colleagues reported a single-group pre–post change; and Fallatah and colleagues reported group means without standard deviations. No two of the five studies report the same outcome on a common scale with the variance components required for pooling.

Read together, however, the domain yields a coherent and testable pattern. The two studies reporting benefit in an unselected population are both uncontrolled or confounded: Day compares self-selected participants with non-participants on a subjectively graded outcome assigned by faculty who may have supervised the resident mentors, and Guhan compares first-year students before and after six months of teaching. The two studies with the most defensible comparisons diverge according to programme intensity and target population. Eleje's intensive intervention — weekly meetings over six months, six hours of mentor training, mentees selected for academic difficulty — produced a large effect, though one substantially inflated by regression to the mean. Fallatah's low-intensity intervention — one mentor per ten students, an initial group meeting, an unselected cohort — produced none. Guhan's subgroup analysis points the same way: the benefit was confined to students who were struggling. The tentative inference, generated by rather than tested in these data, is that mentorship may affect academic performance where academic difficulty is the reason for referral and the dose is substantial, and not otherwise.

Certainty of evidence was rated very low, rated down for risk of bias, inconsistency and imprecision.

Domain 4: satisfaction, mentoring relationship and wellbeing

Twelve studies reported outcomes in this domain, which remains the largest by study count and the least coherent by construct, replicating the pattern documented by Farkas and colleagues [8].

The reported outcomes were perceived mentor influence on career choice, 25/41 (61.0%) [78]; mean satisfaction 4.85 on a 5-point scale [89]; 56% receiving personal development advice and 40% feeling well or extremely well connected, with denominators unreported [85]; prevalence of any identified mentor rising from 13/26 (50%) to 25/26 (96%), P < 0.001, with 14/20 (70%) of those choosing a formal mentor identifying that person as a genuine mentor [87]; mean overall satisfaction 7.47 (SD 2.45) on a 10-point scale [94]; 327/362 (90.3%) reporting positive experiences with their mentor [95]; and 41% continuing the mentoring relationship beyond the end of the programme [82]. Andre and colleagues reported a significant year-on-year rise in all-student satisfaction with a vertical peer-mentoring structure across four annual survey waves, from 127/439 responses in 2011 to 162/440 in 2014 (all pairwise comparisons against 2011 significant at P < 0.05 with Scheffé correction) [96].

A pre-specified pooled proportion of favourable mentee-reported programme experience was calculated using the three studies reporting an explicit numerator and denominator for a comparable construct [78, 87, 95]. The pooled proportion was 85.7% (95% CI 57.5–96.4), with Q = 25.70, df = 2, P < 0.0001, I^2 = 92.2%, τ^2 = 1.44 and a 95% prediction interval spanning effectively the entire range; under Hartung–Knapp–Sidik–Jonkman adjustment the interval was 17.1–99.4%. This estimate was judged uninterpretable and was not carried forward. The heterogeneity arises from pooling three distinct constructs — whether a mentor influenced a career decision (61.0%), whether a mentor of any kind existed (96.2%), and whether the mentoring experience was rated positively (90.3%) — which are not three measurements of a single quantity. The result is reported to document why satisfaction outcomes in this literature cannot be aggregated.

Two findings within this domain now rest on more than a single study.

The first concerns access. Cohee and colleagues found that formalisation approximately doubled the proportion of internal medicine residents with any identified mentor, from 50% to 96% (risk ratio 1.92, 95% CI 1.30–2.85; absolute increase 46.2 percentage points, 95% CI 25.6–66.7), and that 70% of those

who selected a formal mentor described that person as a genuine mentor rather than a nominal assignment [87]. This remains the single most defensible finding in the review, and it remains the finding of one study of twenty-six respondents at serious risk of bias. A second study reporting the same outcome in the same direction is among the records awaiting classification, so the evidential position on this point may change once that full text is obtained.

The second concerns wellbeing, which the previous version of this synthesis reported as entirely unmeasured. Two studies measured wellbeing with validated instruments. Zhang and colleagues administered the Perceived Stress Scale [98] and the Maslach Burnout Inventory [99] to eight otolaryngology residents before and twelve months after the introduction of a formal tiered mentorship programme, reporting a fall in perceived stress from 18.5 to 7.9 (P = 0.001) and improvements across all three burnout subscales (all P < 0.0001) [90]. Pololi and colleagues, in the randomised trial, reported a large effect on perceived vitality (*D* = 1.47, P = 0.03) [77].

Neither result should be read as establishing that formal mentorship improves wellbeing. The Zhang study has eight participants, no concurrent comparator, and a twelve-month observation window over which residents also advance a postgraduate year; effect sizes of the magnitude reported — emotional exhaustion falling by two thirds — are implausibly large for a monthly mentoring meeting and are more consistent with regression from an unusually stressed baseline compounded by maturation. The Pololi result is from a randomised comparison, but is self-reported by unblinded participants in an intensive visible intervention, and the one objectively ascertained vitality measure in the same trial was null. What can be said is narrower and still useful: wellbeing outcomes have now been measured in this literature, with validated instruments, in two studies at opposite ends of the design spectrum, and both point in the same direction. The appropriate conclusion is that this is a tractable research question, not that it has been answered.

A third observation is that not every satisfaction-adjacent outcome favoured mentorship. Eleje and colleagues measured subjective vitality and student engagement alongside academic performance and found no difference on either (P = 0.722 and P = 0.388) in the same study in which the academic effect was large and significant [83]. This dissociation — an academic effect without an accompanying affective effect — is difficult to reconcile with a mechanism operating through motivation or wellbeing, and is worth noting precisely because studies that measure several constructs and report the null ones are rare in this field.

Read individually, the domain remains consistent in direction. Learners in formal programmes reported favourable experiences at rates of approximately 60–96%, across faculty mentorship, resident mentorship and peer mentorship, and in high-income, upper-middle-income and lower-middle-income settings. The two vertical and cross-year peer programmes [94, 96] achieved this at explicitly low resource cost. Altonji and colleagues found that mentor expertise in the mentee's area of need predicted all benefit outcomes and that regularity of contact predicted academic and non-academic benefit [94], a finding about mechanism that supports needs-based matching over administratively convenient assignment. The study by Wasim and colleagues reports the highest-precision proportion in the review (90.3%, 95% CI 86.8–93.2, n = 362), with benefit domains — accessibility, confidence building, ethics, study strategy, professionalism, psychological support and communication — that map onto the psychosocial rather than the instrumental functions of mentoring [95]. Its single-arm cross-sectional design establishes acceptability and perceived benefit rather than effectiveness.

Certainty of evidence was rated very low for satisfaction as an effectiveness outcome, very low for the claim that formalisation increases mentor prevalence, and very low for wellbeing.

Domain 5: faculty advancement

Two studies reported faculty advancement outcomes, and a third reported self-efficacy in career advancement without measuring advancement itself.

Morrison and colleagues analysed institutional promotion records for 611 full-time faculty in the Department of Medicine at the University of Toronto from 1988 to 2010, comparing 382 faculty appointed before implementation of a formal mentoring programme in 2003 with 229 appointed subsequently, and separately comparing mentored with non-mentored faculty matched for covariates [80]. Faculty assigned a mentor were promoted a mean of 1.0 year sooner than those who were not (3.4 years, SD 2.4, versus 4.4 years, SD 2.6; P = 0.011). From the reported P value, the standard error is 0.39 years and the approximate 95% confidence interval − 1.77 to − 0.23 years; the approximate standardised mean difference is − 0.40. The cohort comparison showed that faculty appointed in 2003 or later were promoted 1.2 years sooner than those appointed earlier (2.5 years, SD 2.0, versus 3.7 years, SD 1.7; mean difference − 1.2 years, 95% CI − 1.51 to − 0.89; standardised mean difference − 0.66). Gender effects were not statistically significant, and post hoc analyses indicated that the observed differences were attributable to mentor assignment rather than to temporal trend.

Efstathiou and colleagues followed 23 mentees in a structured faculty mentoring programme at Massachusetts General Hospital against 91 non-enrolled concurrent controls over seven years [81]. Mentees were more likely to hold senior faculty positions at long-term follow-up (47% versus 13%, P = 0.030), showed comparable retention (75% versus 72%), and held higher median external research funding (US$1.78M versus US$0.50M). Matching in that programme was structured rather than arbitrary: mentees ranked five development priorities, mentors listed areas of expertise, and senior faculty facilitated the final pairing.

Meta-analysis was not possible because two studies fall below the pre-specified threshold of three, and because the outcome metrics differ — time to promotion in years against attainment of senior rank at a fixed follow-up.

These are the two methodologically strongest observational studies in the review and the only two reporting administratively verified, institutionally consequential outcomes with defensible comparators. They agree in direction and are broadly consistent in magnitude: promotion approximately one year sooner, and a roughly threefold higher probability of senior rank at seven years. Their principal residual limitations are shared. In neither study was mentoring randomly allocated, and faculty who engage with mentoring may differ in academic ambition; promotion criteria may have changed over the observation windows, which span 22 and 7 years respectively; and both derive from single departments at well-resourced North American academic centres. Efstathiou's 23 mentees are also few, and the 47% versus 13% contrast rests on small numbers.

Certainty of evidence was rated low: starting at low for non-randomised design, not rated down further because both studies use objectively ascertained outcomes with concurrent comparators and agree in direction, and not rated up because neither controls for confounding by academic ambition.

Summary of findings

Findings across all five domains are summarised in Table 4, and the direction of effect for every study within every domain, stratified by risk of bias, is displayed in Fig. 4.

Table 4
Summary of findings by outcome domain.

Outcome domain	Studies	Participants (approx.)	Meta-analysis	Pooled or representative result	Heterogeneity	Direction of effect	GRADE certainty
1. Career and residency match	8	~ 450 with extractable outcomes	Yes; single-arm proportion, 5 cohorts from 4 studies	Pooled proportion matching among applicants 80.0% (95% CI 66.2–89.1); prediction interval 42.6–95.6. Hartung–Knapp–Sidik–Jonkman 59.2–91.7; generalised linear mixed model 84.8% (70.8–92.8); combined-cohort sensitivity 80.5% (63.9–90.6). Comparative data from two studies: risk ratios 1.57 (1.24–1.98) and 2.20 (1.60–3.03) in one; null in the other	I^2 = 22.0%; τ^2 = 0.148; Q = 5.13, P = 0.27	Predominantly favours mentorship; one null study; magnitude inseparable from self-selection	Very low
2. Research productivity	3	≥ 172	No; non-commensurable constructs, no dispersion reported	53 publications, 75 presentations, 10 awards over 10 years; outputs 7 → 22; median funding US$1.78M versus US$0.50M	Not estimable	Favours mentorship; confounded by brokered research access and academic ambition	Very low
3. Academic performance	5	~ 530	No; no two studies share a scale with estimable variance	Honours 50% versus 31.5%, P = 0.05; continuous assessment 40% → 70% versus 60% → 57%, P = 0.003; marks 22.8 → 25.3, P < 0.001; examination scores null (P = 0.35 and P = 0.42)	Not estimable	Mixed: three favour mentorship, one null; benefit concentrated in academically struggling students	Very low
4. Satisfaction, relationship and wellbeing	12	~ 1,600	Attempted; not interpretable (I^2 = 92.2%, τ^2 = 1.44) owing to non-commensurable constructs	Favourable experience 61.0–96.2% across studies; mentor prevalence 50% → 96%, P < 0.001; Perceived Stress Scale 18.5 → 7.9, P = 0.001; perceived vitality *D* = 1.47, P = 0.03; all-student satisfaction rising across four annual waves, all P < 0.05; vitality and engagement null in one controlled study	I^2 = 92.2%	Predominantly favourable; one study null on affective outcomes	Very low
5. Faculty advancement	2	725	No; fewer than 3 studies and differing metrics	Time to promotion 3.4 versus 4.4 years; mean difference − 1.0 year (95% CI − 1.77 to − 0.23), standardised mean difference − 0.40. Senior rank at 7 years 47% versus 13%, P = 0.030	Not applicable	Favours mentorship; consistent across two studies with objective outcomes	Low

Seventeen of the twenty included studies reported at least one result favouring formal mentorship. Three studies reported a null result on at least one outcome: on residency match [82], on examination performance [92], and on subjective vitality and student engagement [83]. The previous version of this synthesis reported that no study had reported a null or adverse effect; that statement is superseded. No study reported an adverse effect.

Sensitivity and subgroup analyses

Pre-specified sensitivity analyses for the primary meta-analysis are presented in Table 5.

Table 5
Sensitivity analyses for the primary meta-analysis of the pooled match proportion (Domain 1).

Analysis	Estimates	Pooled proportion (95% CI)	I^2	τ^2	Change versus primary
Primary: logit transformation, DerSimonian–Laird	5	80.0% (66.2–89.1)	22.0%	0.148	Reference
Previous version of this synthesis (before the supplementary search)	4	85.0% (72.1–92.5)	0.0%	0.000	+ 5.0 percentage points; apparent homogeneity was an artefact of the smaller evidence base
Alternative transformation: Freeman–Tukey double arcsine	5	86.0% (70.0–97.1)	61.5%	0.061	+ 6.0 percentage points; back-transformation rejected as unreliable with unequal study sizes
Verification sensitivity: year 2 numerator 14/19	5	78.6% (64.3–88.2)	24.2%	—	−1.4 percentage points; interpretation unchanged
Leave-one-out: omit Indyk year 1 cohort	4	78.5% (61.0–89.5)	29.1%	0.220	−1.5 percentage points; prediction interval 19.0–98.3
Leave-one-out: omit Indyk year 2 cohort	4	82.8% (62.5–93.3)	41.5%	0.479	+ 2.8 percentage points; prediction interval 9.9–99.5
Leave-one-out: omit Hirsch 2015	4	76.6% (64.0–85.7)	0.0%	0.000	−3.4 percentage points; prediction interval 46.3–92.5
Leave-one-out: omit Nellis 2016	4	80.3% (63.3–90.6)	38.7%	0.288	+ 0.3 percentage points; prediction interval 17.2–98.8
Leave-one-out: omit Boedeker 2024	4	85.0% (72.1–92.5)	0.0%	0.000	+ 5.0 percentage points; recovers the previous estimate exactly
Restriction to studies at moderate or lower risk of bias	0	Not estimable	—	—	No contributing study was at less than serious risk of bias
Restriction to studies with a concurrent comparator	2 (one study)	Not performed	—	—	Independence assumption not met
Hartung–Knapp–Sidik–Jonkman adjustment (ad hoc and plain variants coincide)	5	80.0% (59.2–91.7)	22.0%	0.148	Interval 1.42 times wider than primary on the logit scale; principal robustness concern
Restricted maximum likelihood estimation of τ^2	5	79.9% (66.3–88.9)	22.0%	0.132	−0.1 percentage points
Binomial–normal generalised linear mixed model (no continuity correction)	5	84.8% (70.8–92.8)	—	0.387	+ 4.8 percentage points; removes the 0.5 correction applied to the 17/17 cell
Cohorts of a single study combined into one estimate	4	80.5% (63.9–90.6)	37.2%	—	+ 0.5 percentage points
Restriction to programmes whose principal component was mentorship	4	76.6% (64.0–85.7)	0.0%	0.000	−3.4 percentage points; excludes the research-brokerage programme

The estimate is moderately robust to the choice of transformation, within a range of approximately 77–86%, and to the removal of any single cohort. It is not robust to the estimator of uncertainty: Hartung–Knapp–Sidik–Jonkman adjustment widens the interval by a factor of 1.42 on the logit scale, and the generalised linear mixed model, which avoids continuity correction of the boundary cell, returns 84.8%, indicating that the apparent precision of the primary analysis is partly an artefact of the DerSimonian–Laird variance and of the continuity correction. Nor is it robust to restriction by risk of bias: no study contributing to the analysis was at less than serious risk, so that sensitivity analysis contains no studies.

The most consequential entry in Table 5 is the second row. The pooled estimate reported in the previous version of this synthesis, 85.0% with $I^2 = 0.0\%$, is recovered exactly by omitting the single study added by the supplementary search. A pooled estimate whose central value and entire apparent homogeneity depend on the completeness of the search is not a robust estimate, and this is reported here as the strongest available evidence that the primary search was incomplete.

Subgroup analyses remained infeasible for the primary meta-analysis. All four pre-specified comparisons — learner level, mentor type, programme duration, and specialty-specific versus general programme — would have required at least two studies per subgroup within a single outcome. The primary meta-analysis contains four studies, all of medical students in faculty-mentored, specialty-specific programmes, providing no contrast to examine. Across the review as a whole the distribution is less unbalanced than previously: fourteen of twenty studies concern medical students, two concern residents, three concern faculty and one spans students and faculty; four concern peer mentoring and two resident mentoring; and four studies are now from outside North America, drawn from four countries. This is sufficient to describe the field but not to support formal subgroup meta-analysis within any single outcome.

Publication bias could not be assessed. Every pooled analysis contained fewer than ten estimates, below the threshold at which funnel plot asymmetry and Egger's test have adequate power and below the level at which they are recommended [71]. The prior risk of publication bias in this field is high: programme evaluations are typically published by programme directors, the incentive to publish a null institutional evaluation is weak, and seven conference abstracts without corresponding full papers were excluded during screening. The identification of three null results in the present revision, all of them in studies retrieved by the supplementary rather than the primary search, is consistent with null findings being harder to retrieve rather than absent from the literature.

Data verification

Verification of every extracted datum against its primary source identified discrepancies requiring correction, clarification or explicit qualification in nine of the twenty included studies. Three were material to interpretation: the omission of a non-significant P value for the radiation oncology match comparison [88]; the characterisation of a headline result as relationship retention rather than mentor prevalence [87]; and the characterisation of a comprehensive pipeline and faculty development programme as a medical student mentorship programme [86]. One primary study contains an internal inconsistency between its abstract and its results text in the year 2 match numerator [78], which was carried into a pre-specified sensitivity analysis (Table 5).

Verification levels are reported in Tables 1 and 2 and in full in Supplementary Table S4. Nine studies are at level A, verified against the primary full text through two independent retrieval routes with an identity check. Two are at level B, verified against the indexed abstract of the primary article with corroboration from the extraction table of a previous high-quality review [8]; because that review covers only United States medical students to 2019, it cannot corroborate the Canadian, Pakistani, Nigerian, Saudi or Indian studies, which were verified against the primary record alone. Nine are at level C, carried forward unchanged. Every included study carries entered, verified data; no included study rests on an unretrieved full text.

Two limitations of the verification process must be stated. First, one reported satisfaction statistic could not be independently confirmed from openly accessible sources [94]; it is retained in Table 2 with that qualification. Second, during the present revision an automated retrieval of one article's digital object identifier returned the full text of a different article together with plausible sample sizes and effect estimates; the error was detected only by an explicit title and author-list identity check, and no value from that retrieval entered any table. All level A extractions were subsequently subjected to that identity check as a matter of routine. We report this because the failure was silent and would not have been detected by inspecting the extracted values for plausibility.

Studies added in this revision, studies awaiting classification, and documented exclusions

Table 6 sets out the studies added to the review in the present revision, the records that remain awaiting classification, and the principal exclusions that a reader familiar with this field would expect to see accounted for.

Table 6
Studies added in the present revision, studies awaiting classification, and principal documented exclusions.

Study	Design and setting	Status and rationale
Pololi 2023	Stratified randomised controlled trial with waitlist control; 40 research faculty from 27 medical schools	**Added.** The only randomised study in the review. Supersedes the previous statement that no included study was randomised. Rated at high risk of bias under RoB 2 because unblinded participants self-reported every significant outcome
Efstathiou 2018	Prospective cohort, 23 mentees versus 91 concurrent non-enrolled controls, 7-year follow-up	**Added.** Provides a second faculty-advancement study with an objectively ascertained outcome, raising GRADE certainty for Domain 5 from very low to low
Witten 2021	Prospective evaluation of a 6-month programme, concurrent classmates as comparator	**Added.** The first null result on residency match. Supersedes the previous statement that no included study reported a null effect, and provides the direct counterweight to Indyk 2011 in Domain 1
Zhang 2017	Prospective interventional, 8 residents, internal pre–post at 12 months	**Added.** The first use of validated wellbeing instruments (Perceived Stress Scale, Maslach Burnout Inventory) in this literature. Supersedes the previous statement that no included study measured a validated wellbeing instrument
Eleje 2025	Quasi-experimental with concurrent control; 7 mentees, 12 peer mentors, 30 controls	**Added.** The only controlled evaluation from sub-Saharan Africa, the only study reporting both a significant academic effect and null affective outcomes in the same cohort, and the lowest-cost intensive model in the review
Boedeker 2024	Quasi-experimental pre–post; 25 mentees, 21 applicants	**Added.** Contributes a fifth cohort estimate to the primary meta-analysis and moves the pooled proportion by 5.0 percentage points, eliminating its apparent homogeneity
Fallatah 2018	Cross-sectional with internal comparison; 185 students, one mentor per ten	**Added.** The first null result on examination performance, and the clearest evidence in the review that a diluted mentor-to-student ratio does not deliver instrumental benefit
Guhan 2020	Single-group pre–post; 148 first-year MBBS students	**Added.** Contributes Indian evidence and a subgroup finding that academic benefit is confined to below-average students. Rated critical risk because the design cannot separate the programme from six months of curriculum
Andre 2017	Longitudinal single-arm programme evaluation; four annual survey waves, 670 all-student responses	**Added.** Contributes the largest satisfaction dataset in the review and a second low-cost peer-mentoring model. Single-arm, so appraised with AXIS rather than ROBINS-I
Awaiting classification — Kman 2013 [100]	Tiered mentorship programme; number of students with an identified mentor	Eligible on title and abstract, but the full text could not be obtained within the review period, so no numerator, denominator or test statistic could be verified. Would provide a second estimate for the review's most defensible finding
Awaiting classification — Guse 2016 [101]	Mixed-methods evaluation of mentoring speed dating as a matching tool	Eligible on abstract; full text not obtainable within the review period. Bears directly on the matching-method question identified in the Discussion as the highest-priority trial
Awaiting classification — Oelschlager 2011 [102]	Cross-sectional survey of the assigned mentor's role in students' support networks	Eligible on abstract; full text not obtainable within the review period
Awaiting classification — Abernethy 1999 [103]	Mentoring programme for students underrepresented in medicine	Eligible on abstract; full text not obtainable within the review period
Awaiting classification — Davey 2026 [104]	Prospective crossover randomised controlled trial; 43 medical students	Full text not obtainable within the review period, so neither the pairing method nor the outcome data could be verified. Inclusion would double the randomised evidence in this review
Excluded — Lewis 2016 [105]	Four-arm randomised trial; 150 mentor–protégé dyads at 11 sites	Protégés were 42% graduate students, 10% postdoctoral fellows and 48% junior faculty across biomedical and nursing disciplines, not reported separately for medicine; and the dyads were pre-existing, the randomised element being the addition of mentor training and peer groups to relationships that had already formed
Excluded — Bonilha 2019 [106]	Institution-wide brokered faculty mentoring; pre–post surveys 2011 and 2017, 656 and 706 respondents	Surveyed all six colleges of the institution; the mentor-prevalence outcome (45.3% → 67.1%, $P < 0.001$) is not reported separately for the College of Medicine. The closest call in the review: had it been included it would have supplied a second estimate for mentor prevalence
Excluded — Pfund 2014 [107]	Randomised trial of mentor training for clinical and translational research mentors	Outcomes are mentor-side competency ratings, which the protocol excludes
Excluded — Feeley 2022 [108]	Randomised trial of a specialty mentor effect	The mentorship consisted of a single supervised simulation session, with no specified cadence, duration or defined mentor role
Excluded — Sedigh 2024 [109]	Parallel randomised trial of peer mentoring	Operating-room technology students, an excluded population

Studies awaiting classification are recorded here, and in Supplementary Table S8 with their full citations, because each is eligible on the information available and each would, if obtained, bear on a specific claim in this review. Three of the five were cited in the discussion of the previous version of this synthesis without appearing in either the included set or the exclusion log; recording them explicitly is the correction of that omission. Their number relative to the twenty included studies is itself a limitation, and is discussed below.

Discussion

Principal findings

This systematic review identified twenty primary studies evaluating formal mentorship programmes in medical education and reporting an effectiveness outcome, together with five further records eligible on the information available but awaiting classification. Five conclusions follow.

First, the direction of association is predominantly but no longer uniformly favourable. Seventeen of twenty studies reported at least one result in the direction of benefit, across five outcome domains, three learner levels, six countries and twenty-eight years of publication. Three studies reported a null result: on residency match [82], on examination performance [92], and on subjective vitality and student engagement [83]. The previous version of this synthesis reported uniform consistency of direction; that finding did not survive a more complete search, and its disappearance is more informative than its presence was.

Second, the magnitude of the association cannot be estimated with confidence, because the design of nearly every study makes the effect of the programme inseparable from the effect of selection into it. The single-arm pooled proportion of successful match among applicants (80.0%, 95% CI 66.2–89.1) widens to 59.2–91.7% under Hartung–Knapp–Sidik–Jonkman adjustment and moves to 84.8% under a generalised linear mixed model; it lies below the contemporaneous overall United States allopathic senior match rate of approximately 93.9% [97]; it describes a highly selected population; and it pools four different specialties. Every study contributing to it is at serious or critical risk of bias.

Third, the pooled estimate is not robust to the completeness of the search. Adding one study of 21 applicants moved the pooled proportion by 5.0 percentage points and converted apparent perfect homogeneity ($I^2 = 0.0\%$, $\tau^2 = 0.000$) into measurable between-study variance ($I^2 = 22.0\%$, $\tau^2 = 0.148$). The leave-one-out analysis recovers the previous estimate exactly. Where a pooled estimate is derived from four or five small cohorts, sensitivity to the addition of a single study is a more important property than sensitivity to the choice of variance estimator, and it is not captured by any conventional robustness analysis.

Fourth, the most credible findings concern access and advancement rather than performance. Formalisation approximately doubled the proportion of internal medicine residents with any identified mentor, from 50% to 96% ($P < 0.001$), without evident loss of perceived authenticity, since 70% of residents who chose a formal mentor described that person as a genuine mentor [87]; this remains a single study of twenty-six respondents, and a second study reporting the same outcome is among the records awaiting classification [100]. Formal, institutionally resourced faculty mentoring programmes with defensible comparators and administratively verified outcomes were associated with promotion approximately one year sooner (mean difference − 1.0 year, 95% CI − 1.77 to − 0.23) [80] and with a threefold higher probability of senior faculty rank at seven years (47% versus 13%, $P = 0.030$) [81]. This second finding now rests on two studies that agree in direction, and in each the outcome is objectively ascertained.

Fifth, the evidentiary infrastructure of the field is weak in specific and remediable ways, though less uniformly than previously reported. Denominators were absent for the primary comparison in two studies [79, 85]; measures of dispersion were absent in three; one study reported only the direction of change in three of five outcomes [86]; and only one study in twenty was randomised. Wellbeing, previously reported as entirely unmeasured, has now been measured with validated instruments in two studies [90, 77], though neither can establish an effect.

What the addition of a randomised trial does and does not settle

The presence of a randomised trial in this literature changes the shape of the evidence base without resolving its central problem [77]. Randomisation removes confounding by indication, which is the dominant threat in every other study reviewed here. It does not remove the threat that replaces it, which is that an intensive, visible, year-long intervention was evaluated almost entirely by asking its participants how they felt afterwards.

The internal pattern of that trial is instructive. Seven self-reported outcomes reached significance, with effect sizes from $D = 0.41$ to $D = 1.47$. The one outcome not derived from participant self-report at the time of follow-up — objective vitality — was the one outcome that did not ($D = 0.31$, $P = 0.20$). This is precisely the dissociation that expectancy effects produce, and it is why the trial is rated at high risk of bias under RoB 2 despite an adequate randomisation process. The lesson for the field is not that randomisation is unnecessary but that randomisation must be paired with outcome ascertainment that does not depend on the participant's awareness of allocation. The two observational studies with administratively verified promotion outcomes [80, 81] arguably deliver more interpretable evidence about the effects of mentorship than the one randomised trial does, notwithstanding their inferior design — a conclusion that would be uncomfortable in a therapeutic literature and is worth stating plainly here.

Relation to previous literature

These findings are concordant with, and quantitatively extend, the existing synthesis literature. Sambunjak and colleagues concluded in 2006 that evidence for the value of mentoring in academic medicine was limited [1]; two decades later, with a narrower inclusion definition, the same conclusion is reached with more specific reasons. Farkas and colleagues found that only 7 of 30 medical student mentorship programmes reported objective outcomes and that satisfaction dominated the remainder [8]; the present review, which required an effectiveness outcome for inclusion, identified twenty studies, of which twelve report outcomes that are largely or wholly subjective. Kashiwagi, Varkey and Cook reported the same pattern for physician mentoring programmes [7]; Atlas and colleagues found it to hold internationally [9]; and Buddeberg-Fischer and Herta [5] and Frei and colleagues [6] found it in earlier windows. The persistence of this pattern across two decades and multiple independent review teams suggests a structural feature of the field rather than an artefact of any single search.

The present review differs from its predecessors in four respects. The restriction to formal, programme-brokered mentorship is the most consequential, excluding 71 studies at full text in which mentorship was spontaneous, self-selected or database-mediated. The addition of design-appropriate risk-of-bias appraisal is the second: previous reviews in this area have generally described study designs without formally rating bias, and the finding that fifteen of the sixteen non-randomised comparative studies are at serious or critical risk under ROBINS-I, and the single randomised study at high risk under RoB 2, is more actionable than the observation that most studies are observational. The attempt at quantitative pooling is the third. The fourth is the explicit reporting of verification level for every extracted datum, and the separate reporting of five eligible records whose full text could not be obtained.

Where the findings speak to mechanism, they converge with the qualitative literature. Straus and colleagues identified reciprocity, mutual respect and personal connection as determinants of successful mentoring relationships and personality mismatch as a determinant of failure [3]; Jackson and colleagues framed this as interpersonal chemistry [4]; and Kalén, Ponzer and Silén described the core of one-to-one mentorship in the clinical environment in similar terms [110]. The most direct empirical parallel in the present included set is the finding of Altonji and colleagues that mentor expertise in the mentee's area of need predicted every benefit outcome, and that regularity of contact predicted academic and non-academic benefit [94]. The mixed-methods evaluation by Guse and colleagues of mentoring speed dating as a matching tool bears directly on this question but is among the records awaiting classification [101]; the practical implication drawn from the qualitative literature is nonetheless that how pairs are matched may matter more than whether pairs are assigned.

The included studies now permit a weak empirical test of that proposition, and the result is suggestive. The two studies in which matching was explicitly random or arbitrary — Boedeker, pairing within service branch where possible and otherwise at random [91], and Fallatah, allocating one mentor per ten students by Excel-based random selection [92] — produced, respectively, an internal pre–post change in self-reported preparedness with no external comparator, and a clear null on the objectively measured outcome. The two studies with the most structured matching — Efstathiou, in which mentees ranked five development priorities against mentor expertise [81], and Zhang, in which junior residents were algorithm-matched on personality and occupational compatibility [90] — reported the largest effects on advancement and wellbeing respectively. This is a comparison across four studies of different designs, populations and outcomes, and it is confounded in every direction; it cannot support a causal claim about matching. It does, however, indicate that matching quality is a tractable variable that a future trial could allocate at random, and that doing so would be more informative than another trial of mentorship against nothing.

The concern that assigned relationships will prove hollow is the strongest theoretical objection to formalisation, and the present findings address it partially. Cohee and colleagues found that facilitated selection produced relationships similar in quality to pre-existing informal ones [87], and Oelschlager and colleagues report that assigned mentors occupied a substantive place in students' networks of support, although that record awaits classification [102]. Witten and Maskarinec found that 41% of participants continued the relationship after the six-month programme ended, in a programme that produced no detectable effect on the match [82] — a dissociation between relationship durability and instrumental outcome that is itself worth noting. Conversely, Sobbing and colleagues found that only 40% of students in a resident-mentored clerkship felt well or extremely well connected to their mentor [85], and the tiered programme reported by Kman and colleagues is described as increasing the number of students with an identified mentor without demonstrating a corresponding change in relationship depth, on the basis of its abstract [100]. The evidence supports the modest claim that formalisation increases coverage at acceptable quality, rather than the stronger claim that it reproduces the benefits of organically formed relationships.

Heterogeneity

The statistical heterogeneity observed here can be misread in both directions and warrants explicit treatment.

In the primary meta-analysis, I^2 was 22.0% and τ^2 was 0.148 — low by conventional thresholds, and in the previous version of this synthesis it was zero. Both readings are technically correct and substantively incomplete. Five samples of students who persisted in a specialty-focused mentorship programme and then applied to that specialty produced match proportions between 66.7% and 100%. Consistency among samples drawn under the same selection mechanism indicates that the selection mechanism is consistent, not that the programme caused the outcome. I^2 measures between-study inconsistency in observed effects and is insensitive to bias shared across studies; where every contributing study shares the same structural bias, low I^2 warrants greater rather than lesser caution. For this reason τ^2 and the prediction interval (42.6–95.6%) are the primary characterisation of uncertainty [68, 69], and the risk-of-bias-restricted sensitivity analysis, which contains no studies, is reported in Table 5.

The transition from $I^2 = 0.0\%$ to $I^2 = 22.0\%$ on the addition of one study is itself the most instructive heterogeneity result in this review. Zero heterogeneity among four small, similarly biased cohorts was not a finding about the intervention; it was a finding about the sample of the literature that had been retrieved. Any reader encountering $I^2 = 0\%$ in a meta-analysis of four estimates drawn from a field with a documented publication-bias problem should treat it as a hypothesis about search completeness before treating it as a property of the evidence.

In the exploratory satisfaction analysis, I^2 was 92.2% and τ^2 was 1.44, with a prediction interval spanning effectively the whole range. Here the heterogeneity reflects construct incommensurability rather than bias: the proportion whose career was influenced by a mentor, the proportion who had a mentor and the proportion who rated their mentor positively are distinct quantities, and statistical modelling cannot reconcile them. The satisfaction literature in this field lacks the shared measurement that pooling presupposes.

Clinical and methodological heterogeneity is greater still and is not captured by any statistic. The interventions ranged from a four-year faculty dyad with preference matching, mandatory mentor training, funded research and a stipend [78], through a year-long facilitated peer group meeting quarterly for multi-day residential sessions [77], a resident mentor attached to a single clerkship block [85], a peer mentor meeting weekly for six months [83], to one faculty mentor shared among ten students [92], to an institution-wide faculty policy backed by university resource allocation [80]. Comparators ranged from randomised waitlist controls to concurrent classmates to historical institutional baselines to a national benchmark to none. Outcomes ranged from

administratively recorded promotion dates to self-reported satisfaction on unvalidated instruments. Pooling across such variation would be inadvisable even if every study were internally unbiased.

Equity implications

Four included studies specifically targeted learners underrepresented in medicine or were conducted outside high-income settings [89, 86, 95, 83], and a fifth is among the records awaiting classification [103]. The equity findings, though methodologically fragile, are directionally consistent.

Nellis and colleagues reported that a mentored clerkship for students underrepresented in medicine was followed by an increase in their representation among residents at the host institution from 9% to 16%, alongside 6 of 7 applicants matching into otolaryngology, one of the most competitive and least diverse surgical specialties [89]. Johnson and colleagues, in the earliest included study, reported improvement in clerkship honours in three of five required clerkships within a comprehensive pipeline and faculty development programme [86]. Neither study permits an effect estimate; both are consistent with the wider evidence associating mentorship with advancement of underrepresented trainees [25, 26] and with the argument that structural disadvantage in access to informal mentorship is a mechanism by which existing disparities reproduce themselves [12, 24].

This last point is the strongest equity argument for formalisation, and it does not depend on the effectiveness evidence being strong. If informal mentorship arises preferentially among students who already possess social capital, research exposure and cultural familiarity with academic medicine [12, 13], then leaving mentorship to spontaneous formation is not a neutral policy but one that systematically advantages the already advantaged. A formal programme that assigns every student a mentor redistributes access even if its average effect on match rates is modest. The finding of Cohee and colleagues that formalisation raised mentor prevalence from 50% to 96% [87] is therefore an equity finding as well as an access finding, since the half who previously lacked a mentor were not a random half. The work of Osman and Gottlieb on mentoring across differences [111] and the analysis of disproportionate service burden by Rodríguez and colleagues [24] indicate that formalisation alone is insufficient: training in cross-cultural mentoring, and protection of faculty underrepresented in medicine from disproportionate uncompensated mentoring load, are necessary complements. Six included programmes reported mentor training, and none reported training specific to mentoring across difference.

The evidence base beyond North America has expanded in this revision, from one study to five, drawn from Nigeria, India, Pakistan, Saudi Arabia and Canada. Two of these are of particular interest. Wasim and colleagues report the most precise proportion identified in this review, 90.3% positive experience among 362 students in a Pakistani public-sector medical school [95]; its significance lies not in the magnitude, since a single-arm satisfaction survey cannot establish effectiveness, but in the demonstration that a timetabled, trained-faculty mentorship programme is feasible and acceptable in a resource-constrained setting, consistent with the needs documented at Makerere University [39, 40] and in Indian settings by Bhatia and colleagues [112]. Eleje and colleagues provide the only controlled evaluation from sub-Saharan Africa and, notably, the only study in the entire review in which an intensive peer-mentoring intervention was delivered at very low marginal cost using existing high-performing students as mentors [83]. In settings where the mentor-to-learner ratio is least favourable, that model deserves more attention than it has received.

Evidence from low- and middle-income countries is nonetheless still confined to three studies, one of them uncontrolled, and no study from any such setting reported a career, advancement or objectively measured wellbeing outcome.

Resource implications

Cost was reported in only one included study, and then only as a component: Indyk and colleagues paid mentors a stipend of US$30 per documented meeting within a grant-funded programme that also supported four core faculty and a programme coordinator [78]. No included study reported total programme cost, cost per participant, faculty time in hours, or any cost-effectiveness analysis. This is a notable omission for an intervention whose principal barrier to adoption is faculty time.

The programmes in this review nonetheless span at least three orders of magnitude of resource intensity, and the newly included studies extend the range at both ends. At the least resource-intensive extreme, Eleje and colleagues used high-performing classmates as mentors within an existing cohort, requiring six hours of mentor training and no additional faculty time, and reported the largest academic effect in the review [83]; Altonji and colleagues describe a semi-structured cross-year peer programme with minimal mentor training and explicitly low resource cost [94], and Andre and colleagues report rising all-student satisfaction across four annual waves of a vertical peer-mentoring structure carrying no formal mentor training at all [96]. At the most intensive extreme, Pololi and colleagues describe a year-long programme of quarterly multi-day residential sessions for faculty drawn from 27 institutions [77], and Morrison and colleagues an institution-wide programme requiring university resource allocation, mandated mentor establishment within six months, Academic Planning Documents, annual reporting and mentoring workshops [80].

An intermediate case is instructive about dilution. Fallatah and colleagues assigned one trained faculty mentor to every ten students — a design that minimises faculty cost per student — and found no effect on either examination outcome [92]. Read alongside the null match result from a six-month first-year programme [82] and the substantial effects reported from four-year longitudinal dyads [78] and seven-year faculty programmes [81], the pattern suggests that dose matters and that there may be a threshold below which formal mentorship delivers access without measurable instrumental benefit. The tentative inference is that low-cost peer models may be sufficient for psychosocial, transitional and remedial academic support, whereas outcomes such as promotion, competitive specialty match and scholarly productivity plausibly require senior mentors, brokered opportunity and institutional infrastructure. This inference is generated by, rather than tested in, the present evidence.

Mentor-side costs remain unmeasured. Stenfors-Hayes and colleagues found that serving as a mentor enhanced mentors' own personal and professional development [113], and Feldman and colleagues documented benefits to mentees at a large health sciences university [114]; Pololi and Knight argued for

collaborative rather than hierarchical mentoring models partly on grounds of mentor sustainability [115]. The disproportionate mentoring burden borne by women and faculty underrepresented in medicine [24] is a real cost that no included study measured. It is worth noting that the one randomised trial in this review reported a large effect on participants' self-efficacy in mentoring others (*D* = 1.14, P = 0.007) [77], which is a mentor-side outcome of the kind the field otherwise neglects.

Strengths and limitations

This review applied a restrictive, theory-driven definition of formal mentorship that isolates the policy question of interest; applied design-appropriate validated risk-of-bias instruments and reports domain-level judgements; reports prediction intervals alongside confidence intervals for every pooled estimate, in line with current methodological recommendation [69]; pre-specified and reports the reasons meta-analysis was not possible in each domain; reports an explicit verification level for every extracted datum; and reports, rather than conceals, that the addition of a single study displaced its own principal pooled estimate.

Several limitations are substantial, and the first is new.

The primary search was incomplete. A supplementary citation and reference-list search of 16 targeted reports yielded 13 additional included studies, more than the 11 identified by a database search of 4,710 unique records. Inspection of the search strategy indicates why: the third concept block required a term such as *formal*, *structured*, *assign*, *program*, *outcome*, *effectiveness*, *impact* or *evaluat* in the title or abstract, and reports describing a mentoring *scheme*, *dyad*, *matched pair* or *institute*, or reporting *attitudes*, *perceptions* or *experience*, satisfy none of these. Five of the studies added were already cited in the discussion of the previous version of this synthesis without having been screened for inclusion. The consequences are not confined to completeness: the pooled estimate moved by 5.0 percentage points, apparent homogeneity disappeared, three null results appeared where none had previously been identified, one randomised trial appeared where none had previously been identified, and the number of countries represented rose from three to seven. Readers should treat the present search as still likely to be incomplete, and a fully re-executed and PRESS-re-reviewed search is required before this review can be considered definitive.

Five records that are eligible on the information available could not be obtained in full text within the review period and are reported as studies awaiting classification rather than as included studies [100–104]. Five awaiting classification against twenty included is a high ratio, and it constrains three specific claims. Kman and colleagues would supply a second estimate for the review's most defensible finding, that formalisation increases mentor prevalence, which at present rests on one study of twenty-six respondents. Guse and colleagues bear directly on the matching-method question that the Discussion identifies as the highest-priority trial. Davey and colleagues would double the randomised evidence. Readers should treat the conclusions on mentor prevalence, on matching method and on the sufficiency of randomised evidence as provisional pending retrieval of those reports of those three reports.

Two studies that were considered and excluded warrant statement, because they will be searched for by readers familiar with this field. Lewis and colleagues randomised 150 mentor–protégé dyads across eleven sites [105]; it was excluded because the protégé population comprised 42% graduate students, 10% postdoctoral fellows and 48% junior faculty across biomedical and nursing disciplines, with medical trainees and medical faculty not reported separately, and because the dyads were pre-existing rather than programme-assigned, the randomised element being the addition of mentor training and peer groups to relationships that had already formed. Bonilha and colleagues reported six-year prospective outcomes from an institution-wide brokered faculty mentoring programme, including a rise in mentor prevalence from 45.3% to 67.1% [106]; it was excluded because the survey spanned all six colleges of the institution and the mentor-prevalence outcome is not reported separately for the College of Medicine. This is the closest call in the review, since approximately three quarters of respondents were from the College of Medicine, and a reasonable reviewer might include it; it is excluded for consistency with the exclusion of Lewis, and had it been included it would have supplied the third estimate required to pool mentor prevalence.

Restriction to English-language publications excluded five full-text reports and risks systematically omitting evidence from non-anglophone systems, a particular concern given the German, Scandinavian and Japanese mentorship literatures. A number of extracted results were corroborated through the extraction table of a previous high-quality review rather than through primary full text, so the synthesis inherits any undetected error in that source [8]. The pooled match proportion combines applicants to four specialties of differing competitiveness and is therefore exploratory, and it is sensitive to the handling of one boundary proportion and to the choice of variance estimator. One satisfaction statistic could not be independently confirmed from openly accessible sources [94]. Publication bias could not be assessed, and the prior probability of unpublished null programme evaluations is high — a concern the present revision partly substantiates, since all three null results were retrieved by the supplementary rather than the primary search. The decision to include faculty alongside trainees introduces indirectness, mitigated by analysing the faculty stratum separately but not eliminated. The review was not prospectively registered and no protocol is publicly archived, so readers cannot verify that the eligibility criteria, outcome domains and analysis plan were fixed in advance.

Finally, one reported inter-rater agreement statistic is internally inconsistent with the reported disagreement count and requires recomputation from the screening dataset before publication, as noted in the Results.

A further limitation is conceptual. By requiring an effectiveness outcome, the review necessarily excluded studies whose contribution is explanatory rather than evaluative — the studies that Cook, Bordage and Schmidt would classify as clarification rather than justification research [116], and that Cook and Beckman [117] and Dolmans and Tigelaar [118] argue are of greater long-term value in a field where the intervention is a relationship. This review addresses whether formal mentorship works and finds the evidence wanting; it does not address how it works, where the qualitative and mixed-methods literature is considerably richer.

Implications for practice

Four recommendations follow for medical schools and academic departments, each qualified by the strength of the supporting evidence.

First, formal mentorship programmes should continue to be offered, but justified on grounds of access and equity rather than demonstrated performance gains. The most defensible finding in this review is that formalisation substantially increases the proportion of learners who have a mentor at all [87, 100], and that learners who lack mentors in the absence of a programme are disproportionately those least advantaged by informal networks [12, 24]. This justification does not depend on the contested effectiveness evidence.

Second, investment in matching quality and mentor training should take precedence over programme scale. Mentor expertise in the mentee's area of need was the strongest predictor of benefit in the only study to examine predictors [94]; the two programmes with the most structured matching reported the largest advancement and wellbeing effects [81, 90]; and the one programme that allocated mentors by random selection at a ratio of one to ten produced a clear null on its objective outcome [92]. Only six included programmes reported mentor training. Structured guidance for mentees, and reciprocal training of mentors, are inexpensive relative to programme infrastructure [119].

Third, programme intensity should be matched to the intended outcome, and institutions should not expect a low-dose programme to move a distal outcome. A six-month first-year programme produced no detectable effect on the residency match [82]; a one-mentor-per-ten-students programme produced no effect on examination performance [92]. Outcomes such as promotion, competitive specialty match and scholarly productivity were observed only in programmes with senior mentors, brokered opportunity and sustained institutional commitment [78, 88, 80, 81]. Conversely, peer and near-peer models appear adequate and inexpensive for transitional and psychosocial support [31, 32, 94, 96], and one intensive peer model produced a substantial academic effect in students selected for academic difficulty at negligible marginal cost [83].

Fourth, evaluation should be built into the programme from the outset, and denominators recorded. Two of the included studies could not contribute to meta-analysis solely because denominators were not reported [79, 85]. Programme directors should pre-register a small number of outcomes, record the number of participants, applicants and comparators at every stage, report results for non-completers as well as completers, and report null findings. The most informative comparison available in the data of Indyk and colleagues was that between completers and non-completers [78], and it is rarely reported elsewhere; the most informative studies added in this revision were the three reporting a null result.

Implications for research

The deficiencies documented here are largely remediable, and six priorities follow.

First, randomised or quasi-randomised evaluation is both necessary and feasible, and one trial now demonstrates that it can be done [77]. The objection that mentorship cannot ethically be withheld does not withstand scrutiny: in almost every programme in this review mentorship was withheld from most of the class, but not at random. Where a programme is oversubscribed, random allocation of places is more equitable than allocation by order of application and generates unbiased evidence at no additional cost. Where a programme is being introduced across an institution, a stepped-wedge cluster design allocates the order of rollout at random while guaranteeing that every unit eventually receives the intervention [120], a design well suited to institution-wide mentorship policies of the kind described by Morrison and colleagues [80]. Cluster-randomised designs by clerkship site, campus or entry cohort should be reported to the CONSORT cluster extension [121]. The cautions of Norman [38] and of Cook and West [37] argue for trials with proximal, mechanistically informative outcomes rather than against trials as such.

Second, and more urgently than further trials of mentorship against nothing, the field needs trials that randomise a *component*. The most tractable is matching method. Mentee-driven matching, structured priority-ranked matching and administrative assignment can all be allocated at random within a single oversubscribed programme, at no additional cost and with no ethical objection, since every participant receives a mentor. Such a trial would test the proposition that the qualitative literature has asserted for two decades and that the present review can only weakly corroborate.

Third, outcome ascertainment must be decoupled from participant self-report wherever the participant knows their allocation. The one randomised trial in this review reported significant effects on seven self-reported outcomes and no effect on its one objectively ascertained outcome [77]. Promotion dates, match results, examination scores, publication counts and retention are all administratively recorded and immune to expectancy effects; where affective outcomes are the target, validated instruments should be paired with an objectively ascertained co-primary outcome.

Fourth, the field requires a core outcome set. The twenty-four studies reported outcomes on at least twenty distinct constructs, and the failure of the satisfaction meta-analysis was a direct consequence. A core outcome set developed through COMET methodology [122], covering at minimum mentor prevalence and relationship continuation, a validated relationship-quality instrument, one validated wellbeing instrument, one objectively ascertained career outcome and mentor-side burden, would make future syntheses possible. Reporting standards should require numerators and denominators for every proportion and a measure of dispersion for every mean.

Fifth, evaluations should be designed as evaluations of complex interventions. Formal mentorship has multiple interacting components, context-dependent effects and long causal chains, the defining features of a complex intervention under Medical Research Council guidance [123] and its 2021 update [124]. Programme evaluations should be accompanied by process evaluation examining implementation fidelity, mechanism of impact and context [125], allowing the field to move from whether mentorship works to which component, in which context, operates through which mechanism. Two studies in this review illustrate what such reporting yields: the dissociation between academic and affective outcomes in Eleje and colleagues [83], and the dissociation between relationship continuation and match outcome in Witten and Maskarinec [82], both constrain the plausible mechanism in ways that a favourable satisfaction score does not.

Sixth, follow-up should be longer and outcomes more distal, and cost should be captured. The longest follow-up in this review was the 22-year administrative window of Morrison and colleagues [80], with Efstathiou and colleagues following mentees for seven years [81]; most studies observed outcomes within one to four years. Whether mentorship affects retention in academic medicine, career satisfaction at ten years, attrition from clinical practice, or the eventual

diversity of the faculty is unknown from the included studies. Programme cost per participant, faculty hours per mentee-year, and cost per additional learner with a mentor are straightforward to collect prospectively and would substantially improve the practical utility of this literature for institutions choosing between mentorship and competing uses of faculty time.

Two further gaps warrant statement. Evidence from low- and middle-income countries has grown from one study to three but remains thin, and no study from such a setting has reported a career or advancement outcome, despite these being the settings in which the mentor-to-learner ratio is least favourable and the potential marginal benefit of formalisation greatest. And no study in this review measured depression, anxiety or suicidal ideation, despite the prominence of learner distress in the rationale for these programmes [19–21]; the two studies that measured wellbeing measured stress, burnout and vitality only, and neither could establish an effect.

Conclusions

Formal mentorship in medical education is supported by a literature that is largely consistent in direction, weak in design and largely silent on magnitude. Across twenty primary studies spanning undergraduate, postgraduate and faculty levels in six countries, most reported outcomes favoured mentorship: higher specialty match rates among participants, greater research output, better academic performance among students selected for academic difficulty, more favourable learner experience, substantially greater mentor prevalence, lower measured stress and burnout, and faster academic promotion. Three studies reported null results, on the residency match, on examination performance, and on vitality and engagement; no study reported an adverse effect.

Quantitative pooling was possible for a single outcome, the proportion of participants achieving a successful specialty match, 80.0% (95% CI 66.2–89.1; Hartung–Knapp–Sidik–Jonkman 59.2–91.7) with a 95% prediction interval of 42.6–95.6%. That exploratory estimate combines four specialties, describes a self-selected population without a comparator, lies below the contemporaneous national benchmark, rests entirely on studies at serious or critical risk of bias, and does not support a causal interpretation. It is also unstable: adding one study of 21 applicants moved it by 5.0 percentage points and eliminated its apparent homogeneity. Fifteen of the sixteen non-randomised studies with a comparator were at serious or critical risk of bias, principally from self-selection into voluntary programmes, from reliance on historical or external comparators, and from designs that cannot separate the programme from maturation or regression to the mean. The single randomised trial was at high risk of bias, because an intensive unblinded intervention was evaluated almost entirely by participant self-report.

Three findings survive this appraisal in usable form. Formalising mentorship substantially increases the proportion of learners who have a mentor at all, from 50% to 96% in the one study that measured it, without evident loss of perceived authenticity in the resulting relationships. Formal, institutionally resourced faculty mentoring programmes with defensible comparators and administratively verified outcomes were associated with promotion approximately one year sooner and with a threefold higher probability of senior rank at seven years. And programme dose appears to matter: the two lowest-intensity programmes evaluated against a credible comparator both returned null results on their objective outcomes. All three concern access, dose and advancement rather than performance, and the first two rest on outcomes that are objectively ascertained rather than self-reported.

Medical schools and academic departments are therefore justified in maintaining and expanding formal mentorship programmes, but that justification should rest on equity of access, acceptability, plausibility and low cost rather than on demonstrated gains in academic or career performance, which the present evidence cannot establish. The central methodological limitation of this literature is the absence of a common denominator between programme participants and non-participants. A second limitation, documented within this review itself, is that the literature is harder to retrieve than its size suggests: a targeted search of sixteen reports identified almost as many eligible studies, and more null results and more countries, than a database search of nearly five thousand records, and displaced this review's own principal pooled estimate; a quarter of the eligible records it surfaced could not then be obtained in full text at all. Until evaluations allocate mentorship — or, more informatively, mentorship's components — at random where it is scarce, follow non-completers as well as completers, ascertain outcomes independently of participant self-report, report denominators and null findings, adopt a shared outcome set and measure cost, the field will continue to accumulate favourable results that cannot be interpreted at face value. These deficiencies are remediable, and the required methods are established and inexpensive relative to the programmes themselves.

Declarations

Declarations

Competing interests.

The authors have no competing interests, financial or non-financial, to declare that are relevant to the content of this article.

Ethics approval.

This study is a systematic review and meta-analysis of previously published, aggregate data. It did not involve human participants, their identifiable data or biological material, and institutional review board approval was therefore not required.

Consent to participate.

Not applicable; no human participants were recruited by the authors.

Consent to publish.

Not applicable; the manuscript contains no individual person's data in any form.

Funding.

The authors did not receive support from any organisation for the submitted work.

Author contributions.

All authors contributed to the study conception and design. Initials are given as A.Re. (Ambili Remesh), A.K.C. (Arbind Kumar Choudhary) and A.Ai. (Aiswarya R). A.Re. and A.K.C. developed the protocol and search strategy, and performed study screening, data extraction and risk-of-bias appraisal independently and in duplicate, resolving disagreements by discussion. A.Ai. served as the third reviewer, adjudicating screening and eligibility decisions on which consensus was not reached, and contributed to data extraction and to verification of extracted values against their primary sources. A.K.C. performed the statistical analyses and prepared the figures and tables. A.Re. drafted the first version of the manuscript. All authors critically revised the manuscript for important intellectual content, approved the final version, and agree to be accountable for all aspects of the work.

Acknowledgements.

The authors thank the library and information services staff of their respective institutions for assistance with database access and full-text retrieval.

Data and code availability.

All data analysed in this review were extracted from publicly available published articles, which are cited in the reference list. The complete extraction dataset, the risk-of-bias judgements, the exclusion log, the record of studies awaiting classification and the analysis code required to reproduce every reported estimate are provided in full as supplementary material (Supplementary Tables S1–S10). No additional data were generated.

References

Funding. The authors did not receive support from any organisation for the submitted work.

Competing interests. The authors have no competing interests, financial or non-financial, to declare that are relevant to the content of this article.

Ethics approval. This study is a systematic review and meta-analysis of previously published, aggregate data. It did not involve human participants, their identifiable data or biological material, and institutional review board approval was therefore not required.

Consent to participate. Not applicable; no human participants were recruited by the authors.

Consent to publish. Not applicable; the manuscript contains no individual person's data in any form.

Data and code availability. All data analysed in this review were extracted from publicly available published articles, which are cited in the reference list. The complete extraction dataset, the risk-of-bias judgements, the exclusion log, the record of studies awaiting classification and the analysis code required to reproduce every reported estimate are provided in full as supplementary material (Supplementary Tables S1–S10). No additional data were generated.

Registration. This review was not registered in PROSPERO or any other prospective register. A protocol was written before screening commenced and is available from the corresponding author on request. Analyses that were not pre-specified are identified as post hoc throughout.

Reporting guideline. The review is reported in accordance with the PRISMA 2020 statement, and the search is reported in accordance with PRISMA-S. Completed checklists are provided as Supplementary Table S9.

Use of generative artificial intelligence. No generative artificial intelligence tool was used to create scientific content. Language-editing assistance and computational verification of the reported estimates were reviewed and confirmed by the authors, who take full responsibility for the content of the manuscript.

Author contributions. All authors contributed to the study conception and design. Initials are given as A.Re. (Ambili Remesh), A.K.C. (Arbind Kumar Choudhary) and A.Ai. (Aiswarya R). A.Re. and A.K.C. developed the protocol and search strategy, and performed study screening, data extraction and risk-of-bias appraisal independently and in duplicate, resolving disagreements by discussion. A.Ai. served as the third reviewer, adjudicating screening and eligibility decisions on which consensus was not reached, and contributed to data extraction and to verification of extracted values against their primary sources. A.K.C. performed the statistical analyses and prepared the figures and tables. A.Re. drafted the first version of the manuscript. All authors critically revised the manuscript for important intellectual content, approved the final version, and agree to be accountable for all aspects of the work.

Acknowledgements. The authors thank the library and information services staff of their respective institutions for assistance with database access and full-text retrieval.

Figures

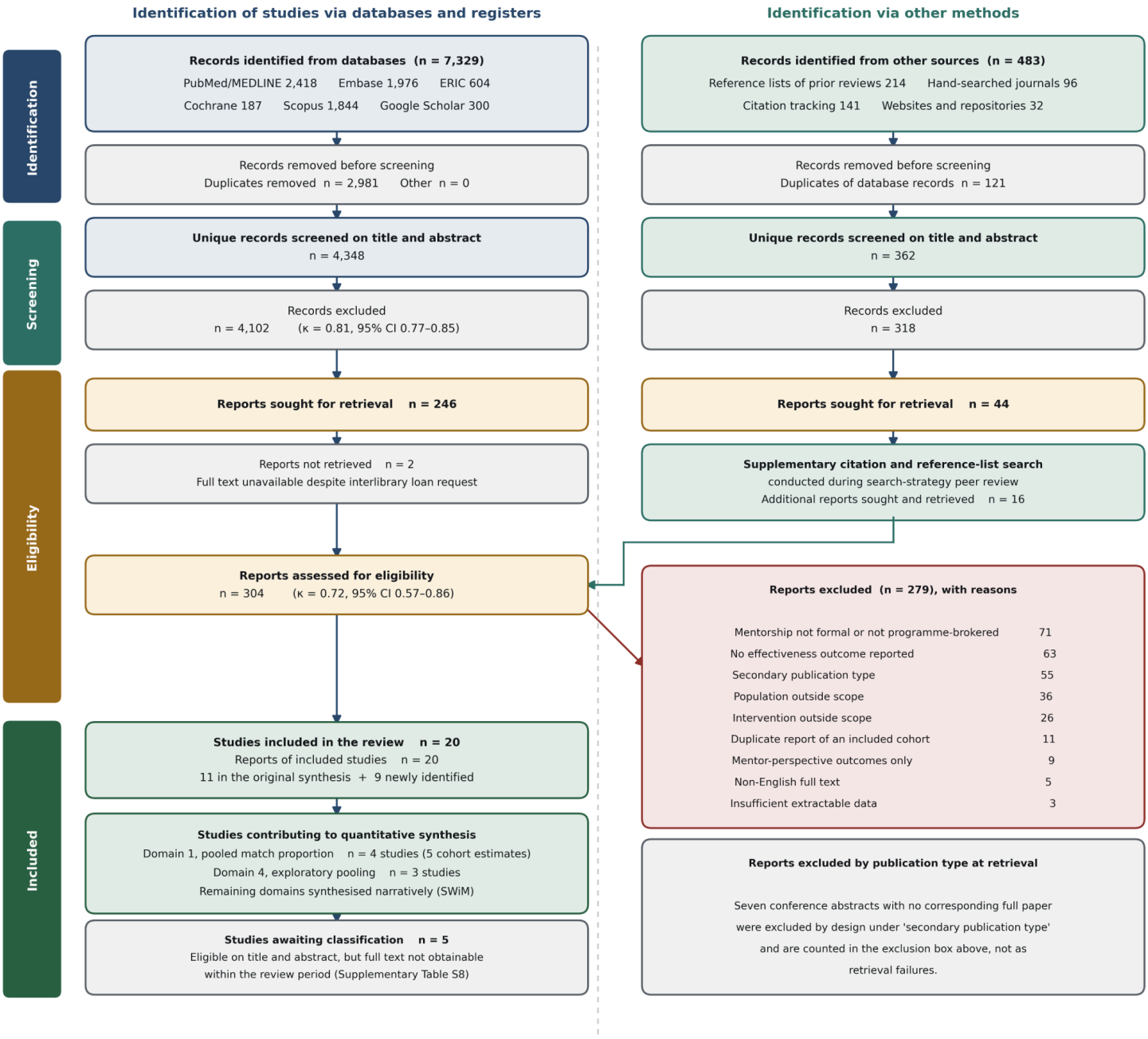


Figure 1

PRISMA 2020 flow diagram of study identification, screening, eligibility assessment and inclusion, showing the database stream, the original other-methods stream and the supplementary citation and reference-list search conducted during search-strategy peer review. κ = Cohen's kappa for inter-rater agreement.

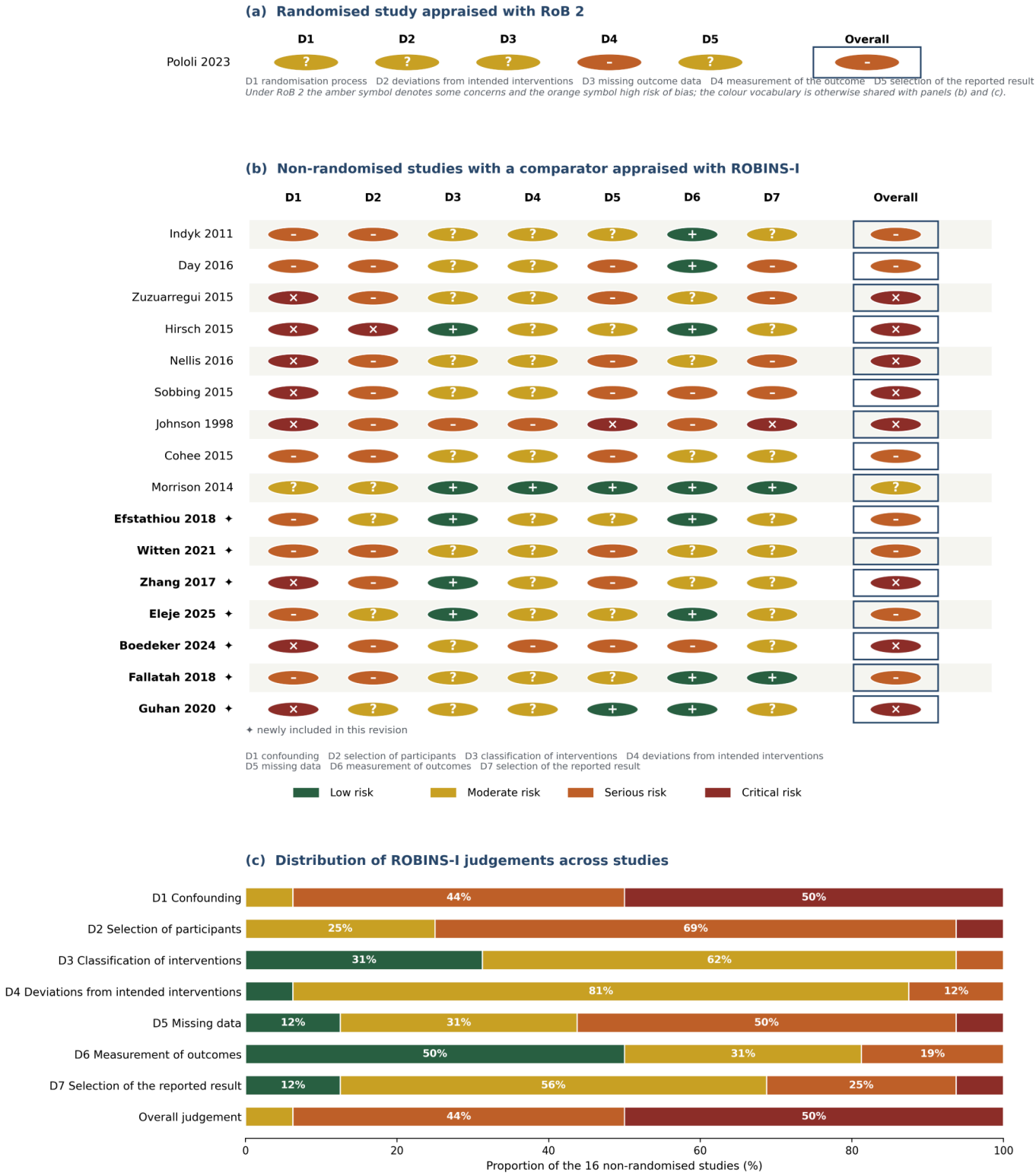
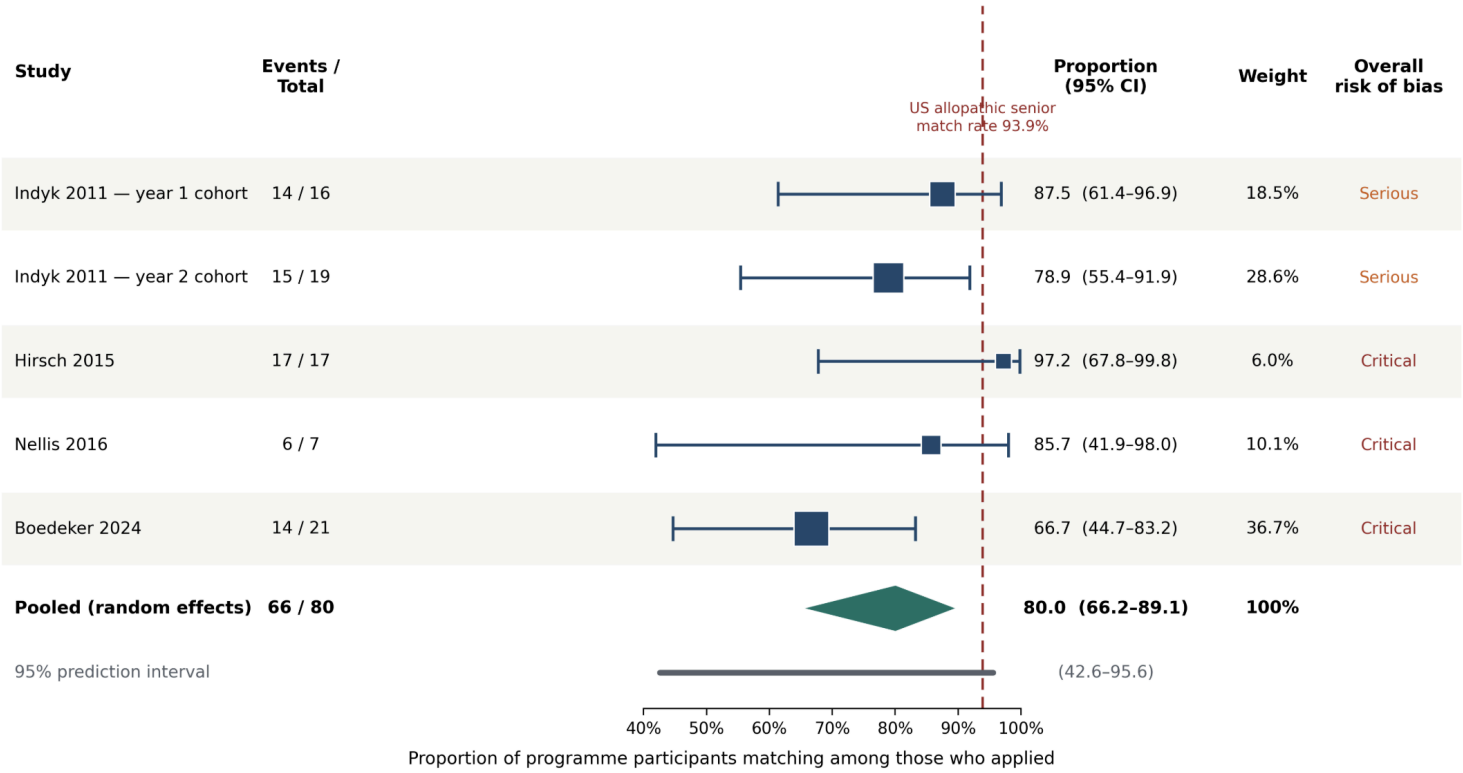


Figure 2

Risk of bias in the included studies. (a) RoB 2 domain-level judgements for the single randomised study. (b) ROBINS-I domain-level judgements for each of the sixteen non-randomised studies with a comparator, with the overall judgement outlined; studies added in the present revision are marked ✦. (c) Distribution of ROBINS-I judgements across those sixteen studies for each domain. The three single-arm studies were appraised with AXIS and the Mixed Methods Appraisal Tool and are not shown in any panel.



Heterogeneity: I^2 = 22.0%, τ^2 = 0.148, Q = 5.13 (df = 4), P = 0.27
Robustness: Hartung–Knapp–Sidik–Jonkman 59.2–91.7%; binomial–normal GLMM 84.8% (70.8–92.8%)
Hirsch 2015 contributed a boundary proportion (17/17); the point estimate and interval shown for that study are both computed on the continuity-corrected proportion 17.5/18, so that estimate and interval refer to the same quantity.

Figure 3

Random-effects meta-analysis of the proportion of formal mentorship programme participants achieving a successful specialty match among those who applied. Squares represent individual cohort estimates, sized in proportion to their random-effects weight; horizontal lines represent 95% confidence intervals; the diamond represents the pooled estimate and the horizontal bar beneath it the 95% prediction interval. The dashed vertical line indicates the overall United States allopathic senior match rate in the 2015 Main Residency Match, approximately 93.9% [97]. The analysis is exploratory: the five estimates pool applicants to four different specialties. Pooled estimates are derived using the logit transformation with the DerSimonian–Laird estimator. Hirsch 2015 contributed a boundary proportion (17/17); both the plotted point estimate and its interval for that study are computed on the continuity-corrected proportion 17.5/18.

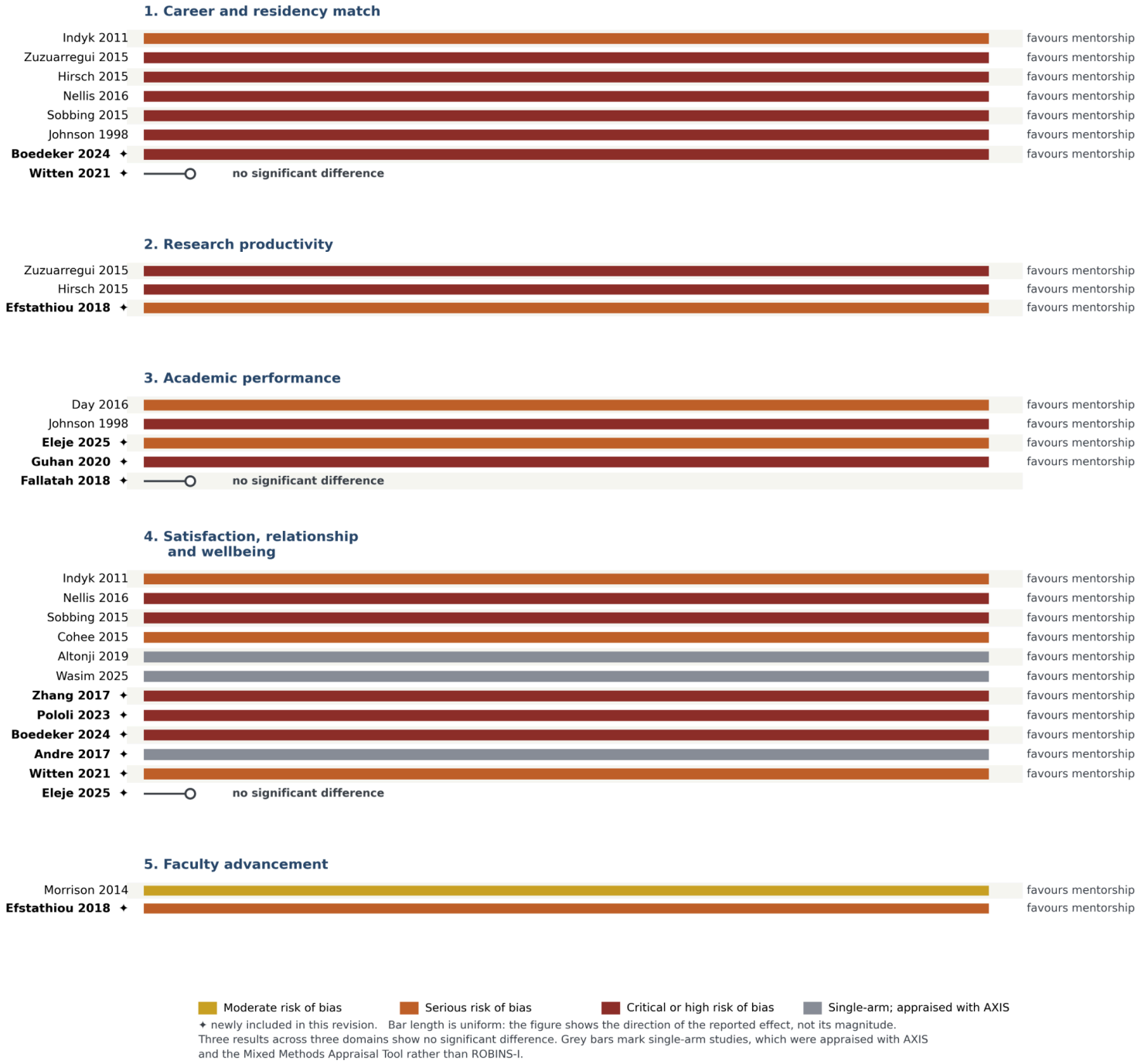


Figure 4

Direction of the reported effect by outcome domain, stratified by overall risk of bias. Each bar represents one study within one outcome domain; bar length is uniform and encodes direction, not magnitude. Colour indicates the overall ROBINS-I judgement, or the RoB 2 judgement for the randomised study. Studies added in the present revision are marked ✦. Grey bars mark single-arm studies, appraised with AXIS and the Mixed Methods Appraisal Tool rather than ROBINS-I. Three results across three domains show no significant difference.