OPEN ACCESS

Evidence-based medical leadership development: a systematic review

Oscar Lyons [1], Robynne George,[2] Joao R Galante,[3,4] Alexander Mafi,[5] Thomas Fordwoh,[5] Jan Frich [6], Jaason Matthew Geerts [7,8]

► Additional material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/leader-2020-000360).

[1]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK
[2]Royal United Hospital Bath NHS Trust, Bath, UK
[3]Department of Medical Education, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[4]Cardiology Department, Buckinghamshire Healthcare NHS Trust, Amersham, UK
[5]University of Oxford Medical School, University of Oxford, Oxford, UK
[6]Department of Health Management and Health Economics, University of Oslo, Oslo, Norway
[7]Research and Leadership Development, Canadian College of Health Leaders, Ottawa, Ontario, Canada
[8]The Business School (formerly Cass), University of London, London, UK

Correspondence to
Dr Oscar Lyons, Nuffield Department of Surgical Sciences, University of Oxford, Oxford OX3 9DU, UK;
oscar.lyons@nds.ox.ac.uk

Received 6 August 2020
Revised 16 September 2020
Accepted 28 September 2020
Published Online First 16 November 2020

Check for updates

To cite: Lyons O, George R, Galante JR, *et al*. *BMJ Leader* 2021;**5**:206–213.

ABSTRACT

Health systems invest significant resources in leadership development for physicians and other health professionals. Competent leadership is considered vital for maintaining and improving quality and patient safety. We carried out this systematic review to synthesise new empirical evidence regarding medical leadership development programme factors which are associated with outcomes at the clinical and organisational levels. Using Ovid MEDLINE, we conducted a database search using both free text and Medical Subject Headings. We then conducted an extensive hand-search of references and of citations in known healthcare leadership development reviews. We applied the Medical Education Research Study Quality Indicator (MERSQI) and the Joanna Briggs Institute (JBI) Critical Appraisal Tool to determine study reliability, and synthesised results using a meta-aggregation approach. 117 studies were included in this systematic review. 28 studies met criteria for higher reliability studies. The median critical appraisal score according to the MERSQI was 8.5/18 and the median critical appraisal score according to the JBI was 3/10. There were recurring causes of low study quality scores related to study design, data analysis and reporting. There was considerable heterogeneity in intervention design and evaluation design. Programmes with internal or mixed faculty were significantly more likely to report organisational outcomes than programmes with external faculty only (p=0.049). Project work and mentoring increased the likelihood of organisational outcomes. No leadership development content area was particularly associated with organisational outcomes. In leadership development programmes in healthcare, external faculty should be used to supplement in-house faculty and not be a replacement for in-house expertise. To facilitate organisational outcomes, interventions should include project work and mentoring. Educational methods appear to be more important for organisational outcomes than specific curriculum content. Improving evaluation design will allow educators and evaluators to more effectively understand factors which are reliably associated with organisational outcomes of leadership development.

INTRODUCTION

Health systems invest significant resources in leadership development for physicians and other health professionals.[1] Competent leadership is considered vital for team effectiveness, for clinical and financial performance and for maintaining and improving quality and patient safety.[1–5] Clinical leadership development involves activities to promote leadership competencies among clinicians, while medical leadership development refers to activities centred on doctors.

Research suggests that medical leadership development can improve outcomes at individual, organisational and clinical levels.[6–11] Evidence backing medical leadership development activities has, however, been variable in quality.[1 7–10 12–15] There has been a particular lack of research and evaluation that goes beyond individual learner feedback and subjective outcomes.[6–9] One systematic review of 45 studies evaluating leadership development interventions for doctors found that effective interventions were characterised by the use of multiple learning methods, including seminars and group work, alongside action learning projects in multidisciplinary teams.[8] These findings were echoed in a recent study by Geerts *et al*,[9] who emphasised that plans need to be in place for transferring learning from the intervention into the working environment.

We undertook this systematic review to synthesise recent empirical evidence regarding medical leadership development programme factors associated with outcomes at the clinical and organisational levels. We specifically investigated links between aspects of programme design, delivery and evaluation and improved outcomes. Given the variable quality of studies highlighted in previous reviews,[7–9] we applied two validated critical appraisal instruments[16 17] to isolate higher reliability findings. This review is the first to apply both instruments in order to identify and synthesise the highest quality empirical evidence in medical leadership development.

METHODS

The design of this review was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses[18] and the Best Evidence in Medical Education (BEME) guide for systematic reviews.[19] Our methods were based on the review conducted by Frich *et al*,[8] with methodological changes drawn from other reviews.[7 9 10 14 15 20] Following the BEME recommendations for systematic reviews,[19] we hand-searched references and citations of known reviews extensively to supplement our database search. In line with recommendations from Geerts *et al*[9] and Rosenman *et al*,[7] we assessed study quality using the Medical Education Research Study Quality Indicator (MERSQI), which is designed to measure the methodological quality of quantitative medical education research studies.[16] We added

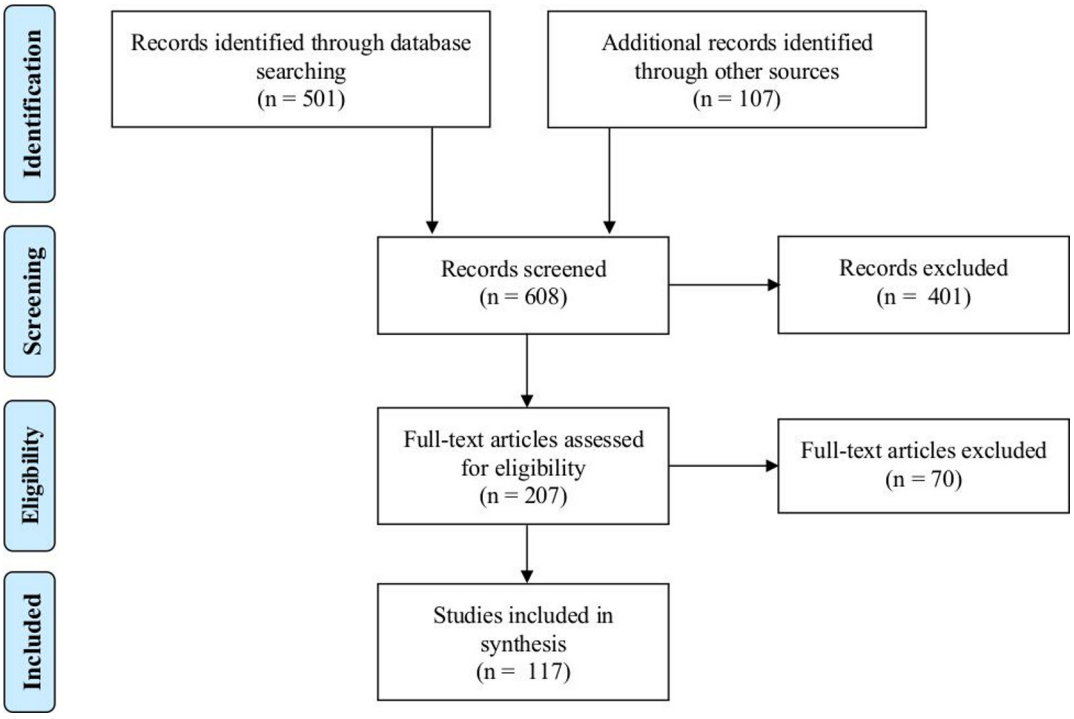

Figure 1 PRISMA diagram. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

the Joanna Briggs Institute (JBI) Critical Appraisal Checklist,[17] which is designed for meta-aggregation of qualitative research and is well-established in healthcare research.[21]

Search strategy

We began this review by re-examining the data set identified in the review of leadership development for physicians by Frich *et al*.[8] With assistance from a specialist librarian at the University of Oxford, we then based our search strategy on Frich *et al*'s review.[8] Using the Ovid MEDLINE database, we conducted a search using both free text and Medical Subject Headings. The full search terms are listed in the online supplemental material. This search identified 501 unique publications. We then conducted an extensive hand-search of references and of citations in known healthcare leadership development reviews using Web of Science and Google Scholar. This identified an additional 107 studies for possible inclusion, for a total of 608 records for screening (figure 1).

Inclusion criteria

We included any peer-reviewed study published in English between January 2000 and January 2020 which:

1. Describes a leadership development intervention (programme, workshop, course and so on).
2. Includes physicians as learners (defined here as any practising doctor post-qualification).
3. Evaluates the leadership development intervention.

Qualitative, quantitative and mixed evaluations were included. We excluded studies where leadership development was a minor focus or where the proportion of physicians was lower than 10% of intervention participants.

Screening process

Two members of the review team (OL and TF) independently screened all study titles and abstracts for eligibility. Articles that were approved by either reviewer progressed to full-text review. Two members of the review team independently reviewed for inclusion the full text of all 207 articles that passed the title and abstract screen (TF and RG reviewed half each, OL reviewed all). Where there was disagreement about inclusion, all three reviewers (OL, TF, RG) reached consensus by discussion, with the third reviewer (TF or RG) arbitrating where required.

Data abstraction

After screening and reviewing for eligibility, 117 unique studies were included for abstraction and analysis. Data were abstracted and coded for educational setting, methods, content, evaluation methods and outcomes. Outcome data were categorised according to an adapted version of Kirkpatrick's Framework for evaluation of training programmes (see table 1).[19 22] One reviewer abstracted and coded all 117 included studies (OL). The second reviewers (RG/JRG/AM/TF) each abstracted and coded at least five studies in full to ensure consistency between reviewers. Data abstraction and coding for all 117 studies was then cross-checked by the second reviewers. Any differences were resolved by consensus, with a third reviewer arbitrating where required. Where possible, statistical tests performed in studies were replicated and checked for accuracy.

Study quality appraisal

Previous reviews have shown marked variation in the quality of studies of medical leadership development.[7 9 10 14 15 20] To isolate the most reliable evidence linking medical leadership programmes to improved outcomes, two researchers independently critically appraised each included study using the MERSQI and JBI Instruments.[16 21] Differences in MERSQI and JBI quality score were resolved by consensus, and a third researcher arbitrated where needed.

The MERSQI was applied to all 117 studies. The MERSQI is a validated appraisal tool consisting of 10 items in six domains which relate to design, sampling, type of data collected, validity of evaluation methods, analysis and outcomes.[16] Each domain is

Table 1 Kirkpatrick's Framework for evaluation of training programmes, with adaptations from Frich *et al*[8]

Kirkpatrick level	Description
Level 1 Reaction	Participants' satisfaction with the learning experience, its organisation, presentation, content, teaching methods and quality of instruction
Level 2A Change in attitudes	Changes in the attitudes or perceptions among participant groups towards leadership, management and/or administration
Level 2B Change in knowledge or skills	For knowledge, this relates to the acquisition of concepts, procedures and principles; for skills, this relates to the acquisition of thinking/problem-solving, psychomotor and social skills
Level 3A Behavioural change (self-reported)	Transfer of learning to the workplace and changes to professional practice, as noted by participants themselves
Level 3B Behavioural change (observed)	Transfer of learning to the workplace and changes to professional practice, as noted by a third party or by promotions
Level 4a Results (self-reported)	Organisational outcomes perceived by respondents and group effectiveness perceived by subordinates
Level 4b Results (observed)	Tangible organisational outcomes, such as reduced costs, improved quality and safety, impact of projects

scored to a maximum of 3, for a total score of 5–18. In line with Geerts *et al*,[9] studies with scores of 12 or higher were categorised as higher reliability studies (see the Data analysis section).

The JBI Checklist for Qualitative Studies was also applied where a study used mixed methods (k=53) or qualitative methods (k=10). Fundamental differences in study design, sampling, evaluation instruments and analysis preclude summative comparison of mixed-methods or qualitative studies to quantitative studies using the MERSQI.[16 21 23 24] The JBI Checklist is considered the most appropriate qualitative critical appraisal tool for use in pragmatic meta-aggregation of qualitative research.[24] It includes 10 items which regard the study's research questions, methods, analysis and reporting, for a total score of 0–10. Following recommendations from the JBI Reviewers' Manual,[17] a cut-off score for higher reliability studies was predetermined at 6/10. This score was chosen as studies obtaining six or more points included most key elements of high-quality design.

Data analysis

MERSQI and JBI Scores were used to establish which studies presented more reliable evidence of outcomes. Summary statistics were calculated for all 117 studies. In line with Geerts *et al*,[9] studies with a final MERSQI Score of 12/18 or higher were also analysed separately to isolate the most reliable evidence, as were qualitative and mixed-methods studies which achieved the pre-determined JBI Score of 6/10 or higher. As there was substantial methodological heterogeneity, study characteristics and outcomes were synthesised using a meta-aggregation approach.[25] All study quality appraisal scores are presented in the Online supplemental table 1, and full data extraction tables are available on request.

RESULTS

Study reliability (MERSQI and JBI)

Twenty-eight of 117 studies (25%) were categorised as higher reliability. Two studies were categorised as higher reliability by both the MERSQI and the JBI tool,[26 27] 14 studies (12%) by the MERSQI only and 12 studies (10%) by the JBI tool only. The median critical appraisal score according to the MERSQI was 8.5 (range 5–16 from possible range of 5–18) and the median critical appraisal score according to the JBI was 3 (range 0–9 from possible range of 0–10). Online supplemental table 1 includes the MERSQI and JBI Scores for all included studies.

Study design showed considerable room for improvement, as shown in online supplemental tables 2 and 3. Nearly half the of studies (46%) relied on post-programme evaluations only, and 92% did not include a control group. Of the nine studies that did include control groups, most had substantial methodological flaws in their selection of control groups. One common method for control group recruitment was to use unsuccessful course applicants.[28–30] In terms of evaluation design, the median evaluation instrument score was 0 (range 0–3). The majority of studies (59%) did not fulfil any of the MERSQI requirements for evaluation instruments, including reporting questionnaire design, wording and content. Objective outcome measures were used in only a minority of studies, with 60% relying solely on self-reported measures.

Data analysis and reporting likewise showed considerable limitations. Only one in five studies (20%) met criteria for comprehensive analysis and reporting of data. Few studies analysed their data beyond descriptive statistics to consider the generalisability and implications (13%). In many cases, studies omitted basic statistical significance tests.

Many studies did not contain key reporting elements for qualitative research as outlined in the JBI tool (see online supplemental table 3). There was clear congruity between research methodologies chosen and the research objectives and methods employed in 60% of studies. A minority of studies adequately reported their analysis (28%) and interpretation of data (25%), the potential for the researcher to have influenced data collection and interpretation (23%) and the researcher's cultural or theoretical orientation (15%). Participant voices were clearly represented through quotes in only 16/53 (30%) of mixed-methods studies and 5/10 (50%) of qualitative studies. There was a statement of ethical approval or ethics exemption in only 26 of 63 studies (40%) which used qualitative methods. No study included a statement of philosophical perspective (normally expected for qualitative research).[17]

Programme design

There was considerable heterogeneity in leadership development intervention design. It was often unclear whether established good practice for development of medical education interventions was followed, as shown in figure 2.[9 31] Only 52 studies (44%) reporting having conducted a needs assessment before

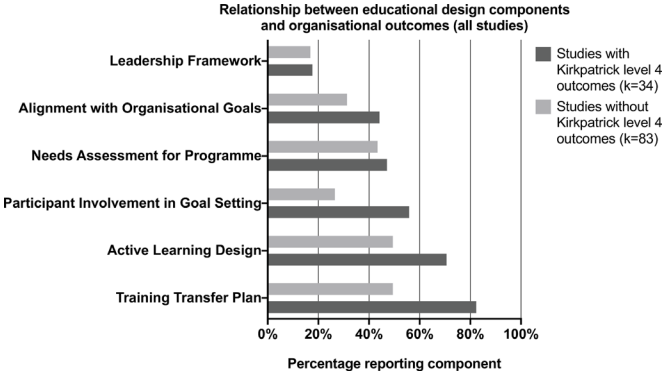

Figure 2 Educational design components: studies which reported Kirkpatrick level 4 outcomes (k=34) compared with studies that did not report Kirkpatrick level 4 outcomes (k=83).

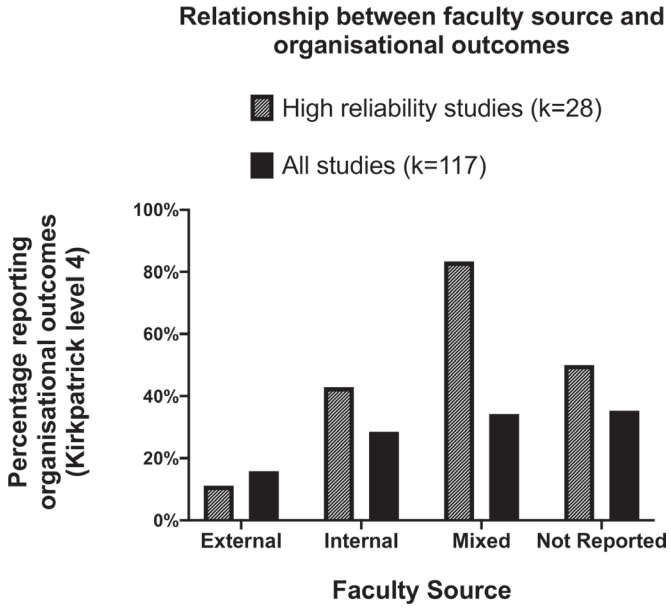

Figure 3 Relationship between faculty source and programme outcomes. Higher reliability studies were those with Medical Education Research Study Quality Indicator Score of at least 12/18 or Joanna Briggs Institute Score of at least 6/10. NR, not reported.

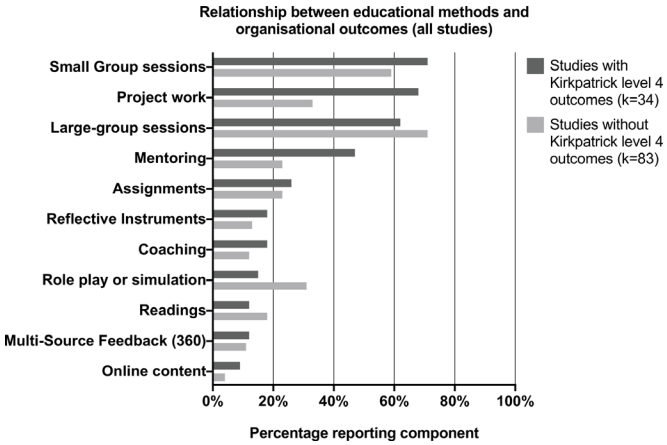

Figure 4 Educational methods: studies which reported Kirkpatrick level 4 outcomes (k=34) compared with studies that did not report Kirkpatrick level 4 outcomes (k=83).

their intervention, and only 20 studies (17%) explicitly reported using an established capability or competency framework to inform leadership programme goals and objectives. There was, however, a plan for training transfer reported or built into 68 of 117 interventions (59%).

The majority of interventions were carried out in a single hospital department (27%), single hospital (22%) or a single university (12%). Just under a quarter (23%) of interventions were conducted in multiple healthcare centres. A further 15% of studies were conducted within a specialty training programme outside healthcare centres.

Most of the studies took place in the USA (67%) or the UK (16%). The remainder of studies were in other European countries (7%), Canada (4%) or Australia (3%), with a single study each from Africa,[32] India,[33] Israel[34] and Qatar.[35]

Programmes ranged in length from 2 hours to 4 years. The median intervention length was 6 months, and the most common length was 1 year (19%). Only 18 interventions (15%) lasted longer than 1 year. Five interventions (4%) were shorter than 1 day.

Programme faculty

Programmes were predominately delivered by either in-house faculty (36%) or a mix of in-house and external faculty (32%). Programmes delivered by mixed faculty were most likely to show organisational outcomes, as shown in figure 3. The professional backgrounds, qualifications and experience of faculty were generally not reported.

Participants

The majority of programmes included doctors only (76%). Physician learners ranged from residents (60%) to full specialists (30%) and academic medical faculty (19%). Only nine studies of 117 involved doctors from more than one category. Behavioural outcomes were reported in a similar percentage of higher reliability studies for each category (85%–92%), while organisational outcomes were more commonly reported in programmes with academic medical faculty (50%) or full specialists (44%) than in programmes with only residents (20%). The 26 studies (24%) reporting multidisciplinary programmes included a combination of nurses (12%), managers (15%) and allied health professionals (9%). Most studies did not report the gender of participants (74%) or the age of participants (87%).

In terms of participant selection criteria, the majority of interventions included participants who volunteered (27%), were nominated (19%) or who applied to the programme (16%). In some cases the application process was highly competitive. Interventions were mandatory in one-fifth of studies (20%). A considerable proportion of all studies (23%) did not report the selection process for their learners, including one quarter (25%) of the studies categorised as higher reliability by MERSQI criteria.

Educational methods

A wide range of educational methods were employed in various combinations across the reviewed studies, as shown in figure 4. Most interventions included lectures (68%) and small group work (61%). Project work was included in the majority of studies with organisational outcomes (68%), but only in a minority of studies which did not report organisational outcomes (33%). Individual or team mentoring was also more prevalent where organisational outcomes were reported (47% vs 23%).

Educational content

Educational content varied considerably among interventions. The most consistent content area was leadership theory (reported in 65% of interventions). The other common content areas were performance management (44%), self-management (41%), change management (39%), communication (36%), teamwork (33%), quality improvement (30%), healthcare policy (27%), healthcare finance (26%) and leadership behaviours (20%). There were no notable educational content differences in higher reliability studies or in studies which reported organisational outcomes (Kirkpatrick level 4).

Evaluation methods

A wide range of evaluation methods were employed across the included studies. Nearly half used quantitative methods only for their evaluation (46%). Of the remainder, most studies used mixed methods (45%), with 10 studies (9%) using purely

qualitative methods. These proportions were similar in the higher reliability studies (41% quantitative, 48% mixed methods, 10% qualitative).

Four out of every five studies (82%) used questionnaires in their evaluation. Almost all of these employed Likert Scale items (92%) and one-third included open questions (34%). Only 8% used content or construct validated questionnaires. The proportion of higher reliability studies using validated questionnaires was slightly higher at 20% (MERSQI) and 18% (JBI). An additional six studies (6%) had conducted an expert review of their questionnaire for content validity only.

More than two-thirds of the included studies relied solely on self-ratings (69%). A minority of studies included ratings from subordinates (3%), peers (7%), superiors (12%) or experts (20%). The proportion of higher reliability studies which relied on self-ratings was lower (39%), with increased use of ratings from peers (14%), superiors (25%) or experts (39%).

The majority of studies (72%) included the collection of outcome data regarding behavioural changes (Kirkpatrick level 3, 57%) or organisational outcomes (Kirkpatrick level 4, 24%). Only three studies relied solely on Kirkpatrick level 1 outcomes (reaction).[36–38]

Nearly half of the studies used single group post-programme only designs (46%), with most of the other half using single group pre-programme and post-programme designs (46%). Most studies included a post-programme evaluation completed immediately at the end of the programme (90%). Only 18 studies (15%) included a longer-term evaluation. In higher reliability studies, longer-term evaluations were associated with increased reporting of organisational outcomes (56%) when compared with immediately-post designs (31%). All 16 higher reliability studies as assessed by the MERSQI used pre and post designs. Six of these included a non-randomised control group (38%), and one study included a randomised control group (6%). This was the only randomised control group used in any of the 117 studies.

Behavioural and organisational outcomes in higher reliability studies

A full summary of outcomes from all 117 studies is provided in online supplemental table 1.

There was a range of behavioural (Kirkpatrick level 3) and organisational (Kirkpatrick level 4) outcomes demonstrated in higher reliability studies.

Behavioural changes were objectively demonstrated in higher reliability studies through observed changes in behaviour,[26 27 39–43] promotions,[44 45] increased responsibilities or titles[28 46–49] and project completion.[50–52] Subjective changes in behaviour included improved communication,[39] influence,[50] delegation,[27] collaboration,[53] involvement in service improvement[47] and application of skills learnt or improved leadership in general.[39 40 54–57] These changes were indicated through interviews, free text questionnaire responses and behavioural self-assessments.

Organisational outcomes in higher reliability studies (Kirkpatrick level 4) were defined prospectively and in most cases were objectively demonstrated through leadership project impact evaluations. Projects achieved a range of outcomes, including reduced waiting times,[50] improved patient care[46 50] and cost savings.[27 46 47 50] By assessing the financial impact of projects completed during the intervention and relating this to programme costs, one higher reliability study reported a 364% financial return-on-investment (ROI).[27] Other objective outcomes included reduced organisational turnover of participants,[28] improved departmental working climate,[39] reduced sick leave[44] and increased promotion of women.[45] Organisational outcomes were subjectively indicated through reports of increased staff retention[56] and improvement in organisational effectiveness.[27] One study reported that 'intangible benefits' resulted in a 106% financial ROI.[51]

Organisational outcomes in higher reliability studies were reported more frequently from programmes delivered by a mix of internal and external faculty than from programmes delivered by only external faculty (83% vs 11%), as shown in figure 2. Organisational outcomes were also more frequently reported from interventions conducted in a whole hospital (57%) or multiple hospitals (40%), compared with interventions conducted in a single specialty (conference or outside-hospital training programme) (33%), single university (25%) or in a single department (0%). There were no notable differences in outcomes related to specific educational content.

Higher reliability studies that reported organisational outcomes were more likely have included project work (70% vs 44%), mentoring (50% vs 22%), coaching (22% vs 11%) and reflective instruments such as personality type assessments (22% vs 6%) than higher reliability studies that did not report organisational outcomes. Organisational outcomes were reported less frequently in higher reliability studies that included simulation or role play (10% vs 33%).

DISCUSSION

The aim of this review was to synthesise recent empirical evidence and explore factors associated with higher level outcomes in physician leadership development.

We found a substantial increase in the number of studies which evaluate medical leadership development interventions compared with previous reviews.[6–10 14 15] In many studies, it is still not clear whether best practices for design, delivery and evaluation are being followed.[31] It is also not clear whether there are sufficient behavioural and organisational outcomes to justify the considerable and increasing investments in medical leadership development.

Compared with previous reviews, we found an increase in the proportion of studies which report the use of active learning methods such as project work, simulation, discussions and reflections, which are widely accepted to be a vital component of leadership development[58] and which were associated in our review with increased Kirkpatrick level 4 outcomes.

No single leadership development content area was particularly associated with improved outcomes. With respect to educational methods, however, there was an association between the inclusion of individual or group project work and of mentoring with organisational outcomes. This may support the established position that educational methods are more important than specific curriculum content for leadership development.[1 58] Simulation and role play were less common in higher reliability studies which reported organisational outcomes that those that did not report organisational outcomes. This unexpected finding could result from these studies being situated in a training environment rather than a working environment. Alternatively, it could result from the evaluation process and study designs rather than from a lack of organisational impact. Studies which included simulation and role play tended to focus their evaluations on objective changes in behaviour at the expense of evaluating organisational outcomes (see online supplemental table 1). Interestingly, lacking a leadership development framework did not seem to impede programmes from reporting organisational

outcomes. This may indicate that programmes which are designed as bespoke solutions to local needs are more likely to achieve organisational impact than pre-packaged approaches to leadership development.

There was an additional association of more senior participant level with organisational outcomes. This may be related to the wider scope of influence or practice of senior physicians compared with resident physicians. It could also indicate that there is a longer post-programme development period before residents are able to have an impact on organisational outcomes. This would align with the finding that programmes which evaluated longer-term outcomes were more likely to report organisational outcomes.

Importantly, our findings indicated that leadership development interventions which used a combination of internal and external faculty were most likely to report organisational outcomes, and those interventions which used external faculty only were least likely. This could have significant implications for procurement and design of leadership development interventions across healthcare, particularly as courses run internally are associated with significantly reduced costs.[59 60]

As in previous physician leadership development reviews that used critical appraisal instruments,[7 9] we found that studies frequently did not meet criteria for high reliability. Many studies failed to report important methodological features, which restricts readers' ability to appraise studies and learn from their findings. This was particularly notable in terms of questionnaire design, with fewer than one in 10 studies using validated questionnaires or reporting their questionnaire content in detail. Most studies also did not report or analyse outcome evaluation data comprehensively. Many study designs were biased towards obtaining positive results, particularly in terms of the absence of control groups, having stringent or undisclosed selection criteria, including leading questions on questionnaires and relying solely on self-ratings. This is likely to have resulted in improved reported outcomes. The lack of evaluation quality seems to indicate perfunctory attention paid to evaluation design and precludes confident conclusions from these studies. Future studies could benefit from consulting study quality appraisal checklists such as the MERSQI and JBI in advance, in order to effectively design their evaluations.

This review does indicate that certain recommendations for improved programme evaluation are beginning to be applied into research. Whereas only 29% of the studies reviewed by Frich *et al*[8] included qualitative components, 63 (54%) of the 117 studies included in our review used mixed or qualitative methods. In a nascent and complex field such as medical leadership development research,[1 8 9 61] qualitative methods can have value in terms of establishing effective programme design features to achieve desired outcomes,[21 25 31] as well as helpful nuances of how, for whom, to what extent or in what circumstances interventions are effective or not.[9 10 62]

Additionally, many studies in this systematic review evaluated outcomes at Kirkpatrick level 3 behavioural change (57%) or level 4 organisational outcomes (24%). This is a significant improvement from previous reviews.[7 8 14] Changes in behaviour (level 3) and organisational outcomes (level 4) are more closely associated with transfer of learning to the working environment than participant reaction (level 1) and learning (level 2).[63–65]

Limitations and strengths

This review was limited by the reliability of the studies included. We attempted to control for study reliability using critical appraisal tools with cut-off scores for higher reliability studies. To the best of our knowledge, this is the first systematic review of healthcare leadership development interventions to use the JBI critical appraisal tool to critically appraise qualitative studies. The JBI tool enabled us to identify 12 additional higher reliability qualitative and mixed-methods studies which were not identified using the MERSQI. Marked heterogeneity of studies and evaluations precluded a formal meta-analysis, therefore, we adopted a meta-aggregation approach. This enabled us to highlight design components that are correlated with behavioural and organisational outcomes in higher reliability studies.

A substantial majority of studies reported only positive outcomes, which could represent a publication bias, and we limited our review to English language peer-reviewed studies. In line with Frich *et al*,[8] our database search was limited to MEDLINE, however, we augmented our database search with an extensive hand-search of reference lists and citations using Web of Science and Google Scholar. The hand-search revealed that many relevant empirical studies were absent from recent reviews despite some of those reviews searching a greater range of research databases. This could indicate flaws in healthcare leadership development literature tagging and filing procedures within medical and educational databases.

CONCLUSION

Our review has practical implications for those commissioning, designing and evaluating medical leadership development programmes in healthcare. No specific area of curriculum content and no particular leadership development framework were clearly associated with behavioural or organisational outcomes. While relevance and appropriateness of educational content is important,[31] this systematic review has more clear implications for leadership development methods than for specific content. Where possible, interventions should include projects and individual or group mentoring. Transfer of learning from the programme into learners' daily work and their organisations should be planned into the programme and where possible active learning educational designs should be employed, including opportunities for learners to set their own goals for development. External faculty should be judiciously used to supplement in-house faculty, not as a replacement for in-house expertise.

In terms of evaluation design, efforts should be made to ensure that evaluations are cost-effective and produce data that is useful for both practitioners and researchers.[66 67] Effective mixed-methods evaluation strategies should be integrated into evaluation designs. Study quality checklists such as the MERSQI and JBI could be consulted in the programme design phase to help build high quality quantitative and qualitative evaluation methods into programmes. At the minimum, evaluation design should include consideration of assessment at multiple time points, inclusion of control groups and collection of objective data, as well as collection of qualitative data from interviews, focus groups, questionnaires or observations. Programme goals and intended organisational outcomes should be explicitly considered during evaluation design[67] so that measures of organisational outcomes (including project outcomes) can be incorporated into the evaluation design. Improving study design and building robust evaluation methods into programmes will allow evaluators and educators to more effectively understand factors which are reliably associated with high level programme outcomes. This could both inform the improvement of individual programmes and

contribute to the medical leadership literature as a whole. It is only through more considered and thorough evaluation of physician leadership development programmes that we will be able to justify the investment they represent.

Twitter Oscar Lyons @oscarlyonsnz, Jan Frich @J_Frich and Jaason Matthew Geerts @jaasongeerts

Acknowledgements We would like to thank Tatjana Petrinic, University of Oxford Health CareHealthcare Librarian, for her invaluable assistance and advice in the search process.

Contributors OL, RG and JRG planned the review. OL, RG and TF screened studies for inclusion. OL, RG, JRG, AM and TF abstracted and coded studies. OL, RG, JRG, AM, TF, JF and JMG contributed to analysis, writing and editing the manuscript.

Funding Oscar Lyons was supported during this work by a Rhodes Scholarship, a Goodger and Schorstein Research Scholarship (University of Oxford) and the Shirtcliffe Fellowship (Universities New Zealand)

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iDs
Oscar Lyons http://orcid.org/0000-0001-5809-7173
Jan Frich http://orcid.org/0000-0001-9079-7508
Jaason Matthew Geerts http://orcid.org/0000-0001-6672-3859

REFERENCES

1 West M, Armit K, Loewenthal L, *et al*. Leadership and leadership development in health care: the evidence base. *Kings Fund* 2015;19022015:1–36.
2 Goodall AH. Physician-Leaders and hospital performance: is there an association? *Soc Sci Med* 2011;73:535–9.
3 Falcone RE, Satiani B. Physician as Hospital chief executive officer. *Vasc Endovascular Surg* 2008;42:88–94.
4 Spurgeon P, Long P, Clark J, *et al*. Do we need medical leadership or medical engagement? *Leadersh Health Serv* 2015;28:173–84.
5 Tasi MC, Keswani A, Bozic KJ. Does physician leadership affect Hospital quality, operational efficiency, and financial performance? *Health Care Manage Rev* 2019;44:256–62.
6 Husebø SE, Akerjordet K. Quantitative systematic review of multi-professional teamwork and leadership training to optimize patient outcomes in acute hospital settings. *J Adv Nurs* 2016;72:2980–3000.
7 Rosenman ED, Shandro JR, Ilgen JS, *et al*. Leadership training in health care action teams: a systematic review. *Acad Med* 2014;89:1295–306.
8 Frich JC, Brewster AL, Cherlin EJ, *et al*. Leadership development programs for physicians: a systematic review. *J Gen Intern Med* 2015;30:656–74.
9 Geerts JM, Goodall AH, Agius S. Evidence-based leadership development for physicians: a systematic literature review. *Soc Sci Med* 2020;246:112709.
10 Steinert Y, Naismith L, Mann K. Faculty development initiatives designed to promote leadership in medical education. A BEME systematic review: BEME guide No. 19. *Med Teach* 2012;34:483–503.
11 Stoller JK. Developing physician-leaders: a call to action. *J Gen Intern Med* 2009;24:876–8.
12 Clark J, Armit K. Attainment of competency in management and leadership: no longer an optional extra for doctors. *Clin Gov* 2008;13:35–42.
13 Leslie LK, Miotto MB, Liu GC, *et al*. Training young pediatricians as leaders for the 21st century. *Pediatrics* 2005;115:765–73.
14 Straus SE, Soobiah C, Levinson W. The impact of leadership training programs on physicians in academic medical centers: a systematic review. *Acad Med* 2013;88:710–23.
15 Sadowski B, Cantrell S, Barelski A, *et al*. Leadership training in graduate medical education: a systematic review. *J Grad Med Educ* 2018;10:134–48.
16 Reed DA, Cook DA, Beckman TJ, *et al*. Association between funding and quality of published medical education research. *JAMA* 2007;298:1002–9.
17 Aromataris E, Munn Z. *JBI Reviewer's Manual*. Joanna Briggs Institute, 2019.
18 Liberati A, Altman DG, Tetzlaff J, *et al*. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;339:b2700.
19 Hammick M, Dornan T, Steinert Y. Conducting a best evidence systematic review. Part 1: from idea to data coding. BEME guide No. 13. *Med Teach* 2010;32:3–15.
20 Husebø SE, Olsen Øystein Evjen, Olsen OE. Impact of clinical leadership in teams' course on quality, efficiency, responsiveness and trust in the emergency department: study protocol of a trailing research study. *BMJ Open* 2016;6:e011899.
21 Lockwood C, Munn Z, Porritt K. Qualitative research synthesis. *Int J Evid Based Healthc* 2015;13:179–87.
22 Kirkpatrick DL. Techniques for evaluating training programs. *Train Dev J* 1979:178–92.
23 Côté L, Turgeon J. Appraising qualitative research articles in medicine and medical education. *Med Teach* 2005;27:71–5.
24 Hannes K, Lockwood C, Pearson A. A comparative analysis of three online appraisal instruments' ability to assess validity in qualitative research. *Qual Health Res* 2010;20:1736–43.
25 Munn Z, Porritt K, Lockwood C, *et al*. Establishing confidence in the output of qualitative research synthesis: the ConQual approach. *BMC Med Res Methodol* 2014;14:108.
26 Cooper S. Developing leaders for advanced life support: evaluation of a training programme. *Resuscitation* 2001;49:33–8.
27 Orme D, Campbell C. How leadership training saves money 'service line leadership' at Nottingham University Hospitals. *Leader* 2019;3:29–36.
28 Fassiotto M, Maldonado Y, Hopkins J. A long-term follow-up of a physician leadership program. *J Health Organ Manag* 2018;32:56–68.
29 Berkenbosch L, Muijtjens AMM, Zimmermann LJI, *et al*. A pilot study of a practice management training module for medical residents. *BMC Med Educ* 2014;14:107.
30 Day DV, Sin H-P, *et al*. Longitudinal tests of an integrative model of leader development: charting and understanding developmental trajectories. *Leadersh Q* 2011;22:545–60.
31 Thomas PA, Kern DE, Hughes MT. *Curriculum development for medical education: a six-step approach*, 2016.
32 Nakanjako D, Namagala E, Semeere A, *et al*. Global health leadership training in resource-limited settings: a collaborative approach by academic institutions and local health care programs in Uganda. *Hum Resour Health* 2015;13:87.
33 Gulati K, Singh AR, Kumar S, *et al*. Impact of a leadership development programme for physicians in India. *Leadersh Health Serv* 2019;33:73–84.
34 Maza Y, Shechter E, Pur Eizenberg N, *et al*. Physician empowerment programme; a unique workshop for physician-managers of community clinics. *BMC Med Educ* 2016;16:269.
35 Al-Mutawa N, Elmahdi H, Joyce P. The implementation of a practice management programme for family medicine residents in Qatar. *Educ Prim Care* 2016;27:380–5.
36 Ringdahl EN, Tarwater KD, Lindbloom EJ. A longitudinal curriculum to address the gender gap in physician leadership. *J Grad Med Educ* 2014;6:361–2.
37 Johnson JM, Stern TA. Teaching residents about emotional intelligence and its impact on leadership. *Acad Psychiatry* 2014;38:510–3.
38 Bhatia K, Morris CA, Wright SC, *et al*. Leadership training for residents: a novel approach. *Physician Leadersh J* 2015;2:76–80.
39 Boyle DK, Kochinda C. Enhancing collaborative communication of nurse and physician leadership in two intensive care units. *J Nurs Adm* 2004;34:60–70.
40 Ruston A, Tavabie A. Fostering clinical engagement and medical leadership and aligning cultural values: an evaluation of a general practice specialty trainee integrated training placement in a primary care trust. *Qual Prim Care* 2010;18:263–8.
41 Cole DC, Giordano CR, Vasilopoulos T, *et al*. Resident physicians improve Nontechnical skills when on operating room management and leadership rotation. *Anesth Analg* 2017;124:300–7.
42 Ten Have ECM, Nap RE, Tulleken JE. Quality improvement of interdisciplinary rounds by leadership training based on essential quality indicators of the interdisciplinary rounds assessment scale. *Intensive Care Med* 2013;39:1800–7.
43 Gilfoyle E, Gottesman R, Razack S. Development of a leadership skills workshop in paediatric advanced resuscitation. *Med Teach* 2007;29:e276–83.
44 von Vultée PJ, Arnetz B. The impact of management programs on physicians' work environment and health. A prospective, controlled study comparing different interventions. *J Health Organ Manag* 2004;18:25–37.
45 Dannels SA, Yamagata H, McDade SA, *et al*. Evaluating a leadership program: a comparative, longitudinal study to assess the impact of the executive leadership in academic medicine (ELAM) program for women. *Acad Med* 2008;83:488–95.
46 Agius SJ, Brockbank A, Baron R, *et al*. The impact of an integrated medical leadership programme. *J Health Organ Manag* 2015;29:39–54.

47 McKimm J, Hickford D, Lees P, *et al*. Evaluating the impact of a national clinical leadership fellow scheme. *Leader* 2019;3:37–42.
48 Tsoh JY, Kuo AK, Barr JW, *et al*. Developing faculty leadership from 'within': a 12-year reflection from an internal faculty leadership development program of an academic health sciences center. *Med Educ Online* 2019;24:1567239.
49 Haftel HM, Swan R, Anderson MS, *et al*. Fostering the career development of future educational leaders: the success of the association of pediatric program directors leadership in educational academic development program. *J Pediatr* 2018;194:5–6.
50 Hopkins J, Fassiotto M, Ku MC, *et al*. Designing a physician leadership development program based on effective models of physician education. *Health Care Manage Rev* 2018;43:293–302.
51 Throgmorton C, Mitchell T, Morley T, *et al*. Evaluating a physician leadership development program - a mixed methods approach. *J Health Organ Manag* 2016;30:390–407.
52 Levine SA, Chao SH, Brett B, *et al*. Chief resident immersion training in the care of older adults: an innovative interspecialty education and leadership intervention. *J Am Geriatr Soc* 2008;56:1140–5.
53 Pradarelli JC, Jaffe GA, Lemak CH, *et al*. A leadership development program for surgeons: first-year participant evaluation. *Surgery* 2016;160:255–63.
54 Wurster AB, Pearson K, Sonnad SS, *et al*. The patient safety leadership Academy at the University of Pennsylvania: the first cohort's learning experience. *Qual Manag Health Care* 2007;16:166–73.
55 Bergman D, Savage C, Wahlstrom R, *et al*. Teaching group dynamics--do we know what we are doing? An approach to evaluation. *Med Teach* 2008;30:55–61.
56 Monkhouse A, Sadler L, Boyd A, *et al*. The improving global health fellowship: a qualitative analysis of innovative leadership development for NHS healthcare professionals. *Global Health* 2018;14:69.
57 Cohen D, Vlaev I, McMahon L, *et al*. The Crucible simulation: behavioral simulation improves clinical leadership skills and understanding of complex health policy change. *Health Care Manage Rev* 2017;44:246–55.
58 Rabin R. *Blended learning for leadership: the CCl approach*. 12, 2014.
59 MacPhail A, Young C, Ibrahim JE. Workplace-based clinical leadership training increases willingness to lead: appraisal using multisource feedback of a clinical leadership program in regional Victoria, Australia. *Leadersh Heal Serv* 2015;28:100–18.
60 Gagliano NJ, Ferris T, Colton D, *et al*. A physician leadership development program at an academic medical center. *Qual Manag Health Care* 2010;19:231–8.
61 Frye AW, Hemmer PA. Program evaluation models and related theories: AMEE guide No. 67. *Med Teach* 2012;34:e288–99.
62 Kwamie A, van Dijk H, Agyepong IA, *et al*. Advancing the application of systems thinking in health: realist evaluation of the leadership development programme for district manager decision-making in Ghana. *Health Res Policy Syst* 2014;12:29.
63 Klein KJ, Kozlowski SWJ. *Multilevel theory, research, and methods in organizations: foundations, extensions, and new directions*. San Francisco: Jossey-Bass, 2000.
64 Saks AM, Burke LA. An investigation into the relationship between training evaluation and the transfer of training. *Int J Train Dev* 2012;16:118–27.
65 Kennedy PE, Chyung SY, Winiecki DJ, *et al*. Training professionals' usage and understanding of Kirkpatrick's level 3 and level 4 evaluations. *Int J Train Dev* 2014;18:1–21.
66 Patton MQ. *Utilization-focused evaluation*. 4th ed. Thousand Oaks, Calif, London: SAGE, 2008.
67 Edmonstone J. Healthcare leadership: learning from evaluation. *Leadersh Health Serv* 2013;26:148–58.