City Research Online

City, University of London Institutional Repository

Citation: Goodall, A. H., Geerts, J. & Agius, S. (2019). Evidence-Based Leadership Development for Physicians: A Systematic Literature Review. Social Science and Medicine, 246, 112709. doi: 10.1016/j.socscimed.2019.112709

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/23371/

Link to published version: https://doi.org/10.1016/j.socscimed.2019.112709

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Evidence-Based Leadership Development for Physicians: A Systematic Literature Review

Jaason M. Geerts[1]
Faculty of Education
University of Cambridge

Amanda H. Goodall,
Cass Business School,
City, University of London & IZA

Stevie Agius
School of Medicine
University of Nottingham

Social Science & Medicine (forthcoming)

Abstract

Interest in leadership development in healthcare is substantial. Yet it remains unclear which interventions are most reliably associated with positive outcomes. We focus on the important area of physician leadership development in a systematic literature review of the latest research from 2007 - 2016. The paper applies a validated instrument used for medical education, MERSQI, to the included studies. Ours is the first review in this research area to create a tiered rating system to assess the best available evidence. We concentrate on findings from papers in the highly-rated categories. First, our review concludes that improvements in individual-level outcomes can be achieved (e.g. knowledge, motivation, skills, and behavior change). Second, development programs can substantially improve organizational and clinical outcomes. Third, some of the most effective interventions include: workshops, videotaped simulations, multisource feedback (MSF), coaching, action learning, and mentoring. Fourth, the evidence suggests that objective outcome data should be collected at baseline, end of program, and retrospectively. An outcomes-based approach appears to be the most effective design of programs. We also make recommendations for future research and practice.

Key words: leadership development; physicians; systematic literature review; outcomes; impact.

[1] Corresponding author jaasongeerts@cantab.net

1. Introduction

Leadership development is a burgeoning global enterprise, with an expanding number of program providers (Lacerenza, Reyes, Marlow, Joseph, and Salas, 2017) and an annual spend estimated at $50 billion (USD) (Kellerman, 2012, 2018). This equals nearly half of all funds allocated annually to cancer treatment (Quintiles IMS Institute, 2017). Research from healthcare shows that leadership development interventions can improve the capabilities of individuals and contribute to better organizational and benefit to patient outcomes (Burke and Day, 1986; Husebø and Akerjordet, 2016; Komives, Nance, and McMahon, 1998; Pfeffer, 2016; Rosenman, Shandro, Ilgen, Harper, and Fernandez, 2014; Steinert, Naismith, and Mann, 2012; Stoller 2008, 2009). However, the evidence on the effectiveness of different programs is conflicting. Indeed, questions about efficacy have led to pressure to demonstrate the effect of interventions by linking them to outcomes and establishing a clear return on investment (Avolio, Avey, and Quisenberry, 2010; Kellerman, 2018).

Our study provides a new kind of systematic review. With a focus on healthcare, we use an innovative methodology to advance understanding about which elements of design, delivery, and evaluation of leadership development interventions are most reliably linked to outcomes at the level of the individual, the organization, and of benefit to patients. Analysis is problematic when it is not immediately apparent which studies' findings are substantiated by objective evidence and which are not (Antonakis et al., 2011; Hamlin, 2010; Husebø and Akerjordet, 2016; Pfeffer, 2016; Rousseau and McCarthy, 2007; Rynes et al., 2014). This can be confusing and misleading for readers. We applied the Medical Education Research Study Quality Instrument (MERSQI), a validated instrument used for medical education, to all the included studies to assess the reliability of evidence reported in each. Ours is the first review to produce a transparent category rating system that clearly presents the strength of evidence

linking elements of programs to outcomes. This study addresses the research gaps, builds on previous review articles, and highlights the preeminent leadership development strategies.

2. Background

Formal leadership development programs are being offered in military academies, business schools, international corporations, and relatively recently, in medical schools (McKimm, Spurgeon, Needham, and O'Sullivan, 2011; Solansky, 2010). A desire for capable leaders is to be expected given the consequences of disturbingly common leadership failures (Caulkin, 2015; Kellerman, 2018; Pfeffer, 2016) and the positive influence that leaders and managers have on employee job satisfaction (Artz, Goodall, and Oswald, 2017; Rynes et al., 2014; Sellgren, Ekvall, and Tomson, 2008), employee performance (Lazear, Shaw, and Stanton, 2015), and organizational performance (Goodall, 2011; Spurgeon, Long, Clark, and Daly, 2015; Spurgeon, Mazelan, and Barwell, 2011).

In healthcare, leadership development initiatives are found to benefit patients through reduced clinical errors and mortality rates, shorter lengths of stay in hospital (Husebø and Akerjordet, 2016; Rosenman et al., 2014), and reduced costs (Mountford and Webb, 2009). However, the extraordinary growth of leadership development programs has sparked some to suggest that a "great training robbery" is occurring (Beer, Finnström, and Schrader, 2016), while others argue the investment is wasted and potentially even harmful (Blume, Ford, Baldwin, and Huang, 2010; Kellerman, 2012; McDonald, 2017; Pfeffer, 2015; Watkins, Lysø, and deMarrais, 2011). The financial outlay, significant time commitments in delivering and undertaking development programs, and the ensuing opportunity cost for those involved, make leadership development a 'high-stakes game' (Antonakis et al., 2011). Surprisingly, most leadership program designs do not incorporate robust evaluation processes (Alimo-Metcalfe and Lawlor, 2001; Geerts, 2018). Those who do follow up often rely on surveys of participants'

satisfaction (Kellerman, 2012), which offer no indication of application to the workplace (immediate or sustained) (Straus, Soobiah, and Levinson, 2013). Others frequently confine assessment to the individual, which neglects the broader organizational-level impact (Avolio, 2005). It is also potentially worrying if program designs or funding decisions are made using unsubstantiated evidence (Phillips, Phillips, and Ray, 2015; Rousseau, 2006; Zaccaro and Horn, 2003), particularly in fields such as healthcare where people's lives, safety, health, or well-being are on the line (Bruppacher et al., 2010; Hannah, Uhl-Bien, Avolio, and Cavarretta, 2009; Salas, Tannenbaum, Kraiger, and Smith-Jentsch, 2012). Thus far, the evidence on the effectiveness of different programs is equivocal, with effect sizes ranging from -1.4 to 2.1 (Collins and Holton III, 2004). The majority of leadership interventions are generally well-rated (Frich, Brewster, Cherlin, and Bradley, 2015; Steinert et al., 2012); however, worryingly, some authors report that the transfer of learning to the workplace is low, with as few as five per cent of trainees claiming to have successfully applied their skills on the job (Gilpin-Jackson and Bushe, 2007). More concerning are accounts of programs that have apparently failed altogether (DeNisi and Kluger, 2000; Kwamie, van Dijk, and Agyepong, 2014; Malling, Mortensen, Bonderup, Scherpbier, and Ringsted, 2009).

This highlights an underlying problem and a need, addressed by this paper, to clarify explicitly which elements of leadership development interventions are empirically linked to improved outcomes at different levels (Day and Sin, 2011; Hannum and Bartholomew, 2010; Ireri, Walshe, Benson, and Mwanthi, 2011; Klimoski and Amos, 2012; Powell and Yalcin, 2010).

2.1. Physician leadership development

Our study examines physician leadership development for several reasons. First, healthcare organizations employ 234 million people globally with a spend projected to reach

8.7 trillion (USD) by 2020 (World Health Organization, 2018). Second, formal leadership development programs for doctors are relatively recent and attempts to examine their impact, while growing, is limited (Dine, Kahn, Abella, and Shea, 2011; Ireri et al., 2011; Lee and Hall, 2010; McAlearney, 2010; Stoller, 2008, 2009). Third, physicians determine how a considerable portion of resources are allocated, and they play a vital role in driving improvement initiatives (Bohmer, 2011; Byrnes, 2016; Chadi, 2009; Daly, Jackson, Mannix, Davidson, and Hutchinson, 2014; Denis and van Gestel, 2016; Dickson and Van Aerde, 2018; Geerts, 2019). Fourth, there is growing evidence showing that engaging physicians in leadership is linked to improved patient care and organizational performance (Falcone and Santiani, 2008; Goodall, 2011; Spurgeon et al., 2015, 2011; Tasi, Keswani, and Bozic, 2017). Finally, given the common contention that the currency of success in leadership development *is* the transfer or application of learning to the workplace, this field provides an additional and important measurable outcome: benefit to patients (Day, Fleenor, Atwater, Sturm, and McKee, 2014; Edmonstone, 2013; Lacerenza et al., 2017; Reed et al., 2007).

With reference to the generalizability of our focus on healthcare, the challenges that physicians face may be viewed as analogous to those faced by leaders in other sectors (Edler, Adamshick, Fanning, and Piro, 2010). For example, physician leaders often function as decision-makers in large, complex, high-intensity environments with constrained budgets, as well as alternating leadership and membership positions within teams or coalitions (Perry, Mobley, & Brubaker, 2017; Taylor, 2010). Similarly, the core skills that physician leaders require, such as adaptability, enabling and motivating others to realize a common vision, and shaping organizational culture, are common among most leaders. Finally, there is a considerable overlap between approaches to leadership development in healthcare and in other domains (Geerts, 2018). In summary, the features of physician leadership development and

the potential generalizability to leadership development in other sectors make it an interesting case.

2.2. *Previous reviews*

Previous literature reviews on leadership development for physicians have found a link between interventions and self-reported individual-level outcomes, such as increased knowledge, skills, behaviours, and competence, as well as increased self-awareness, motivation, and attitudes toward leadership (Frich et al., 2015; Rosenman et al., 2014; Straus et al., 2013). Leadership programs are also reported to contribute to objective individual outcomes, such as promotions, assuming leadership roles, and improved multisource feedback scores and performance ratings (Frich et al., 2015; Rosenman et al., 2014; Straus et al., 2013). Studies have shown that leadership programs can also influence organizational outcomes such as reduced absenteeism (Straus et al., 2013), as well as clinical outcomes, such as length of stay in hospital, decreased clinical error rates, decreased morbidity and mortality rates, and increased scores on quality of care indicators (Frich et al., 2015; Husebø and Akerjordet, 2016; Rosenman et al., 2014).

Lectures have been identified as the most common developmental activity, and, though less frequent, simulations and action learning are also commonly included in development programs (Frich et al., 2015; Rosenman et al., 2014). In terms of post-program evaluation, a limited number of studies involved a control group, and many relied on self-reports, rather than assessment sometime after the intervention, which leads to greater understanding about the transfer or application of learning to the workplace (Frich et al., 2015; Husebø and Akerjordet, 2016; Rosenman et al., 2014). Finally, while two reviews reported studies of interventions that failed, they did not investigate the suggested causes, which could be helpful to inform the design of future programs (Husebø and Akerjordet, 2016; Straus et al., 2013).

Importantly, and of relevance to our study, the authors of previous reviews indicated that given the relatively low quality of evidence, their ability to determine best practice for leadership development was limited (Husebø and Akerjordet, 2016; Rosenman et al., 2014). They suggest that a higher degree of evidence is needed that links elements of programs to outcomes, particularly at the organizational and benefit to patients' levels (Husebø and Akerjordet, 2016; Rosenman et al., 2014). Our study has striven to do this.

Our study attempts to address common methodological flaws that previous review authors highlight, such as relying exclusively on subjective data and incomplete reporting of study details, which restrict readers' ability to learn from articles' findings (Husebø and Akerjordet, 2016; Reed et al, 2008; Straus et al., 2013).

3. **Methods**

The research question guiding our systematic review was: what reliable evidence exists of approaches to the design, delivery, and evaluation of leadership development for physicians that are associated with improved outcomes?

After establishing a formal research protocol to guide the process (Petticrew and Roberts, 2006), two researchers worked independently to enhance the quality and objectivity of the study (Moher, Liberati, Tetzlaff, and Altman, 2009). The design of this review was informed by three resources: The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al., 2009), the Cochrane Review Handbook for Systematic Reviews of Interventions (The Cochrane Collaboration, 2011), and the Cook and West (2012) strategy for conducting systematic reviews in medical education. We then applied the Participants, Interventions, Comparison, Outcomes, and Study Design (PICOS) framework (Richardson, Wilson, Nishikawa, and Hayward, 1995) to specify the key study elements (see Table 1).

Insert Table 1 about here

We categorized the reported outcomes of leadership programs according to a modified version of the four-level Kirkpatrick and Kirkpatrick (2006) model for development programs (see Table 2). Level 1 describes participant satisfaction (Post-Program Evaluations (PPE's)), Level 2a reflects changes in participants' attitudes or perspectives, such as increased engagement and aspirations to lead, Level 2b concerns improved knowledge and skills, and Level 3a denotes self-reported changes in participants' behaviour. Level 3b, which refers to objective indications of behaviour change, such as improved Multisource Feedback (MSF) results (pre and post), was added, and Level 4a, organizational impact, such as decreased absenteeism or implementing a new program, and Level 4b, benefit to patients outcomes, such as a decrease in patient mortality rates, were separated.

Insert Table 2 about here

A final preliminary step was to undertake a search for existing literature reviews on leadership development for physicians, through which we identified six articles. We used these reviews to inform the design of this study. To our knowledge, no other review appeared to systematically isolate the most reliable evidence of elements of interventions that are linked to outcomes based on the methodological quality of the included studies. This novel approach separates the more robust evidence from somewhat limited or uncertain reports (Geerts, 2018), which is why we elected to take this approach. Instead, most review authors tended to present raw data (such as the demographic characteristics of the samples) with descriptions of the studies' reported outcomes and overarching observations.

3.1. Criteria for inclusion and exclusion

The next step involved conducting a systematic literature review of leadership development interventions. Studies were included in our review if they met the following

criteria: a) they evaluated the effectiveness of a leadership development intervention (rather than simply presenting a model, theory, or a program that was not evaluated); and b) their sample included physicians, and/or physicians and other healthcare professionals. In addition to the focus on physicians mentioned above, this inclusion criteria enabled comparisons between the outcomes of leadership development programs with physician-only versus interdisciplinary samples, which is an important consideration in the field (Frich et al., 2015).

Reports that focused on one individual task, such as making a business plan, a single capability, such as innovation, or programs where leadership was only one of many learning outcomes, were not included.

3.2. Literature search

The search strategy was guided by two specialist librarians from the University of Cambridge (Cahill, Robinson, Pettigrew, Galvin, and Stanley, 2018). The search was conducted using the following electronic databases: Business Source Complete, ABI, ERIC, Pubmed/Medline, Embase, Scopus, and Web of Science, as well as the Cochrane Central Registry. Articles were limited to those published in English-language peer-reviewed journals from January 2007 to December 2016. The terms used in all the searches were: "lead*" AND ("educat*" OR "develop*" OR "teach*" OR "taught" OR "train*"). The population was not specified because various synonyms of "physician" (e.g. doctor, resident, consultant, oncologist) are used in the titles and keywords of articles. Unpublished studies and popular leadership literature were not included. The initial search yielded a provisional sample of 18,999 records, which was predictably large. Scanning the titles of each article reduced the potentially relevant studies to 600 records. In the next stage, we examined the abstracts and, if necessary, the full texts. Relevant studies not identified in the initial search were added via

citation chasing after reviewing the bibliographies. Twenty-five unique empirical studies met the inclusion criteria (see Fig. 1).

Insert Figure 1 about here

3.3. Coding

Details of the 25 included studies were then coded in four broad categories: study design, sample, program, and evaluation (see Table 4 and Table 5). The program goals, leadership theories, and the topics or curricular content addressed in the included studies were also coded, but were too heterogenous to include in the analysis.

Insert Table 3 and Table 4 about here

3.4. MERSQI

In order to isolate the most reliable evidence of leadership development effectiveness, we began by assessing the methodological quality of the included studies using a validated instrument, the Medical Education Research Study Quality Instrument (MERSQI) (Reed et al., 2007). This tool was designed in response to widespread acknowledgement of deficiencies in the quality of medical education research and to a desire for increased methodological rigor (Dauphinee and Wood-Dauphinee, 2004; Lurie, 2003), reminiscent of the Terpstra (1981) classification system. Other instruments were considered for this review, including the Cochrane Assessment Tool for Non-Randomized Studies of Interventions (Sterne, Higgins, and Reeves, 2014), which is not specific to leadership development programs, and the Newcastle-Ottawa Scale (NOS) (Wells et al., 2016), which does not feature numerical score components.

Quantifying aspects of study design highlights those that contribute to robust and reliable findings and enables quick tabulation of these key quality indicators, as well as comparisons among studies. To our knowledge, our review is unique because it includes the

complete MERSQI scores for each section, rather than just the total for each study or the mean for the included study set (Rosenman et al., 2014) (see Table 7). This breakdown increases transparency and provides readers with a detailed assessment of the methodological strengths and weaknesses of each reviewed study, which can then be used to assess the validity of their key findings. MERSQI includes ten items pertaining to six domains of study quality: design, sampling, type of data (subjective or objective), validity, data analysis, and outcomes (see Table 5). Each of these aspects is scored on a three-point ordinal scale and the results are summed to produce a total score out of 18, with a minimum of 5. We replaced the instrument term "appropriateness" of data analysis with "comprehensiveness" in order to avoid placing a value judgment on the included studies.

Insert Table 5 about here

3.5. Category groupings

To focus our key findings, we created three categories of methodological reliability (gold, silver, bronze) (see Table 6). This is similar in concept to the five levels of methodological rigor of organizational development (OD) studies that Terpstra (1981) created. Our groupings were created by analyzing combinations of the total MERSQI scores for each study, plus combinations of key methodological indicators of robust and reliable findings. For example, studies that included a control group (where there was no 'treatment' given), collected objective data, used detailed forms of data collection, conducted sophisticated and comprehensive data analysis, and targeted Level 4b outcomes would receive a minimum of 14 points, the lowest mark of the gold category. Placing studies in this category would indicate a high level of reliability of their findings. Leaving out one of these elements would tend to result in a minimum score of 12, the lowest mark of the silver category. Leaving out two of those elements would further challenge the reliability of a study's findings and consequently

result in a score lower than 12, which would be categorized as a bronze study. Bronze studies also tended to rely on self-ratings, which have been shown to be potentially unreliable as single sources of data (Berg and Karlsen, 2012; Blume et al., 2010; Taylor, Russ-Eft, and Taylor, 2009). The intention in creating these groupings was to address the need suggested by previous review authors (Day and O'Connor, 2003), and to identify aspects of program design, delivery, and evaluation that are empirically linked to improved outcomes. These are unique features of this review.

Insert Table 6 about here

The gold category contains MERSQI scores from 14 to 18. Gold studies correlate objective outcome information by using pre-program data and assessment sometime after the intervention (post-post) to measure whether there has been a relative and sustained change in outcomes. Gold studies also compared the outcomes of the intervention participants to those of a control group (e.g. Ten Have, Nap, and Tulleken, 2013). The silver category includes studies that obtained MERSQI scores of between 12 and 13.5. Silver studies correlated objective data with outcomes, but they had methodological limitations or omitted study details (e.g. Kuo, Thyne, Chen, West, and Kamei, 2010). Lastly, studies in the bronze category received MERSQI scores from 5 to 12. Bronze studies were characterized by the inclusion of findings typically based on subjective ratings or authorial perceptions, rather than objective data, or, again, by methodological limitations or omission of study details.

Finally, once we had isolated the most reliable elements of leadership development program design, delivery, and evaluation based on the gold and silver studies, we tracked the frequency with which programs described in the included studies implemented them. The intention was to ascertain the extent to which there was a research/practice divide.

4. **Results**

The MERSQI scores for the 25 included studies ranged from 5 to 15 with a median score of 10 and a mean of 10 (+2.6), which places the mean in the bronze category (see Table 7). These results mirror those in the Rosenman et al. (2014) review.

Insert Table 7 about here

4.1. Raw data findings

The raw data collected from the included studies can be found in Table 3 and Table 4. The variable "k" is used to denote an included article, whereas "n" is used to represent a respondent in an included study. To summarize, among the 25 studies, the most common research design was case study (k = 14), at a single site (k = 24), featuring one iteration of the intervention (k = 19), that involved collecting both qualitative and quantitative data (k = 12). Eight studies used only qualitative information. Data were collected most often using questionnaires (k = 21) and self-ratings (k = 22), with more than half of the studies (k = 13) relying on single raters, which prevents data triangulation and can increase response bias (Malling et al., 2009; Solansky, 2010). Evaluation most often involved assessments at the end of interventions (post) (k = 14); however, half of the studies (k = 12) did not include a measure sometime after (post-post), which precludes assessing the sustained impact of the intervention in the workplace. Similarly, 56 per cent of studies (k = 14) did not include a pre-intervention or baseline measurement, which can serve as a reference point for relative change, and only two studies (8%) combined pre, post, and post-post assessments. Most programs took place in North America (k = 15), with only one study from Africa and none from Asia, the Middle East, or Central or South America. The durations of the interventions were highly variable, ranging from one day to four years, with the majority being longer than eight months (k = 14). Programs were often delivered in-house (k = 18), and frequently combined internal and

external faculty (k = 9). Women made up 66 per cent of the samples and 53 per cent of the physician-only samples were women (k = 15). The most common participant samples were junior physicians (k = 9), and those who were nominated (k = 7) or volunteered (k = 6).

Thirty-three different developmental activities were included in a heterogenous series of combinations, with only one study utilizing a single activity. The most common activities were workshops (k = 14), reading assignments (k = 11), small group discussion/work (k = 11), 360-degree assessments (360s)/multisource feedback (MSF) (k = 9), and simulations/role plays (k = 9). Only eight programs involved lectures, possibly suggesting that the traditional didactic default is shifting to include more experiential methods (Blumenthal et al., 2014; Steinert et al., 2012). The most frequently reported outcomes and benefits were Post-Program Evaluations (PPEs) (n = 21, Level 1), self-reported increased skills (n = 13, Level 2b), knowledge (n = 12, Level 2b), and behaviours (n = 10, Level 3a). Nearly a third of studies (k = 7) relied exclusively on subjective, individual outcomes at the Kirkpatrick Levels 1 – 3a. Only five studies (20%) reported organizational outcomes (Level 4a) and only six (24%) claimed impact on patient outcomes (Level 4b). No study enabled participants to set their own evaluation outcome metrics, despite the potential benefit of increased perceived relevance to their specific professional contexts that personalization offers (Burke and Hutchins, 2007; Knowles, 1984).

4.2. Elements of leadership development programs reliably linked to outcomes and impact

In this section, we isolate the elements of leadership development from our key findings in studies that obtained a gold or silver rating (see Table 8). Evidence from these studies suggests that leadership development can facilitate increased self-ratings of competence, self-efficacy, self-awareness, and leadership knowledge and skills (Dannels et al., 2008; Day, Tabrizi, Kramer, Yule, and Ahn, 2010; MacPhail, Young, and Ibrahim, 2015; Patel et al., 2015). Interventions are also associated with increased frequency of observable leadership

behaviours, including those related to technical performance, decision-making, communication, and teamwork (Ten Have et al., 2013). Development programs have been found to have a positive impact on career progression (Dannels et al., 2008; Day et al., 2010; Kuo et al., 2010) and on increased aspirations to lead following an intervention (Dannels et al., 2008; MacPhail et al., 2015; Patel et al., 2015). Finally, and importantly, leadership development is associated with organizational benefits (Level 4a) (Husebø and Akerjordet, 2016; Patel et al., 2015; Rosenman et al., 2014) and benefit to patients (Level 4b) (Ten Have et al., 2013; Weaver, Dy, and Rosen, 2014).

Several developmental activities were reliably correlated with outcomes. These include workshops, followed by videotaped simulations with expert and peer feedback, which have been shown to be effective in improving technical, teamwork, communication, and leadership skills, as well as in enhancing self-awareness (Patel et al., 2015; Ten Have et al., 2013). Action learning, or leadership impact projects, can facilitate a variety of outcomes at the organizational and the benefit to patients levels and these activities include the application of learning directly to the workplace (Kuo et al., 2010; MacPhail et al., 2015; Patel et al., 2015). Though a common term in the academic literature, 'action learning' in this context differs from action learning sets, common in the UK, which are akin to group coaching.

Experiential approaches to leadership development are becoming more common. The suggestion that they are more effective for transfer of learning than the traditional default of lecture-centric programs was supported in this review (Kolb, 1984; Steinert et al., 2012). For example, interactive workshops were cited in 14 interventions, and the same number of interventions included action learning projects as those that used lectures (k = 8). Coaching, 360s/multisource feedback, and mentoring were also shown to increase performance, enhance self-awareness, and effectively support other developmental activities (Bowles, Cunningham,

Rosa, and Picano, 2007; Day et al., 2010; Leskiw and Singh, 2007; McCauley, 2008; Watkins et al., 2011).

An outcomes-based program design that explicitly links the goals, desired outcomes, content, delivery, and evaluation seems to be optimal (Geerts, 2018; Kuo et al., 2010; MacPhail et al., 2015; Nabi, Liñán, Fayolle, Krueger, and Walmsley, 2017). It is most beneficial if these desired outcomes are informed by a pre-intervention needs and gap analysis (Kuo et al., 2010; Malling et al., 2009) and a capability framework (Garman, McAlearney, Harrison, Song, and McHugh, 2011; Kuo et al., 2010; Ten Have et al., 2013). Furthermore, incorporating Knowles's (1984) principles of adult learning into the design of leadership interventions is reported to enhance their effectiveness (MacPhail et al., 2015; Ten Have et al., 2013).

Certain factors of organizational culture can significantly enhance or corrode the effectiveness of programs and, most importantly, the transfer of learning to the workplace (Cheng and Hampson, 2008; Kuo et al., 2010; Malling et al., 2009; Tracey and Tews, 2005). While both physician-only (Day et al., 2010; Ten Have et al., 2013) and interdisciplinary (Dannels et al., 2008) leadership development programs have been shown to be effective, these two approaches have not been directly compared. Embedding a leadership program in a medical residency program has been shown to work (Patel et al., 2015) and can contribute to self-reports of increased clinical skills (Kuo et al., 2010).

Our findings support several effective evaluation components, such as targeting objective behavior change (Level 3b), organizational impact (Level 4a), and benefit to patients (Level 4b) outcomes by collecting data at different points. These time points comprise measuring at baseline, at the end of an intervention (post), and retrospectively (post-post) to assess the relative and sustained outcomes (Dannels et al., 2008; Ten Have et al., 2013). Other key factors include collecting quantitative and objective data through external raters and by

using formal statistics (Dannels et al., 2008; Malling et al., 2009; Ten Have et al., 2013), and comparing individual or team performance to those in a control group or a non-intervention population (Dannels et al., 2008; Day et al., 2010). This application can take time (Abrell, Rowold, Weibler, and Moenninghoff, 2011; Dannels et al., 2008). Participants' self-ratings of leadership capabilities can also decrease from baseline to the end of an intervention because of having developed a deeper understanding of leadership or as a result of increased self-awareness. In other situations, participants' self-ratings of their leadership capabilities, confidence, and self-efficacy can decrease from the end of the intervention to post-post ratings when they experience challenges applying their learning to the workplace (Fernandez, Noble, Jensen, and Chapin, 2016; Sanfey, Harris, Pollart, and Schwartz, 2011). This may also extend to substantial declines in team performance and clinical outcomes over time (Kwamie et al., 2014). Evaluation is not only beneficial from a research and demonstrating ROI perspective, it has been shown to enhance the outcomes of programs (Latham and Locke, 1983; Watkins, Lysø, and deMarrais, 2011).

Program outcomes measurements tend to be focused at the level of the individual, with only 20 per cent of included studies measuring and reporting organizational outcomes. However, it is important to assess outcomes beyond the individual to organizational (Level 4a) and benefit to patients (Level 4b) levels (Ten Have et al., 2013). Mixed methods are arguably preferable for analyzing the complexities of leadership development. Quantitative data can substantiate findings and track frequency distribution among responses, while qualitative responses can illuminate the nuances of how, for whom, to what extent, or in what circumstances interventions were effective or not (Kwamie et al., 2014; Steinert et al., 2012).

Insert Table 8 about here

Many questions are not addressed in these studies. For example, evidence supporting optimal development strategies for different levels of seniority, domains, or professions, is absent. Research that examines the providers of programs is also missing. It might be interesting to compare programs designed and facilitated by university academics, to those designed by consultants and to those created by in-house human resources or organizational development professionals. Other factors for future research inquiries include location and length of program, ideal combinations of developmental activities, and innovative approaches to leadership development. Finally, leadership development research tends to focus exclusively on individual interventions in isolation, rather than considering leadership development more broadly across an organization as a combination of programs and formal and informal activities.

5. **Discussion**

This study was inspired by the growing cost and associated skepticism about the true yield of leadership development interventions in the field of healthcare. We have attempted to respond to demands from previous review authors and practitioners for further empirical clarification about specific, effective approaches to, and benefits of, different types of leadership programs. In our study, we investigate the most reliable elements of design, delivery, and evaluation of interventions that are linked to improved outcomes, including at the organizational and benefit to patients levels. To isolate these, our literature review uses a validated instrument, MERSQI, to evaluate the quality of evidence. To our knowledge, this is the only review of leadership development for physicians to publish transparently the full set of MERSQI scores and to base the key findings on the most reliable evidence.

As suggested above, our review findings demonstrate that leadership interventions can positively influence outcomes at the individual level and can help to facilitate organizational and clinical outcomes, including quality improvement in a healthcare setting.

5.1. Implications for future research

Our study highlights several areas for future research consideration. To continue to address the need to isolate the most reliable available evidence, we suggest a strategy to augment an empirical approach for future studies. We propose that investigators design and conduct studies that can answer outstanding questions in the field using a rigorous methodology, akin to the gold standard criteria used in this review. In addition, it may be helpful if authors, when submitting manuscripts for publication, provide clear and transparent descriptions of the methods, analysis, findings, and conclusions in a standardized form and report the MERSQI (or other) scores for each domain, as well as the total score. This process could precipitate two potential benefits: it would offer a standardized approach for publications, allowing readers to easily appraise the reliability of evidence; and this transparency may also motivate researchers to include key elements of high-quality research. For example, knowing that they will need to report their outcomes score explicitly may encourage researchers to include Level 4a or 4b outcome metrics in their studies to receive the highest points for this category. This process could make the evidence supporting the knowledge base in the field more transparent and potentially, more robust.

Our second recommendation is to reevaluate MERSQI, particularly the categories that are not immediately objective, such as the "appropriateness of data analysis". Making each category objective could potentially minimize bias when assessing future studies. It would also be valuable to assess the extent to which MERSQI is generalizable to other fields. The majority of categories, such as collecting objective versus subjective data and including a control group, do not appear to be field-specific (Geerts, 2018; Terpstra, 1981). Other

components require further consideration, such as whether there are comparable Level 4b outcomes in other domains. For example, would organizations in the financial sector really consider benefit to clients to be the ultimate outcome of leadership development? It seems that in many private sector corporations, this is not the case (Kellerman, 2018). Similarly, who would the "clients" be for the military? Clarifying the generalizability of a study quality assessment instrument could facilitate comparisons of studies and their findings across sectors.

Our third recommendation identifies some of the unanswered questions about designing optimal development programs for physicians and other professionals. Many aspects of leadership development warrant further investigation, such as the value of physician-only programs, which proponents suggest enhance learning, since they provide a safe space to discuss issues to which fellow learners can relate, given their similar backgrounds and responsibilities (McAlearney, Fisher, Heiser, Robbins, and Kelleher, 2005; Vimr and Dickens, 2013). A comparable argument is made in favor of role- or level of seniority-specific interventions, such as executive leaders. Conversely, many believe that interdisciplinary or mixed leadership programs are advantageous, mainly for the purpose of mirroring the collaboration that is needed in the workplace and breaking down silos. Should interventions be modified according to different models of leadership in healthcare organizations, such as unitary, dyadic, or team-based approaches? And what strategies are there for maximizing "training transfer", the transfer of learning from leadership interventions to the workplace, a need which was also highlighted in the Rosenman et al. (2014) review. A further question concerns the extent to which principles of optimal leadership development are generalizable to different professional domains and cultures, and how those principles for individual interventions relate to leadership development more broadly?

Another consideration is how leadership development can contribute to health promotion, community development, and capacity-building beyond individual organizations.

This can advance culture change aimed at tackling health needs or inequalities in populations by way of the multiplier effect (Hawe and Shiell, 2000).

Finally, in a field where a substantial portion of formal leadership development is offered by private companies, we believe that healthcare and management scholars in universities could play a more central role by serving as arbiters of quality in both the design of optimal leadership development and in the assessment of program effectiveness.

5.2. Implications for practice

Practitioners could attempt to apply and thereby test the key findings of this review in their organizations. It seems surprising that so few are being consistently implemented. For example, fewer than half of the studies included in our review reported conducting a needs analysis (k = 10), and only two collected baseline, post, and post-post data. Benefit to patients is recognized as an important goal of leadership development for physicians; and yet, only five included studies used it as an outcome metric. One possible explanation for why these elements are included so seldom is feasibility linked to time and cost, or suspicions regarding attribution between programs and outcomes. However, the potential impact of leadership development programs and the pressure to justify the return on investment (ROI) may in future prompt providers to consider these principles more carefully. Further concrete examples of best practice in both research and program design, delivery, and evaluation may demonstrate convincingly the viability of these strategies. This level of testing is beneficial for the academic and practitioner worlds alike. Experiments in local contexts can provide further evidence, insights, nuances, and collective learning (Dietz et al., 2014). This application includes attempting to apply Knowles's principles of adult learning to leadership development programs, while concomitantly investigating if these principles should be modified or if a new list should be created specifically for leadership development.

Given that interventions have the potential to underperform or fail, it is important to take an evidence-based approach to leadership development. Rather than prescribing one program for all, which is not an optimal approach, we believe that the process involved in applying an outcomes-based theoretical model, along with the evidence-based program components identified in this review, is most effective (Geerts, 2018).

The first step involves conducting a needs and gap analysis with relevant stakeholders to inform the selection of the program's desired outcomes.

Step two is selecting the desired outcomes for the program. These outcomes should include enhancing self-awareness, self-efficacy, and leadership knowledge, skills, and behaviours. In healthcare leadership development, it is important to also include organizational (Level 4A) and benefit to patients outcomes (Level 4B). Enabling participants to select their own goals and desired outcomes aligns with the principles of adult learning and can enhance their perception of the relevance and utility of the program.

Step three is selecting explicit goals for the program that are aligned with organizational strategy or priorities and are linked directly to the desired outcomes and the evaluation framework.

Step four is selecting the participants intentionally based on their suitability to fill the needs and gaps identified during step one.

Step five is selecting the program structure, content, faculty, and developmental activities using an outcomes-based design that are included specifically according to their intended efficacy in facilitating the achievement of the desired outcomes (Dale, 1969). In terms of the developmental activities, offering a variety is important to accommodate different learning preferences. Consideration should be given to experiential activities, including action learning projects and interactive workshops involving video-taped simulations and peer and

expert feedback, as well as to coaching, 360 assessments or multisource feedback, and mentoring.

Step six is devising an evaluation framework for the program overall, for individual activities, and for participants. The framework should include subjective and objective data at the individual level, as well as metrics pertaining to organizational and benefit to patients outcomes. Data should be collected at baseline, at the end of the intervention, and six to nine months following to assess relative and sustained improvements.

Step seven is conducting a barriers assessment of organizational cultural factors that may hinder the achievement of desired outcomes and attempting to circumvent or remove them by incorporating "training transfer" or application of learning strategies.

The final step is ensuring that the principles of adult learning or principles of leadership development have been addressed in the program design.

Therefore, the most effective approach to designing leadership development programs is to incorporate an outcomes-based model using the most reliable evidence-based components identified in this review.

6. **Limitations**

Our systematic review has some limitations. Restricting it to published peer-reviewed articles may have limited the scope; however, this choice aligned with our goals of isolating the highest-quality evidence (Cook and West, 2012). We chose to focus on physicians, which may have reduced the generalizability to other groups; however, common principles are known to apply across the professions in leadership development (Bryson, Forth, and Stokes, 2017). A further possible limitation is the high level of heterogeneity of the designs, reporting, interventions, and assessments of the included studies. This heterogeneity, as well as the small sample size, precluded a formal meta-analysis.

7. Conclusion

The global interest and investment in leadership development calls for clear evidence of impact, especially in sectors as important as healthcare, which is under great financial stress. Because of the growing evidence demonstrating the benefits of physician leadership and having doctors in leadership positions, our study focuses on physician leadership development. Through an innovative methodology, this review clarified the most reliable elements of design, delivery, and evaluation of leadership interventions that are empirically linked to positive outcomes. We hope this review advances understanding about effective leadership development for physicians.

References

Abrell, C., Rowold, J., Weibler, J., and Moenninghoff, M. 2011. Evaluation of a long-term transformational leadership development program. Zeitschrift Für Personalforschung. 25 (3), 205–224.

Alimo-Metcalfe, B., and Lawlor, J. 2001. Leadership development in UK companies at the beginning of the twenty-first century: lessons for the NHS? Journal of Management in Medicine. 15 (5), 387–404.

Antonakis, J., Fenley, M., and Liechti, S. 2011. Can charisma be taught? Tests of two interventions. Academy of Management Learning and Education. 10 (3), 374–396. Retrieved from https://doi.org/10.5465/amle.2010.0012.

Artz, B. M., Goodall, A. H., and Oswald, A. J. 2017. Boss competence and worker well-being. Industrial Labor Relations Review. 70 (2), 419–450. Retrieved from https://doi.org/10.1177/0019793916650451.

Avolio, B. J. 2005. Leadership development in balance: made/born. Oxford: Psychology Press

Avolio, B. J., Avey, J. B., and Quisenberry, D. 2010. Estimating return on leadership development investment. The Leadership Quarterly. 21 (4), 633–644.

Bearman, M., O'Brien, R., Anthony, A., Civil, I., Flanagan, B., Jolly, B., … Nestel, D. 2012. Learning surgical communication, leadership and teamwork through simulation. Journal of Surgical Education. 69 (2), 201–207.

Beer, M., Finnström, M., and Schrader, D. 2016. Why leadership training fails - and what to do about it. Harvard Business Review. 94 (10), 50-57.

Berg, M. E., and Karlsen, J. T. 2012. An evaluation of management training and coaching. Journal of Workplace Learning. 24 (3), 177–199.

Bergman, D., Fransson-Sellgren, S., Wahlstrom, R., and Sandahl, C. 2009. Impact of short-term intensive and long-term less intensive training programmes. Leadership in Health Services. 22 (2), 161–75.

Blume, B. D., Ford, J. K., Baldwin, T. T., and Huang, J. L. 2010. Training transfer: a meta-analytic review. Journal of Management. 38 (4), 1065–1105.

Blumenthal, D. M., Bernard, K., Fraser, T. N., Bohnen, J., Zeidman, J., and Stone, V. E. 2014. Implementing a pilot leadership course for internal medicine residents: design considerations, participant impressions, and lessons learned. BMC Medical Education. 14 (257), 1–11.

Bohmer, R. M. J. 2011. The four habits of high-value health care organizations. The New England Journal of Medicine. 365 (22), 2045–2047. Retrieved from https://doi.org/10.1056/NEJMp1111087.

Bowles, S., Cunningham, C. J. L., Rosa, G. M. D. L., and Picano, J. 2007. Coaching leaders in middle and executive management: goals, performance, buy-in. Leadership and Organization Development Journal. 28 (5), 388–408.

Bruppacher, H. R., Alam, S. K., LeBlanc, V. R., Latter, D., Naik, V. N., Savoldelli, G. L., … Joo, H. S. 2010. Simulation-based training improves physicians' performance in patient care in high-stakes clinical setting of cardiac surgery. Anesthesiology. 112 (4), 985–992. Retrieved from https://doi.org/10.1097/ALN.0b013e3181d3e31c.

Bryson, A., Forth, J., and Stokes, L. 2017. Does employees' subjective well-being affect workplace performance? Human Relations. 70 (8), 1017–1037.

Burke, L. A., and Hutchins, H. M. 2007. Training transfer: an integrative literature review. Human Resource Development Review. 6 (3), 263–296.

Burke, M. J., and Day, R. R. 1986. A cumulative study of the effectiveness of managerial training. Journal of Applied Psychology. 71 (2), 232–245. Retrieved from https://doi.org/10.1037/0021-9010.71.2.232.

Byrnes, J. 2016. The value of physician leaders. Journal of Healthcare Management. 61 (4), 251–255.

Cahill, M., Robinson, K., Pettigrew, J., Galvin, R., and Stanley, M. 2018. Qualitative synthesis: a guide to conducting a meta-ethnography. British Journal of Occupational Therapy. 81 (3), 129–137. Retrieved from https://doi.org/10.1177/0308022617745016.

Caulkin, S. 2015. Have we created an unachievable myth of leadership? Financial Times. Retrieved from https://www.ft.com/content/07ed132e-8fa3-11e5-8be4-3506bf20cc2b.

Chadi, N. 2009. Medical leadership: doctors at the helm of change. McGill Journal of Medicine. 12 (1), 52–57.

Cheng, E. W. L., and Hampson, I. 2008. Transfer of training: a review and new insights. International Journal of Management Reviews. 10 (4), 327–341.

Cherry, R., Davis, D., and Thorndyke, L. 2010. Transforming culture through physician leadership. Physician Executive. 36 (3), 38–44.

Collins, D. B., and Holton III, E. E. 2004. The effectiveness of managerial leadership development programs: a meta-analysis of studies from 1982 to 2001. Human Resource Development Quarterly. 15 (2), 217–248.

Cook, D. A., and West, C. P. 2012. Conducting systematic reviews in medical education: a stepwise approach. Medical Education. 46, 943–952.

Dale, E. 1969. Audio-visual methods in teaching (3rd ed.). New York: The Dryden Press.

Daly, J., Jackson, D., Mannix, J., Davidson, P. M., and Hutchinson, M. 2014. The importance of clinical leadership in the hospital setting. Journal of Healthcare Leadership. 6, 75–83. Retrieved from https://doi.org/10.2147/JHL.S46161.

Dannels, S. A., Yamagata, H., McDade, S. A., Chuang, Y. C., Gleason, K. A., McLaughlin, J. M., … Morahan, P. S. 2008. Evaluating a leadership program: a comparative, longitudinal study to assess the impact of the Executive Leadership in Academic Medicine (ELAM) program for women. Academic Medicine. 83 (5), 488–495.

Dauphinee, W. D., and Wood-Dauphinee, S. 2004. The need for evidence in medical education: the development of best evidence medical education as an opportunity to inform, guide, and sustain medical education research. Academic Medicine: Journal of the Association of American Medical Colleges. 79 (10), 925–930.

Day, C. S., Tabrizi, S., Kramer, J., Yule, A., and Ahn, B. 2010. Effectiveness of the AAOS Leadership Fellows Program for orthopaedic surgeons. Journal of Bone and Joint Surgery of America. 92 (16), 2700–2708.

Day, D. V. 2001. Leadership development: a review in context. The Leadership Quarterly. 11 (4), 581–613.

Day, D. V., and Dragoni, L. 2015. Leadership development: an outcome-oriented review based on time and levels of analyses. In F. P. Morgeson (Ed.), Annual review of organizational psychology and organizational behaviour, vol. 2. 133–156. Palo Alto, CA: Annual Reviews.

Day, D. V., Fleenor, J. W., Atwater, L. E., Sturm, R. E., and McKee, R. A. 2014. Advances in leader and leadership development: a review of 25 years of research and theory. The Leadership Quarterly. 25 (1), 63–82.

Day, D. V., and O'Connor, P. M. 2003. Leadership development: understanding the process. In S. E. Murphy and P. Riggio (Eds.), The future of leadership development. 11–28. Mahwah, NJ: Lawrence Erlbaum.

Day, D. V., and Sin, H.-P. 2011. Longitudinal tests of an integrative model of leader development: charting and understanding developmental trajectories. Leadership Quarterly. 22 (3), 545–560. Retrieved from https://doi.org/10.1016/j.leaqua.2011.04.011.

Denis, J.-L., and van Gestel, N. 2016. Medical doctors in healthcare leadership: theoretical and practical challenges. BMC Health Services Research. 16, 158. Retrieved from https://doi.org/10.1186/s12913-016-1392-8.

DeNisi, A. S., and Kluger, A. N. 2000. Feedback effectiveness: can 360-degree appraisals be improved? The Academy of Management Executive. 14 (1), 129–139.

Dickey, C., Dismukes, R., and Topor, D. 2014. Creating Opportunities for Organizational Leadership (COOL): creating a culture and curriculum that fosters psychiatric

leadership development and quality improvement. Academic Psychiatry. 38 (3), 383-387.

Dickson, G., and Van Aerde, J. 2018. Enabling physicians to lead: Canada's LEADS framework. Leadership in Health Services. 31 (2), 183–194. Retrieved from https://doi.org/10.1108/LHS-12-2017-0077.

Dietz, J., Antonakis, J., Hoffrage, U., Krings, F., Marewski, J. N., and Zehnder, C. 2014. Teaching Evidence-Based Management with a focus on producing local evidence. Academy of Management Learning and Education. 13 (3), 397–414. Retrieved from https://doi.org/10.5465/amle.2013.0197.

Dine, C. J., Kahn, J. M., Abella, B. S., and Shea, J. A. 2011. Key elements of clinical physician leadership at an academic medical center. Journal of Graduate Medical Education. 3 (1), 31–36.

Edler, A., Adamshick, M., Fanning, R., and Piro, N. 2010. Leadership lessons from military education for postgraduate medical curricular improvement. The Clinical Teacher. 7 (1), 26–31. Retrieved from https://doi.org/10.1111/j.1743-498X.2009.00336.x

Edmonstone, J. 2009. Evaluating clinical leadership: a case study. Leadership in Health Services. 22 (3), 210–224.

Edmonstone, J. 2011. The development of strategic clinical leaders in the National Health Service in Scotland. Leadership in Health Services. 24 (4), 337–353.

Edmonstone, J. 2013. Healthcare leadership: learning from evaluation. Leadership in Health Services. 26 (2), 148–158.

Falcone, R. E., and Santiani, B. 2008. Physician as hospital Chief Executive Officer. Vascular and Endovascular Surgery. 42 (1), 88–94.

Fernandez, C. S. P., Noble, C. C., Jensen, E. T., and Chapin, J. 2016. Improving leadership skills in physicians: a 6-month retrospective study. Journal of Leadership Studies. 9 (4), 6–19.

Frich, J. C., Brewster, A. L., Cherlin, E. J., and Bradley, E. H. 2015. Leadership development programs for physicians: a systematic review. Journal of General Internal Medicine. 30 (5), 656–74.

Garman, Andrew N., McAlearney, A. S., Harrison, M. I., Song, P. H., and McHugh, M. 2011. High-performance work systems in health care management, Part 1: development of an evidence-informed model. Health Care Management Review. 36 (3), 201–213. Retrieved from https://doi.org/10.1097/HMR.0b013e318201d1bf.

Geerts, J. M. 2018. Optimal leadership development for professionals. Unpublished doctoral dissertation. University of Cambridge, Cambridge, UK.

Geerts, J. M. 2019. Leadership defined for our generation and the next. Canadian Journal of Physician Leadership. 5 (4), 207–214.

Gilpin-Jackson, Y., and Bushe, G. R. 2007. Leadership development training transfer: a case study of post-training determinants. Journal of Management Development. 26 (10), 980–1004.

Goodall, A. H. 2011. Physician-leaders and hospital performance: is there an association? Social Science & Medicine. 73 (4). 535–539.

Hamlin, B. 2010. Evidence-based leadership and management development. In J. Gold, R. Thorpe, and A. Mumford (Eds.), Gower handbook of leadership and management development (5th ed). 197–220. Surrey, England: Gower.

Hannah, S. T., Uhl-Bien, M., Avolio, B. J., and Cavarretta, F. L. 2009. A framework for examining leadership in extreme contexts. The Leadership Quarterly. 20 (6), 897–919. Retrieved from https://doi.org/10.1016/j.leaqua.2009.09.006.

Hannum, K. M., and Bartholomew, C. S. 2010. Introduction to special issue on leadership development evaluation. The Leadership Quarterly. 21 (4), 581–582.

Hartley, J., and Hinksman, B. 2003. Leadership development: a systematic literature review. A report for the NHS Leadership Centre. Warwick, UK: Warwick Institute of Governance and Public Management. 1–77.

Hawe, P., and Shiell, A. 2000. Social capital and health promotion: a review. Social Science & Medicine. 51 (6), 871–885. Retrieved from https://doi.org/10.1016/S0277-9536(00)00067-8.

Hemmer, P. R., Karon, B. S., Hernandez, J. S., Cuthbert, C., Fidler, M. E., and Tazelaar, H. D. 2007. Leadership and management training for residents and fellows: a curriculum for future medical directors. Archives of Pathology and Laboratory Medicine. 131, 610–614.

Husebø, S. E., and Akerjordet, K. 2016. Quantitative systematic review of multi professional teamwork and leadership training to optimize patient outcomes in acute hospital settings. Journal of Advanced Nursing. 72 (12), 2980–3000. Retrieved from https://doi.org/10.1111/jan.13035.

Ireri, S., Walshe, K., Benson, L., and Mwanthi, M. 2011. A qualitative and quantitative study of medical leadership and management: Experiences, competencies, and development needs of doctor managers in the United Kingdom. Journal of Management and Marketing in Healthcare. 4 (1), 16–29.

Kellerman, B. 2012. The end of leadership. New York: Harper Collins.

Kellerman, B. 2018. Professionalizing leadership. New York, NY: Oxford University Press.

Kirkpatrick, D., and Kirkpatrick, J. 2006. Evaluating training programs (3rd ed.). San Francisco: Berrett Koehler Publishers.

Klimoski, R., and Amos, B. 2012. Practicing evidence-based education in leadership development. Academy of Management Learning and Education. 11 (4), 685-702. Retrieved from https://doi.org/10.5465/amle.2012.0018.

Knowles, M. S. 1984. Andragogy in action. San Francisco: Jossey-Bass.

Kolb, D. A. 1984. Experiential learning: experience as the source of learning and development. Englewood Cliffs, N.J.: Prentice Hall.

Komives, S. R., Nance, L., and McMahon, T. R. 1998. Exploring leadership: for college students who want to make a difference. San Francisco: Jossey-Bass.

Korschun, H. W., Redding, D., Teal, G. L., and Johns, M. M. E. 2007. Realizing the vision of leadership development in an academic health center: The Woodruff Leadership Academy. Academic Medicine. 82 (3), 264–271.

Kuo, A. K., Thyne, S. M., Chen, H. C., West, D. C., and Kamei, R. K. 2010. An innovative residency program designed to develop leaders to improve the health of children. Academic Medicine. 85 (10), 1603–1608.

Kwamie, A., van Dijk, H, and Agyepong, I. A. 2014. Advancing the application of systems thinking in health: Realist evaluation of the Leadership Development Programme for district manager decision-making in Ghana. Health Research Policy and Systems. 12 (29), 1–12.

Lacerenza, C. N., Reyes, D. L., Marlow, S. L., Joseph, D. L., and Salas, E. 2017. Leadership training design, delivery, and implementation: a meta-analysis. Journal of Applied Psychology. 102 (12), 1686–1718. Retrieved from https://doi.org/10.1037/apl0000241.

Latham, G. P., and Locke, E. A. 1983. Goal setting—a motivational technique that works. In J. R. Hackman, E. E. Lawlor, and L. W. Porter (Eds.), Perspectives on Behavior in Organizations. 296–304. New York: McGraw Hill.

Lazear, E. P., Shaw, K. L., & Stanton, C. T. 2015. The value of bosses. Journal of Labor Economics. 33 (4), 823–861. Retrieved from https://doi.org/10.1086/681097.

Lee, T. H., and Hall, K. W. 2010. Turning doctors into leaders. Harvard Business Review. 88 (4), 50–58.

Leskiw, S., and Singh, P. 2007. Leadership development: learning from best practices. Leadership and Organization Development Journal. 28 (5), 444–464. Retrieved from https://doi.org/10.1108/01437730710761742.

Lurie, S. J. 2003. Raising the passing grade for studies of medical education. JAMA. 290 (9), 1210–1212. Retrieved from https://doi.org/10.1001/jama.290.9.1210.

MacPhail, A., Young, C., and Ibrahim, J. E. 2015. Workplace-based clinical leadership training increases willingness to lead. Leadership in Health Services. 28 (2), 100–118.

Malling, B., Mortensen, L., Bonderup, T., Scherpbier, A., and Ringstead, C. 2009. Combining a leadership course and multisource feedback has no effect on leadership skills of leaders in postgraduate medical education. An intervention study with a control group. BMC Medical Education. 9 (72), 1–7.

McAlearney, A. S. 2010. Executive leadership development in U.S. health systems. Journal of Healthcare Management. 55 (3), 206–224.

McAlearney, A. S., Fisher, D., Heiser, K., Robbins, D., and Kelleher, K. 2005. Developing effective physician leaders: Changing cultures and transforming organizations. Hospital Topics. 83 (2), 11–18.

McCauley, C. D. 2008. Leader development: a review of research. Center for Creative Leadership.

McDonald, D. 2017. The golden passport: Harvard Business School, the limits of capitalism, and the moral failure of the MBA elite. New York: HarperCollins.

McKimm, J., Spurgeon, P., Needham, G., and O'Sullivan, H. 2011. Clinical leadership: the challenge for senior leaders. British Journal of Hospital Medicine. 72 (9), 525–528. Retrieved from https://doi.org/10.12968/hmed.2011.72.9.525.

Miller, D. L., Umble, K. E., Frederick, S. L., and Dinkin, D. R. 2007. Linking learning methods to outcomes in public health leadership development. Leadership in Health Services. 20 (2), 97–123.

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Journal of Clinical Epidemiology. 62 (10), 1006–1012.

Mountford, J., and Webb, C. 2009. When clinicians lead. McKinsey Quarterly. (February) 1–8.

Murdock, J., and Brammer, C. 2011. A successful model of leadership development for community practice physicians. Physician Executive Journal. 37 (2), 52–56.

Nabi, G., Liñán, F., Fayolle, A., Krueger, N., and Walmsley, A. 2017. The impact of entrepreneurship education in higher education: a systematic review and research agenda. Academy of Management Learning and Education. 16 (2), 277–299. Retrieved from https://doi.org/10.5465/amle.2015.0026.

Nakanjako, D., Namagala, E., Semeere, A., Kigozi, J., Sempa, J., Ddamulira, J. B., … Sewankambo, N. 2015. Global health leadership training in resource-limited settings: a collaborative approach by academic institutions and local health care programs in Uganda. Human Resources for Health. 13 (1), 87.

Nowacki, A. S., Barss, C., Spencer, S. M., Christensen, T., Fralicx, R., and Stoller, J. K. 2016. Emotional intelligence and physician leadership potential: a longitudinal study supporting a link. Journal of Health Administration. 33 (1), 23–41.

Patel, N., Brennan, P. J., Metlay, J., Bellini, L., Shannon, R. P., and Myers, J. S. 2015. Building the pipeline: the creation of a residency training pathway for future physician leaders in health care quality. Journal of the Association of American Medical Colleges. 90 (2), 185–190. Retrieved from https://doi.org/10.1097/ACM.0000000000000546.

Perry, J., Mobley, F., and Brubaker, M. 2017. Most doctors have little or no management training, and that's a problem. Harvard Business Review. Retrieved from https://hbr.org/2017/12/most-doctors-have-little-or-no-management-training-and-thats-a-problem.

Petticrew, M., and Roberts, H. 2006. Systematic reviews in the Social Sciences: a practical guide (Vol. 6). Oxford: Blackwell.

Pfeffer, J. 2015. Leadership BS. New York: HarperCollins.

Pfeffer, J. 2016. Getting beyond the BS of leadership literature. McKinsey Quarterly, 1–7.

Phillips, P. P., Phillips, J. J., and Ray, R. 2015. Measuring the success of leadership development. Alexandria, Virginia: ATD Press.

Powell, S. K., and Yalcin, S. 2010. Managerial training effectiveness. Personnel Review. 39 (2), 227–241.

Pradarelli, J. C., Jaffe, G. A., Lemak, C. H., Mulholland, M. W., and Dimick, J. B. 2016. Designing a leadership development program for surgeons. Journal of Surgical Research. 200 (1), 53–58.

QuintilesIMS Institute. 2017. Global Oncology Trends 2017. Parsippany, NJ: QuintilesIMS Institute.

Reed, D. A., Beckman, T. J., Wright, S. M., Levine, R. B., Kern, D. E., and Cook, D. A. 2008. Predictive validity evidence for medical education research study quality instrument scores: quality of submissions to JGIM's Medical Education Special Issue. Journal of General Internal Medicine. 23 (7), 903–907. Retrieved from https://doi.org/10.1007/s11606-008-0664-3.

Reed, D., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E., and Wright, S. M. 2007. Association between funding and quality of published medical education research. Journal of the American Medical Association. 298 (9), 1002–1009.

Richardson, W., Wilson, M., Nishikawa, J., and Hayward, R. 1995. The well-built clinical question: a key to evidence-based decisions. ACP Journal Club. 123 (3), A12–3.

Rosenman, E. D., Shandro, J. R., Ilgen, J. S., Harper, A. L., and Fernandez, R. 2014. Leadership training in health care action teams: a systematic review. Academic Medicine. 89 (9), 1295–1306.

Rousseau, D. M. 2006. Is there such a thing as 'evidence-based management'? Academy of Management Review. 31 (2), 256–269.

Rousseau, D. M., and McCarthy, S. 2007. Educating managers from an evidence-based perspective. Academy of Management Learning and Education. 6 (1), 84–101. Retrieved from https://doi.org/10.5465/AMLE.2007.24401705.

Rynes, S. L., Rousseau, D. M., and Barends, E. 2014. From the guest editors: change the world: Teach evidence-based practice! Academy of Management Learning and Education. 13 (3), 305–321. Retrieved from https://doi.org/10.5465/amle.2014.0203.

Salas, E., Tannenbaum, S. I., Kraiger, K., and Smith-Jentsch, K. A. 2012. The science of training and development in organizations: what matters in practice. Psychological Science in the Public Interest. 13, 74–101.

Sanfey, H., Harris, I., Pollart, S., and Schwartz, A. 2011. Evaluation of the University of Virginia Leadership in Academic Medicine Program. Teaching and Learning in Medicine. 23 (4), 347–358.

Satiani, B., Sena, J., Ruberg, R., and Ellison, E. C. 2014. Talent management and physician leadership training is essential for preparing tomorrow's physician leaders. Journal of Vascular Surgery. 59 (2), 542–546.

Sellgren, S. F., Ekvall, G., and Tomson, G. 2008. Leadership behaviour of nurse managers in relation to job satisfaction and work climate. Journal of Nursing Management. 16 (5), 578–587.

Shah, P., Cross, V., and Sii, F. 2013. Sailing a safe ship: improving patient safety by enhancing the leadership skills of new consultant specialist surgeons. Journal of Continuing Education in the Health Professions. 33 (3), 190–200.

Solansky, S. T. 2010. The evaluation of two key leadership development program components: leadership skills assessment and leadership mentoring. The Leadership Quarterly. 21, 675–681. Retrieved from https://doi.org/10.1016/j.leaqua.2010.06.009.

Spurgeon, Peter, Long, P., Clark, J., and Daly, F. 2015. Do we need medical leadership or medical engagement? Leadership in Health Services. 28 (3), 173–184. Retrieved from https://doi.org/10.1108/LHS-03-20140029.

Spurgeon, P., Mazelan, P., and Barwell, F. 2011. Medical engagement: a crucial underpinning to organizational performance. Health Services Management Research. 24 (3), 114–120.

Steinert, Y., Naismith, L., and Mann, K. 2012. Faculty development initiatives designed to promote leadership in medical education. a BEME systematic review: BEME guide no. 19. The International Journal of Medical Technology. 34 (6), 483–503. Retrieved from https://doi.org/doi: 10.3109/0142159X.2012.680937.

Sterne, J. A., Higgins, J. P., and Reeves, B. C. 2014. A Cochrane risk of bias assessment tool for non-randomised studies of interventions (Version 1.0.0). Cochrane Colloquium. Retrieved from http://www.riskofbias.info.

Stoller, J. K. 2008. Developing physician-leaders: Key competencies and available programs. The Journal of Health Administration Education. 25 (4), 307–328.

Stoller, J. K. 2009. Developing physician-leaders: a call to action. Journal of General Internal Medicine. 24 (7), 876–878.

Straus, S. E., Soobiah, C., and Levinson, W. 2013. The impact of leadership training programs on physicians in Academic Medical Centers: a systematic review. Academic Medicine. 88 (5), 1–15.

Tasi, M. C., Keswani, A., and Bozic, K. J. 2017. Does physician leadership affect hospital quality, operational efficiency, and financial performance? Health Care Management Review. Retrieved from https://doi.org/10.1097/HMR.0000000000000173.

Taylor, B. 2010. Effective medical leadership. Toronto: University of Toronto Press.

Taylor, P. J., Russ-Eft, D. F., and Taylor, H. 2009. Transfer of management training from alternative perspectives. Journal of Applied Psychology. 94, 104–121.

Ten Have, E. C. M., Nap, R. E., and Tulleken, J. E. 2013. Quality improvement of interdisciplinary rounds by leadership training based on essential quality indicators of the Interdisciplinary Rounds Assessment Scale. Intensive Care Medicine. 39 (10), 1800–1807. Retrieved from https://doi.org/10.1007/s00134-013-3002-0.

Terpstra, D. E. 1981. Relationship between methodological rigor and reported outcomes in organization development evaluation research. Journal of Applied Psychology. 66 (5), 541–543. Retrieved from https://doi.org/10.1037/0021-9010.66.5.541.

The Cochrane Collaboration. 2011. Cochrane Handbook for Systematic Reviews of Interventions (Version 5.1.0) (J. P. Higgins and S. Green, Eds.). Retrieved from http://handbook-5-1.cochrane.org.

Tracey, J. B., and Tews, M. J. 2005. Construct validity of a general training climate scale. Organizational Research Methods. 8 (4), 353–374.

Vimr, M., and Dickens, P. 2013. Building physician capacity for transformational leadership - Revisited. Health care Management Forum. 26 (1), 16–19.

Watkins, K. E., Lysø, I. H., and deMarrais, K. 2011. Evaluating executive leadership programmes: a theory of change approach. Advances in Developing Human Resources. 13 (2), 208– 239.

Weaver, S. J., Dy, S. M., and Rosen, M. A. 2014. Team-training in health care: a narrative synthesis of the literature. British Medical Journal Quality and Safety. 23, 359–372.

Wells, G., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., and Tugwell, P. 2016. The Newcastle-Ottawa Scale (NOS) for assessing the quality of non-randomised studies in meta-analyses. Retrieved from http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

World Health Organization. 2018. Global Health Expenditure Database. Retrieved from http://apps.who.int/nha/database.

Zaccaro, S. J., and Horn, Z. N. J. 2003. Leadership theory and practice: fostering an effective symbiosis. The Leadership Quarterly. 14 (6), 769–806. Retrieved from https://doi.org/10.1016/j.leaqua.2003.09.009.

Table and figure captions

Table 1: PICOS framework

Table 2: Modified version of Kirkpatrick's (2006) model of development program outcomes

Table 3: Coding for the 25 included studies (1/2)

Table 4: Coding for the 25 included studies (2/2)

Table 5: Medical Education Research Study Quality Instrument (MERSQI)

Table 6: Methodological quality groupings

Table 7: MERSQI applied to the 25 included studies

Table 8: Key factors associated with effective leadership development

Figure 1: Literature search process

Table 1

PICOS framework.

PICOS (Participants, Interventions, Comparison, Outcomes, Study Design)	
P	Physicians
I	Leadership development programs or interventions
C	When possible, compare outcomes to those of physicians who did not participate in leadership development
O	Impact on outcomes at the individual, organizational, and benefit to patients levels
S	Qualitative, quantitative, and mixed methods designs were included

Table 2

Modified version of Kirkpatrick's (2006) model of development program outcomes.

Level	Details
1	Participant satisfaction with the program
2a	Changes in participants' attitudes or perspectives
2b	Changes in participants' knowledge and skills
3a	Self-reported changes in participants' behaviour
3b	Objective indicators of changes in participants' behaviour
4a	Organizational impact
4b	Benefit to patients (subjective and objective)

Table 3

Coding for the 25 included studies (1/2).

First author and publication year	Design		Samples					Intervention					Evaluation		Outcomes	
	Methodology	Data collection methods	No. participants	No. control	Female, male	Physicians only?	Level of seniority	Location	Faculty (internal, external, both)	Sites	In-house vs external	Length of intervention	Raters	When	Kirkpatrick levels	Reported outcomes and benefits
Hemmer (2007)	Action research/case study	Questionnaire, document analysis	16	None	NR	No	Residents (internal)	USA	Both	Single	In-house	1 year	Self, faculty	Baseline, post, post-post	1, 2b	1) PPE's; 2b) Increased knowledge tests results
Korschun (2007)	Action research/case study	Questionnaire, interviews	70	None	NR	No	Mixed	USA	Both	Single	In-house	5 months	Self	Post-post	1, 2a, 2b, 3a, 3b, 4a, 4b	1) PPE's; 2a) Increased aspirations to lead, increased engagement, increased commitment; 2b) Increased knowledge and interpersonal and teamwork skills; 3a) Increased leadership effectiveness, networking benefits, have taken on more responsibility; 3b) Retention, promotions, have taken on a leadership role, increased committee involvement; 4a) Having launched a new initiative; 4b) Having implemented action learning projects
Miller (2007)	Case study	Questionnaire	210	None	NR	No	Senior leaders	USA	Unclear	Single	External	1 year	Self	Post-post, Retrospective post	1, 2a, 2b, 3a, 4a	1) PPE's; 2a) Increased confidence and self-awareness; 2b) Increased leadership knowledge and skills; 3a) Increased leadership behaviours; 4a) General organizational benefits, developing and strengthening their organizations' collaborative relationships, and developing or implementing a new program
Dannels (2008)	Experiment	Questionnaire	78	468	78, 0	No	Senior faculty	USA	Unclear	Single	External	Unclear	Self	Baseline, post-post	2a, 2b, 3b	2a) Increased aspirations to lead; 2b) Increased knowledge; 3b) Improved MSF pre and post, promotions, higher participation in further leadership development following programs
Bergman (2009)	Multiple case study	Questionnaire, focus group interviews	109	None	95, 14	No	First-line managers	Sweden	External	Single	In-house	1 week	Self	Pre, post-post	1, 2a, 2b, 3a	1) PPE's; 2a) Increased confidence and self-awareness; 2b) Increased knowledge and communication skills; 3a) Increased leadership behaviours
Edmonstone (2009)	Case study	Questionnaire, interviews, document analysis	218	None	NR	No	Senior leaders	England	Both	Four	External	1 year	Self, faculty	Post-post	1, 2a, 2b, 3a, 4a	1) PPE; 2a) Greater appreciation of others' perspectives, increased engagement, enhanced common identity, increased confidence; 2b) Increased leadership skills; 3a) Increased leadership behaviours, networking benefits, developed PDP's; 4a) Having launched a new initiative
Malling (2009)	Quasi-experiment	Questionnaire, statistical analysis	20	28	NR	Yes	Consultant (education)	Denmark	Unclear	Single	External	6 months	Self, subordinate, peer	Pre, post-post	1, 2b, 3a	1) PPE; 2b) Increased knowledge; 3a) Increased leadership behaviours
Murdock (2009)	Case study	Questionnaire	100	None	NR	Yes	Physicians unspecified	USA	Both	Single	External	20 weeks	Self	Pre, post	1, 2a, 2b, 3a, 3b	1) PPE; 2a) Increased aspirations to lead; 2b) Increased leadership skills; 3a) Increased leadership behaviours; 3b) Having taken on a leadership role
Cherry (2010)	Case study	Unclear	141	None	NR	Yes	Junior physicians	USA	Internal	Single	In-house	9 months	Peer, faculty, superior	N/A	1, 2b, 3b	1) PPE's; 2b) Increased leadership skills; 3b) Research publications
Day (2010)	Case study	Questionnaire, document analysis	100	73	NR	Yes	Surgeons unspecified (Orthopaedics)	USA	Unclear	Single	External	1 year	Self	Pre, baseline	2a, 2b, 3a, 3b	2a) Increased confidence, 2b) Increased knowledge and skills, 3a) Positive impact on their careers; 3b) Having taken on a leadership role, increased committee involvement, research publications, increased academic rank, hospital administrative rank (chair or chief)
Kuo (2010)	Case study	Questionnaire	15	None	NR	Yes	Residents (pediatric)	USA	Internal	Single	In-house	3 years	Self, statistics	Post, post-post	1, 2a, 2b, 3a, 3b	1) PPE's; 2a) Increased aspirations to lead; 2b) Increased leadership competence; 3a) Positive impact on their careers; 3b) Awards won, grants earned, and research publications, having taken on leadership roles, promotions
Edmonstone (2011)	Case study	Questionnaire, document analysis, interviews	125	None	NR	No	Potential senior clinical leaders	Scotland	Unclear	Single	In-house	1 year	Self, subordinate, peer, superior	Pre, post	1, 2a, 2b, 3b, 4a, 4b	1) PPE's; 2a) Increased self-awareness, increased resilience, increased engagement; 2b) Developing interpersonal and networking skills; 3b) Colleagues' feedback on behaviour changes, promotions, 4a) Policy changes, developed organizational capacity; 4b) Implementing action learning projects; Other) Having joined a mentoring network

Note: PPE = Post program Evaluation. MSF = Multisource Feedback. PDP = Personal Development Plan.

Table 4

Coding for the 25 included studies (2/2).

First author and publication year	Design							Intervention						Evaluation		Outcomes	
	Methodology	Data collection methods	No. participants	No. control	Female, male	Physicians only?	Level of seniority	Location	Faculty (internal, external, both)	Sites	In-house vs external	Length of intervention	Raters	When	Kirkpatrick levels	Reported outcomes and benefits
Sanfey (2011)	Grounded theory	Questionnaire	142	None	50, 92	No	Mixed	USA	Both	Single	In-house	10 weeks	Self	Pre, post, post-post	1, 2a, 2b, 3a, 3b	1) PPE's; 2a) Increased aspirations to lead, increased self-awareness, increased leadership self-identity; 2b) Increased knowledge and skills; 3a) Increased leadership behaviours, networking benefits; 3b) Promotions
Bearman (2012)	Case study	Questionnaire, MSF	12	None	NR	Yes	Surgical residents	Australia and New Zealand	Unclear	Single	In-house	2 days	Self	Post	1	1) PPE's
Shah (2013)	Case study	Video analysis, unclear	40	None	NR	Yes	Consultant (ophthalmic surgeons)	UK	External	Single	In-house	2 days	Self	Post	1, 2a, 2b, 3a	1) PPE's; 2a) Increased engagement, 2b) Increased knowledge and skills, 3a) Increased leadership behaviours
Ten Have (2013)	Quasi-experiment	Experiment, questionnaire	9	10	7, 12	Yes	Midlevel surgeons	Netherlands	Internal	Single	In-house	1 day	Peer, faculty	Pre, post-post	3b	3b) Improved MSF pre and post
Vimr (2013)	Action research/case study	Questionnaire	-	None	NR	Yes	Physicians unspecified	Canada	Unclear	Single	In-house	8 months	Self	Post	1, 2a, 3a, 4b	1) PPE's; 2a) Improved self-awareness, developed a systems view; 3a) Increased leadership behaviours; 4b) Having implemented action learning projects
Blumenthal (2014)	Action research	Questionnaire, observation	16	None	10, 6	Yes	Residents (Internal medicine)	USA	Internal	Single	In-house	1 month	Self	Post	1, 2a, 2b	1) PPE's; 2a) Increased confidence, increased self-awareness, increased awareness of different leadership styles, increased interest in further development programs; 2b) Increased knowledge and skills
Dickey (2014)	Action research/case study	Unclear	-	None	NR	Yes	Residents (psychiatry)	USA	Internal	Single	In-house	4 years	N/A	NR	1, 2b	1) Authors' perceptions of program strengths; 2b) Developed negotiation skills
MacPhail (2014)	Case study	Questionnaire, interviews	39	None	32, 7	No	Middle and senior leaders	Australia	Both	Single	In-house	9 - 10 months	Self, superior	Post-post	1, 2a, 2b, 3a, 3b	1) PPE's; 2a) Increased aspirations to lead, increased leadership capacity, plan to change their approach to patient care; 2b) Increased leadership knowledge and skills, developed ideas for improving patient care; 3a) Increased leadership behaviours; 3b) Have taken on a leadership role, retention, promotions.
Satiani (2014)	Case study	Questionnaire, unclear	-	None	NR	Yes	Physicians (early to mid career high potentials)	USA	Both	Single	In-house	18 months	Self, superior	During, post	1, 2a, 2b, 3a, 3b	1) PPE's; 2a) Increased self-awareness and confidence; 2b) Increased leadership skills, increased negotiation skills, developed interpersonal skills; 3a) Networking benefits; 3b) Supervisors' ratings of increased leadership skill levels and changes in behaviour.
Nakanjako (2015)	Case study	Document analysis	15	None	NR	No	Physicians unspecified	Uganda	Both	Single	In-house	1 year	Self	Post	2b, 3a, 3b, 4a, 4b	2b) Increased leadership skills, increased leadership capability; 3a) Have taken on more responsibility; 3b) Retention, awards won, research publications; 4a) General organizational benefits, increased organizational capacity; 4b) Having implemented action learning projects, used innovative approaches to improve healthcare delivery
Patel (2015)	Action research	Questionnaire, unclear, statistical analysis	62	None	NR	Yes	Residents	USA	Internal	Single	In-house	2 years	Self, faculty	Pre, post	1, 2a, 2b, 3b, 4b	1) PPE's; 2a) Increased confidence, increased aspirations to lead; 2b) Increased leadership knowledge; 3b) Have taken on a leadership role, 4b) Having implemented an action learning project
Fernandez (2016)	Quasi-experiment	Questionnaire	37	None	26, 11	Yes	Junior fellows	USA	Both	Single	External	3.5 days	Self	Post, post-post, retrospective pre	1, 2a, 2b, 3a, 3b, 4b	1) PPE's; 2a) Increased confidence; 2b) Increased leadership skills, increased communication and teamwork skills; 3a) Increased leadership behaviours, have taken on more responsibility, positive impact on their careers; 3b) Promotions; 4b) Self-reports of providing better healthcare to patients
Pradarelli (2016)	Case study	Interviews	21	None	2, 5	Yes	Surgeon unspecified	USA	Internal	Single	In-house	8 months	Self, peer, subordinate, superior	Baseline, post	1, 2a, 2b, 3a	1) PPE's; 2a) Increased self-awareness and confidence; 2b) Increase knowledge and skills; 3a) Positive impact on their careers

Note: PPE = Post program Evaluation. MSF = Multisource Feedback. PDP = Personal Development Plan.

Table 5

Medical Education Research Study Quality Instrument (MERSQI).

Domain	Item	Item score	Maximum domain score
1. Study design	Single group cross-sectional or single group post-test only	1	3
	Single group pre and post-test	1.5	
	Non-randomized, two-group	2	
	Randomized controlled experiment	3	
Sampling			
2. Institutions	One	0.5	3
	Two	1	
	>Two	1.5	
3. Response rate	<50% or Not reported	0.5	
	50 – 74%	1	
	≥75%	1.5	
4. Type of data	Assessment by study subject	1	3
	Objective measurement	3	
Validity of evaluation instruments' scores			
5. Internal structure	Not reported	0	3
	Reported	1	
6. Content	Not reported	0	
	Reported	1	
7. Relationships to other variables	Not reported	0	
	Reported	1	
Data analysis			
8. Comprehensiveness[1]	Less comprehensive data analysis given the study design or incomplete data sets	0	3
	Comprehensive data analysis given the study design and provided complete data sets	1	
9. Sophistication	Descriptive analysis only	1	
	Beyond descriptive analysis	2	
10. Outcomes	Satisfaction, attitudes, perceptions, opinions, general facts (Level 1 and 2a)	1	3
	Knowledge, skills (Level 2b)	1.5	
	Behaviours (Level 3a and 3b)	2	
	Benefit to patients outcome (Level 4b)	3	
Total			18

[1]"Appropriateness" in the original instrument, modified to "comprehensive" avoid a strong value judgement

Table 6

Methodological quality groupings.

Evidence	Characteristics	MERSQI scores	k	Studies and MERSQI score
Gold	Correlated objective outcome data Pre and post-post measures Control group	14 - 18	2	Dannels 2008 (14.5) Ten Have 2013 (15)
Silver	Correlated objective outcome data Details or study elements left out	12 - 13.5	4	Malling 2009 (12) Day 2010 (13.5) Kuo 2010 (12) Patel 2015 (12.5)
Bronze	Based largely on participants' or authors' perceptions Or elements of the study left out	<12	19	Hemmer 2007 (7.5) Korschun 2007 (10) Miller 2007 (11.5) Bergman 2009 (10) Edmonstone 2009 (7) Murdock 2009 (8) Cherry 2010 (7) Edmonstone 2011 (11) Sanfey 2011 (11) Bearman 2012 (7.5) Shah 2013 (5) Vimr 2013 (9.5) Dickey 2014 (4.5) MacPhail 2014 (11.5) Satiani (2014) 7.5 Blumenthal 2015 (8.5) Nakanjako 2015 (11) Fernandez 2016 (10.5) Pradarelli 2016 (9.5)

Note. k = The number of included studies (K = 25)

Table 7

MERSI applied to the 25 included studies.

Author	Study design	Sampling		Validity of Evaluation Instruments				Data Analysis			
		No. of Institutions	Response rate	Type of data	Internal structure	Content	Relationships to other variables	Comprehensiveness of analysis[1]	Sophistication of analysis	Outcome level	Total score
Hemmer (2007)	SGPP	1	NR	Objective	NR	NR	NR	Less comprehensive	Descriptive only	2b	7.5
Korschun (2007)	SGCS/P	1	79%	Self-reported	NR	Reported	NR	Less comprehensive	Beyond descriptive	4b	10
Miller (2007)	SGCS/P	1	66%	Self-reported	Reported	Reported	Reported	Comprehensive	Beyond descriptive	3	11.5
Dannels (2008)	NR2GP	1	71%	Objective	Reported	Reported	Reported	Comprehensive	Beyond descriptive	3	14.5
Bergman (2009)	SGPP	1	74%	Self-reported	Reported	Reported	Reported	Less comprehensive	Descriptive only	3	10
Edmonstone (2009)	SGCS/P	1	57%	Self-reported	NR	NR	NR	Less comprehensive	Descriptive only	3	7
Malling (2009)	NR2GP	1	77%	Objective	Reported	NR	NR	Less comprehensive	Beyond descriptive	3	12
Murdock (2009)	SGPP	1	NR	Objective	NR	NR	NR	Less comprehensive	Descriptive only	3	8
Cherry (2010)	SGCS/P	1	NR	Objective	NR	NR	NR	Less comprehensive	Descriptive only	2b	7
Day (2010)	NR2GP	1	53%	Objective	Reported	Reported	NR	Comprehensive	Beyond descriptive	3	13.5
Kuo (2010)	SGCS/P	1	94%	Objective	Reported	Reported	NR	Less comprehensive	Beyond descriptive	3	12
Edmonstone (2011)	SGPP	1	NR	Objective	Reported	NR	NR	Less comprehensive	Beyond descriptive	4b	11
Sanfey (2011)	SGPP	1	50%	Objective	NR	NR	Reported	Less comprehensive	Beyond descriptive	3	11
Bearman (2012)	SGCS/P	1	92%	Self-reported	NR	Reported	NR	Less comprehensive	Descriptive only	2b	7.5
Shah (2013)	SGCS/P	1	NR	Self-reported	NR	NR	NR	Less comprehensive	Beyond descriptive	2b	6
Ten Have (2013)	NR2GP	4	100%	Objective	Reported	Reported	NR	Comprehensive	Beyond descriptive	3	15
Vimr (2013)	SGCS/P	1	NR	Objective	Reported	NR	NR	Less comprehensive	Descriptive only	4b	9.5
Blumenthal (2014)	SGCS/P	1	100%	Self-reported	NR	Reported	NR	Comprehensive	Descriptive only	2b	8.5
Dickey (2014)	SGCS/P	1	NR	Self-reported	NR	NR	NR	Less comprehensive	Descriptive only	1/2a	5
MacPhail (2014)	SGCS/P	1	70%	Objective	NR	Reported	Reported	Less comprehensive	Beyond descriptive	3	11.5
Satiani (2014)	SGCS/P	1	NR	Objective	NR	NR	NR	Less comprehensive	Descriptive only	3	7.5
Nakanjako (2015)	SGCS/P	1	100%	Self-reported	NR	Reported	Reported	Less comprehensive	Beyond descriptive	4b	11
Patel (2015)	SGPP	1	77%	Objective	NR	Reported	Reported	Less comprehensive	Descriptive only	4b	12.5
Fernandez (2016)	SGCS/P	1	60%	Objective	NR	NR	NR	Less comprehensive	Beyond descriptive	4b	10.5
Pradarelli (2016)	SGPP	1	100%	Self-reported	NR	Reported	Reported	Less comprehensive	Descriptive only	3	9.5

Note. RCT = Randomized controlled trial. NR2GP = Non-randomized, two groups. SGPP = Single group, pre and post-test. SGCS/P = Single group, cross-sectional or posttest only. NR = Not reported.

[1]"Appropriateness" in the original instrument

Table 8

Key factors associated with effective leadership development.

Factors associated with effective leadership development			Representation of factors in the included studies	
Major category	**Factor**	**Effective practice**	**All included studies (K = 25)**	**Gold and silver standard studies (k = 6)**
Design	Clearly described goals for the program	Yes	Yes (17), No (8)	Yes (4), No (2)
	Used an outcomes-based approach	Yes	Yes (12), No (13)	Yes (3), No (3)
	Incorporated the Principles of Adult Learning	Yes	Yes (3), No (22)	Yes (1), No (5)
	Conducted a pre-program needs analysis	Yes	Yes (10), No (15)	Yes (2), No (4)
	Used a capability framework	Yes	Yes (8), No (17)	Yes (2), No (4)
Sample	Control group	Yes	Yes (4), No (21)	Yes (3), No (1)
Program	Developmental activities	Workshops, simulations, action learning, coaching, mentoring	Workshops (14), reading assignments (11) small group discussion (11)	Simulation (3), mentoring (3) facilitator feedback (3)
Outcomes	Kirkpatrick level	2a, 2b, 3b, 4a, 4b	Level (n): 1 (21), 2a (39), 2b (41), 3a (23), 3b (29), 4a (9), 4b (9), (2a, 2b, 3b, 4a, and 4b) (3)	Level (n): 1 (3), 2a (5), 2b (7), 3a (3), 3b (10), 4a (0), 4b (2), (2a, 2b, 3b, 4a, and 4b) (0)
	Reported outcomes	Objective outcomes at the individual "(2a - 3b), organizational (4a), and benefit to patients (4b) levels	Confidence (9) (Level 2a), Increased skills (13) (Level 2b), Knowledge (12) (Level 2b), Behaviours (9) (Level 3a)	Aspirations to lead (3) (Level 2a), Increased knowledge (4) (Level 2b), Having taken on a leadership role (3) (Level 3b),
Evaluation	Focus of evaluation (participants, program, or both)	Both	Participants only (5), program only (5) Both (15)	Participants only (3), both (3)
	Type of data collected (quantitative, qualitative, or both)	Both	Quantitative only (5), qualitative only (8) Both (12)	Quantitative only (4), both (2)
	Type of data collected (objective, subjective)	Objective	Objective (15), subjective (10)	Objective (6)
	Raters	Multiple	Self (22), Peer (5), Facilitator (5), Supervisor (5), Single rater (13), Multiple raters (11)	Self (5), Peer (2), Facilitator (2), Single rater (2), Multiple raters (4)
	When data was collected	Pre, post, and post-post	Pre or baseline (11), Post (14), Post-post (13), (Pre, post, and post-post) (2)	Pre or baseline (5), Post-post (5), Post (2), (Pre, post, and post-post) (0)

Note: K = The number of included studies

Fig. 1. Literature search process.