The Effectiveness of Medical Simulation in Teaching Medical Students Critical Care Medicine: A Systematic Review and Meta-analysis

Matt Beal, John Kinnear, Caroline Rachael Anderson, Thomas D Martin, Rachel Wamboldt, Lee Hooper

Affiliations:

1. Norwich Medical School, University of East Anglia, Norwich, United Kingdom NR4 7TJ

Dr. Matt D. Beal, MClinEd
Dr. Caroline R. Anderson, MRes
Dr. Thomas D. Martin, MRes
Rachel Wamboldt
Dr. Lee Hooper, PhD

2. Postgraduate Medical Institute, Anglia Ruskin University, Cambridge & Chelmsford, Bishop Hall Lane, Chelmsford, Essex, CM1 1SQ

Prof. John Kinnear, DMedEd

Correspondence to: matt.beal@doctors.org.uk

SUMMARY

We aimed to assess effectiveness of simulation for teaching medical students critical care medicine and to assess which simulation methods were most useful. We searched AMED, EMBASE, MEDLINE, ERIC, BEI, AEI, plus bibliographies and citations, to July 2013. Randomised controlled trials comparing effectiveness of simulation with another educational intervention, or no teaching, for teaching medical students critical care medicine were included. Assessments for inclusion, quality and data extraction were duplicated and results synthesised using meta-analysis.

We included 22 RCTs (n=1325). Fifteen studies comparing simulation with other teaching found simulation to be more effective (SMD 0.84, 95% CI 0.43 to 1.24; $p<0.001$; I^2 89%). High-fidelity simulation was more effective than low-fidelity and subgrouping supported high-fidelity simulation being more effective than other methods. Simulation improved skill acquisition (SMD 1.01, 95% CI 0.49 to 1.53) but was no better than other teaching in knowledge acquisition (SMD 0.41, 95% CI -0.09 to 0.91).

INTRODUCTION

There is no common medical school curriculum in acute and emergency care,[1] and deficiencies in knowledge are common amongst medical school graduates completing their residency,[2-4] who are often responsible for the early assessment and treatment of patients

who are acutely ill.[3] A review of training in the care of acutely ill patients found medical school training to be suboptimal and to place patients at risk.[3] With the current urgent need to relieve pressure on over-worked acute care specialities, improving the training and preparation of residents may go some way to addressing the shortage of skilled staff to treat patients safely.[5]

At some point in medical education there is a need to refine skills on live patients. However, this must be carefully balanced against the ethical obligation to provide optimal treatment whilst protecting patients from harm.[6] In critical care, this ethical dilemma is intensified as patients are often sedated, or have reduced levels of consciousness, which limits their ability to consent to participating in this kind of education. When trainees do actively participate, the opportunity to correct poor technique is limited,[7] as training is often opportunistic, with limited chance to build expertise by repeated practice. These are some reasons why "learning by doing" has become less acceptable.[8]

There is a growing body of evidence for the use of simulation based medical education,[3] which may go some way to mitigating the ethical tensions that arise from using patients as training tools for clinicians.[6] Simulation is the process of recreating characteristics of the real world,[9] allowing the trainer to carefully control the learning environment and optimise conditions for the skill being taught. This has led the General Medical Council to now recommend that medical schools should utilise simulation technology in the education of undergraduate medical students.[10]

Simulation has been shown to have a positive educational impact in a number of health professional groups,[11-17] but its effectiveness for the medical student is not as clearly defined.[18] Reviews have mainly concentrated on post medical school eduction, and reviews of simulation in medical students consist of a qualitative narrative synthesis, based upon non-systematic identification of literature.[19] The stage of professional development,[20] as well as the varying skills being practiced, may influence the effectiveness of the teaching method employed. Cognitive load theory helps to explain how a learner's prior knowledge may impact on the efficacy of simulation in medical students compared to higher level learners. Where a learning task is too complex, short term memory can rapidly become overloaded, which has the effect of inhibiting learning. [21]

Exposure to simulation during medical school is highly variable and no studies have investigated an ideal amount of exposure time.[18] Simulation is enjoyed by medical students and faculty alike,[22, 23] suggesting that simulation is effective at Kirkpatrick Level 1 (reaction to learning experience). However, its effectiveness at Kirkpatrick levels 2 (Knowledge, skill and attitude acquisition) compared to other teaching methods has been equivocal, with studies reporting no difference, positive or negative effects.[22, 24, 25] This is in contrast to simulation based medical education in other professional groups and following medical school, which demonstrates moderate to large positive effects.[16, 26, 27]

Objectives

The aim of this systematic review and meta-analysis was to assess the effectiveness of simulation for teaching medical students critical care medicine, compared to other teaching methods, and to determine which type of simulation is most effective.

METHODS

The study was undertaken in accordance with a protocol written prior to the commencement of the review process and published on the PROSPERO database (CRD42013005105).[28]

Criteria for selecting studies for review

All included studies were randomised controlled trials that assessed the effectiveness of simulation based teaching compared to other teaching methods, or no teaching, in medical students.

We included studies with teaching interventions directed at critical care, intensive care, anaesthetics, emergency medicine, trauma, or pre-hospital care; studies that used simulation based teaching interventions, which included the use of high and low fidelity mannequins, standardised patients, screen-based computer simulators, and human or animal cadavers; studies which used outcomes of knowledge or skill-based performance in the care of a critically ill patient; and studies whose comparator group was a different type

of simulation technology, a different type of teaching modality, or no teaching. (See Table in Supplemental Digital Content 1, for a list of definitions used for inclusion criteria).

Studies were excluded where participants had already graduated medical school or were other health professionals, studies which had non-randomised designs, those that studied non-acute specialities, or used other types of comparator groups.

Search methods for identification of studies

Studies were identified by systematically searching AMED, EMBASE, MEDLINE, Education Resources Information Centre (ERIC), British Education Index (BEI) and Australian Education Index (AEI) up to July 2013. The search strategy was designed for high sensitivity over precision, to ensure that no relevant studies were lost. The search broadly covered 'medical students', 'simulation' and 'acute specialities' (see Table, Supplemental Digital Content 2, which demonstrates the full search strategy). The reference lists and indexed citations of all included studies were checked for further relevant studies, and authors of included studies were contacted for unpublished literature.

Abstracts of identified studies were independently screened by reviewers MDB and TDM against eligibility criteria. The full texts of potentially eligible studies were obtained and independently screened in full by two reviewers (MDB, and TDM or REW). There were no disagreements at any stage.

Maximal data extraction was carried out independently and in parallel by two reviewers (MDB, and CRA or REW) using a piloted standard format which included methodology, participants, outcome measures, and results. Introducing duplicate extraction of all included studies was the only deviation from the protocol, to reduce the risk of reporter bias. Forms were checked for completeness and discrepancies resolved by reviewing the original article. All discrepancies involved missing information and none were methodological issues or disagreements in interpretation.

Quality Assessment

Individual study quality was assessed using the Cochrane risk of bias assessment tool,[29] which assesses the risk of selection bias during random sequence generation and allocation concealment, performance bias through inadequate blinding of participants and personnel, detection bias through inadequate blinding of outcome assessment, attrition bias through incomplete reporting, and reporting bias through the selective reporting of trials. Additionally we considered any other biases which arose, particularly industry funding by manufacturers of simulation equipment. Authors of studies with unclear risk of bias were contacted for missing information. The quality of evidence for each outcome was assessed using the GRADE framework, which considers five key elements: study design, indirectness of evidence, unexplained heterogeneity, imprecision of results and high probability of publication bias.[30]

Statistical Analysis

We assessed the effectiveness of simulation using outcomes of either knowledge or clinical performance. All of the papers included used continuous outcomes for the measurement of knowledge or clinical performance, including those that used checklists who presented mean checklist scores. Using Cochrane's Review Manager 5.2[31] this mean score was converted to a standardised mean difference, (hedges g) which corrects for differences in measurement scales, making the assumption that variability in the standard deviation arises from differences in the continuous scales used to measure the outcome, rather than variability in the population. Where multiple outcomes were assessed in a study, we determined which outcome measure to include in the review using a hierarchy of outcome measures, based on Miller's Hierarchy,[32] developed by JK who was blinded to the data (Figure 1).

Studies with two intervention arms had both arms combined to form a single intervention group for the main analysis using the standard techniques suggested in the Cochrane Handbook (by combining numbers into a single sample size, mean and standard deviation).[29] For the purposes of sub-group analysis, both arms were examined independently. Where data were unavailable, standard deviations were imputed from p-values by calculating t-values and degrees of freedom to estimate a standard error,[29] or from confidence intervals (CI) using the calculator in Cochranes Review Manager.[31] Paired analysis data from crossover trials were used where there was no evidence of carry-over effect.[29]

For each outcome we assessed heterogeneity using the I^2 test, where an I^2 >75% is sufficient to indicate evidence of considerable inconsistency.[29] In the presence of heterogeneity, we obtained pooled effect estimates and 95% confidence intervals by carrying out an inverse-variance random effects meta-analysis (for all analyses) using the DerSimonian and Laird method in Cochrane's Review Manager 5.2.[31, 33] This type of analysis adds a weighting for within study and between study variations which is helpful in the presence of high heterogeneity. So outlier studies receive less weighting than studies that sit around the mean.

We carried out subgroup analyses to investigate the effects of time to outcome assessment, type of outcome assessment, type of simulation, type of control group, duration of simulation, and year of study. Sensitivity analyses were carried out to examine the effect of exclusion of outliers, high risk of bias, industry funding, imputed standard deviations, and crossover trials. Publication bias was examined using funnel plots.

Results were expressed as standardised mean differences, with 95% confidence intervals and *p*-values, and as percentile change derived from z-scores, which demonstrates the percentile group that the average student in the simulation group would be in when compared to students who received the control intervention.

RESULTS

From the electronic searches we screened 356 abstracts, 326 of which were clearly not relevant, identifying 30 potentially eligible articles. The abstracts of a further 482 references and 437 citations were also screened, identifying a further 14 potentially eligible articles. A total of 44 potentially eligible articles were retrieved in full text and assessed in duplicate for inclusion in the review (Figure 2). Two ongoing studies were also identified (See Table, Supplemental Digital Content 3, for characteristics of excluded and ongoing studies.)

Description of studies

Twenty two papers were included in the review (Table 1), including 1,325 medical students in their second years and above, mainly studying at European or North American medical schools. The number of participants in each study ranged from 28 to 144, with a median of 45. Fifteen studies examined high fidelity simulators, five examined low fidelity simulators, two standardised patient simulations, three screen-based computer simulators, and one study examined a voice advisory mannequin. Eight studies used self-directed learning techniques in their control group (problem based learning (PBL), case based discussion and self-study), six used didactic teaching methods (lecture, video and seminar), one used clinical shadowing and two studies used no teaching. For the purposes of the meta-analysis, studies which gave students material to cover in self-study were categorised as having a comparator teaching intervention as this was guided study. Three studies used low fidelity simulation in their control group, therefore comparing two types of simulation. The median duration of intervention sessions was 2 hours (range 5 minutes to 3 days). The number of

studies do not consistently sum to 22 as several of the studies compared a number of different types of simulation, used a number of different outcome measures, or did not have a non-simulation comparator group.

Eleven studies used knowledge-based assessments (Kirkpatrick level 2a) including Multiple Choice Questions (MCQs), Short Answer Questions (SAQs) and Single Best Answers (SBAs), whilst 15 studies used skill- and performance-based outcome measures including Objective Structured Clinical Examination (OSCE) scores, simulation checklists and time to action (Kirkpatrick level 2b). Eleven studies also used evaluative questionnaires to assess aspects of satisfaction, and three studies used self-efficacy questionnaires to assess participant's confidence (Kirkpatrick level 1). A total of 19 studies assessed students within one week of the intervention, and only three studies followed participants up after three months. There were no studies which assessed for evidence of transfer of learning to clinical practice (Kirkpatrick level 3) or benefit to patients (Kirkpatrick level 4).

The overall risk of bias was high in seven included studies and unclear in the remaining 15 (Figure 3). The majority of studies inadequately reported key risk of bias criteria, making it difficult to precisely judge study quality. We were particularly concerned by studies that did not explain blinding of outcome assessors and randomisation procedures. (See Figure in Supplemental Digital Content 4, which demonstrates the full risk of bias assessment for each study.)

Is simulation effective compared to other teaching methods?

A total of 17 studies compared simulation with other teaching modalities (Figure 4), reporting knowledge- or skill-based performance measures after the teaching session. However, one study reported only median data [34] and in one study participant numbers were unclear,[35] so the 15 remaining studies (1000 participants) were included in the analyses. Simulation was significantly more effective than other teaching methods when data were pooled, with an effect size of 0.84 (95% CI = 0.43 to 1.24; p=<0.001; Z=4.02, I^2=89%) corresponding to a percentile gain of 49.9 percentiles.

However, one study that reported only medians showed no evidence of improved effectiveness of simulation over other teaching methods; medians 37 vs. 38 (Scale 0 to 50; P=0.263) respectively.[34] The study in which participant numbers were unclear showed no evidence of improved effectiveness of simulation over other teaching methods; SMD -0.13 (95% CI -0.72 to 0.47; p=0.47).[35]

Sensitivity analyses excluding studies that were at high risk of bias, with imputed standard deviations, industry funded, or of crossover design retained a statistically significant effect. All studies were of a small size and were therefore grouped together on the funnel plot, making it ineffective for assessing small study bias (Figure 5). Despite this, there was some suggestion of asymmetry as studies with high risk of bias were generally smaller and more positive in effect size. Whilst this suggests that small studies with less positive effects may not have appeared in the literature, sensitivity analysis removing the high risk studies

retained a statistically significant effect (Table 2) and resulted in a symmetrical funnel plot. This suggests that even if it exists, small study or publication bias is of little significance to our overall effect estimates. We carried out sub-group analysis (Table 2) by time to outcome assessment (See Figure in Supplemental Digital Content 5, which shows that simulation was more effective when assessed at less than 72 hours), type of outcome assessment (See Figure in Supplemental Digital Content 6, which shows that simulation was effective in performance based outcomes but no evidence of effect in knowledge based outcomes), type of simulation (Figure 6), duration of simulation (See Figure in Supplemental Digital Content 7, which shows that simulation was more effective when used for over 8 hours), and year of study (See Figure in Supplemental Digital Content 8, which shows that simulation was effective beyond year four of medical school, but no evidence prior to this), type of control group (See Figure in Supplemental Digital Content 9, which shows that simulation was more effective than dependent and independent teaching techniques, but no significant effect compared to self-study). Subgrouping did not explain the heterogeneity (Table 2).

Two studies (78 participants) compared simulation to no teaching. The effect size was 3.41 (95% CI = -2.57 to 9.40; p=0.26, Z=1.12) which was not significant and corresponded to a gain of 36.9 percentiles and was significantly heterogeneous (I^2=*98%*). (See figure in Supplemental Digital Content 10, which demonstrates no evidence of effect compared to no teaching.)

Which type of simulation is most effective?

We examined studies that directly compared different types of simulation teaching (Figure 6). Three studies (173 participants) compared high fidelity simulation with low fidelity simulation. However, one of these studies did not present mean data and the remaining two studies (130 participants) that were included in meta-analysis favoured high fidelity simulation, with an effect size of 1.00 (95% CI 0.63 to 1.37; p<0.001).[36, 37] The other study (43 participants) favoured low fidelity simulation over high fidelity simulation, with a median (Interquartile Range) of 29 (29 to 30) and 26 (25 to 28) respectively (p=0.03).[38] One study (48 participants) compared high fidelity simulation with standardised patients and found no evidence of a difference with an effect size of 0.43 (95% CI -0.14 to 1.01, p=>0.05).[39] One study (28 participants) compared low fidelity simulation with screen based computer simulators and found no evidence of a difference, with an effect size of -0.11 (95% CI -0.85 to 0.63, p=0.77).[40]

Comparisons were also made between types of simulation by sub-grouping studies that compared types of simulation with other teaching methods. Twelve studies (797 participants) reported the use of high fidelity patient simulators [effect size 0.90 (95% CI 0.48 to 1.31; p<0.001; Z=4.25, I^2=86%)] corresponding to a gain of 50.0 percentiles. Two studies (121 participants) reported the use of screen-based computer simulators, with no evidence of an effect [effect size -0.07 (-1.17 to 1.04; p=0.91; Z=0.12)]. Two studies (87 participants) reported the use of low fidelity simulators, with no evidence of an effect [effect size 1.39 (-0.95 to 3.74; p=0.24; Z=1.17)]. One study (46 participants) reported the

use of standardised patients [effect size 1.94 (95% CI 1.23 to 2.65; p<0.001; Z=5.34)]. The results of both the direct and indirect sub-grouped comparisons were resistant to sensitivity analysis that excluded studies with high risk of bias, industry funding, imputed standard deviations and crossover design.

According to the GRADE criteria, (Table 3) the quality of the evidence for simulation against other teaching methods was moderate. The GRADE assessment was downgraded twice to account for the unclear risk of bias across all studies and the unexplained inconsistency indicated by statistically significant heterogeneity. However, the GRADE assessment was upgraded once for the large and practically important effect size that was resilient to sensitivity analysis.

DISCUSSION

Our review suggests that simulation based medical education is more effective for teaching critical care medicine to students than other teaching methods. The size of the effect is large (0.84) according to Cohen who categorises effects of <0.2 as small, 0.2-0.8 as moderate and >0.8 as large.[41] However, this interpretation should be used with caution[42] as in education even small effect sizes have been shown to be important in policy decision making.[43] Using Z-scores to calculate the percentile change we observed an increase of 49.8 percentiles in the simulation group compared to the other teaching groups. This means that the average student in the simulation group would be in the 99.8th percentile of the control

group. Considering a median simulation duration of just 2 hours, we considered this to be a large and practically important effect.

This review builds on a growing body of evidence across a range of healthcare professions. A systematic review by Cook demonstrated that simulation is effective in postgraduate nurses for knowledge and skill acquisition, with an effect size of 1.20 (95% CI 1.04 to 1.35), and 1.09 (95% CI 1.03 to 1.16) respectively.[13] A systematic review by Yuan also demonstrated that simulation is effective in other health professionals for knowledge and skill acquisition, with an effect size of 0.53 (95% CI 0.16 to 0.90 p=0.006), and 1.15 (0.78 to 1.52, p<0.001) respectively.[20] A systematic review by McGaghie found that simulation is effective for clinical skill acquisition across a range of medical seniorities, with an effect size of 0.71 (95% CI 0.65 to 0.76 p<0.001).[16] A systematic review by Lorello found simulation to be more effective in anaesthesiology training across a number of seniorities, with an effect size range of 0.60 to 1.05.[17] The largest systematic review by Ilgen, which incorporated a range of professions and stages of development, found no evidence of an effect for simulation compared to other teaching modalities for knowledge and skills, with an effect size of 0.26 (95% CI -0.08 to 0.60, p=0.14) and 0.19 (95% CI -0.10 to 1.23, p=0.21) respectively.[26] Whilst there is some evidence of inconsistency amongst the existing systematic reviews, this is unsurprising given the differences between participants. Our pooled effect estimates, are statistically consistent with these other studies and demonstrate similar effect sizes to those of Cook, Yuan and McGaghie. This is the first systematic review to describe the effectiveness of simulation for teaching critical care medicine in the medical school setting.

It is perhaps unsurprising that simulation is more effective than other teaching modalities in improving performance-related outcomes (Kirkpatrick 2b) since it is a performance-based method of learning. However, despite adequate power (>0.99) we found no evidence that simulation was more effective than other teaching modalities in preparing for knowledge-based assessments (Kirkpatrick 2a). This is an important finding because simulation based teaching is a resource- and faculty-intensive education technique, which has significant cost implications. To maximise its cost-benefit impact will require defining of the optimal context in which simulation should be used, and this study helps to define that position. The finding is in contrast to the findings of reviews in other trainee groups, which suggests that the type of knowledge or skills gained relates to the level of expertise of the learner.[13, 20] Cognitive Load theory[44] and the Challenge Point framework[45] provide conceptual frameworks to help explain how simulation may impact differently on learning depending on prior level of knowledge. It is therefore important to separate undergraduate from postgraduate learner cohorts when defining the effectiveness of different learning methods, as well as different types of simulation within medical education.

Our finding may go some way to supporting the theory that simulation promotes the transition of knowledge ("knows") into reasoned action ("does"),[24, 46] which would help to explain why we were unable to demonstrate any effect in knowledge based outcomes. This would therefore support the view that simulation is best used as an adjunct to other teaching methods in the undergraduate curriculum, rather than as a standalone method. We would postulate that simulation would be best placed alongside PBL, and didactic

teaching methods in integrated curricula, or following in traditional domain centred curricula.

This study also demonstrated that high fidelity simulation and the use of standardised patients were more effective than other teaching modalities, but that there was no evidence of effectiveness for low fidelity simulation or screen-based computer simulation compared to other teaching modalities. We found that in direct comparisons high fidelity simulation was more effective than low fidelity simulation, which is in contrast to a number of studies in other groups which showed no difference in their efficacy.[47] This finding was not well explained by duration of simulation exposure and is difficult to interpret since the term 'fidelity' is not used consistently by all researchers, which may variously refer to environmental, functional or psychological fidelity.[48]

Although we demonstrated that simulation was more effective than lectures, problem based learning, and other similar techniques when pooled, we could find no evidence of a difference when comparing simulation with independent study or no-teaching. This finding is counter-intuitive, in that if simulation is effective compared to other teaching methods, it would be expected to be more effective than no teaching. This analysis was however limited to only two studies which had significantly heterogenous results, so this result may be due to an outlier study. The study by Ali showed a large and significant effect comparing simulation to no teaching.[49] The study by Hansel showed no evidence of effect comparing simulation to no teaching, and they postulated that the scenarios they used may have been

too complex for their participant group, further supporting the view that simulation may not always be effective in this learner group.[50]

The evidence supports the use of simulation for teaching critical care medicine to medical students. However, this review has been unable to address differences between types of simulation technology, the effect of duration or frequency of simulation teaching (the 'dose' of simulation), the optimal timing by year of study, or retention of skills post simulation. Further work is also needed to categorise the cost effectiveness of simulation based teaching, as equipment and operational costs are high.[51]

Limitations

Despite a thorough literature search using pre-specified criteria and a protocol designed according to methods specified in the Cochrane Handbook,[29] there are limitations to the study. Reviewers were non-blinded throughout the study, which may have biased coding and interpretation of data. However, we felt this was unlikely given the high levels of agreement.

Most of the studies which used skill or behavioural based outcome measures during simulated patient scenarios used the same simulators during the assessment as in the teaching session. This may be considered an important source of bias as the simulation group has the advantage of being assessed on the same simulator used for training. All but one study carried out at least one orientation session on the simulator for all intervention

groups. Our effect size was resilient to the removal of these studies from the metaanalysis and maintained statistical significance.

Another issue was that many requests for further information from authors of the included studies went unanswered, which meant that analysis was limited for a large number of studies. This forced us to include studies with an unclear risk of bias in the meta-analysis, when these studies may have been more appropriately rated as having low or high risk of bias with the additional information. Studies assessed to be at high risk of bias generally had larger and more positive effect sizes[24, 46, 52, 53] than those with low risk of bias. However, most of the included studies favoured simulation and our effect size was resistant to removal of these studies from the meta-analysis.

Furthermore, we identified significant heterogeneity which we were unable to explain through sub-group and sensitivity analyses, suggesting that the results are limited by the quantity and quality of original papers identified. Despite high correlation in measured effects amongst studies, responsiveness of the different outcome measures used may vary substantially between studies which may result in significant heterogeneity and therefore a biased meta-analysis. Inconsistency is a common problem in quantitative educational research which has led some to argue that qualitative methods are more suited in this domain.[54] The use of standardised and validated outcome measures would go some way to helping prevent this inconsistency in future research. Despite the inconsistency in effect size,

the majority of included studies favoured simulation, with only a small number favouring the control interventions.

Conclusions

This systematic review and meta-analysis provides moderate evidence that simulation is effective for teaching critical care medicine to medical students, yielding large favourable benefits over other teaching methods despite relatively short simulated sessions. High Fidelity Simulation appears more effective than Low Fidelity. Simulation was particularly effective in preparing students for clinical performance-based assessments, but not for knowledge-based assessments. However, whether this translates into improved performance in the authentic clinical setting is unproven.

This review is important for medical educators who are responsible for teaching acute care clinical skills to medical students, and are faced with a panoply of educational techniques on the one hand, and a finite budget on the other. The findings also support an educational method that may go some way to mitigating the ethical tensions that arise through teaching critical care medicine to undergraduates. Further high quality research is needed to determine the best way to integrate simulation into undergraduate curriculums, which should also address the broader questions of when, how and why simulation works.

DETAILS OF AUTHOR CONTRIBUTIONS

MDB: study concept and design, search strategy, design of data collection tools, collection and analysis of data, drafting of the manuscript. LH: Study design, search strategy, design of data collection tools, supervision of data collection, analysis and interpretation of data, and critical revision of the article. JK: Critical review of the study protocol, interpretation of data, and critical revision of the article. TDM, RDW and CRA: critical review of the study protocol, design and piloting of data collection tools, collection of data, and critical revision of the article. MDB is the guarantor and affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and registered) have been explained.

ACKNOWLEDGMENTS

We thank Dr. Lesley Bowker and Professor Sam Leinster for commenting on the drafts, providing valuable feedback and discussion of our findings and conclusions; and Dr. Jan Wong for her help screening a full text in the non-English language on our behalf.

We would especially like to thank the authors of both included and excluded studies for their responses to our requests for unpublished literature and for further information: Adel Bassily-Marcus, Mount Sinai School of Medicine; Peyman Benharash and Paul Frank, UCLA; Ester Coolen, Radboud University; Rosemarie Fernandez, University of Washington School of Medicine; Mike Gilbart, University of British Columbia; James Gordon, Harvard Medical

School; Bruce Lo, Eastern Virginia Medical School; Pamela Morgan, University of Toronto; Gavin Perkins, Warwick Medical School; and Raymond Ten-Eyck, Boonshoft School of Medicine.

COMPETING INTERESTS

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

FUNDING

No external funding was sought for this review.

REFERENCES

1 Shen J, Joynt GM, Critchley LAH, Tan IKS, Lee A. Survey of current status of intensive care teaching in English-speaking medical schools. *Critical Care Medicine* 2003; **31**: 293-8

2 Smith GB, Poplett N. Knowledge of aspects of acute care in trainee doctors. *Postgraduate Medical Journal* 2002; **78**: 335-8

3 Smith CM, Perkins GD, Bullock I, Bion JF. Undergraduate training in the care of the acutely ill patient: a literature review. *Intensive Care Medicine* 2007; **33**: 901-7

4 Jensen ML, Hesselfeldt R, Rasmussen MB, et al. Newly graduated doctors' competence in managing cardiopulmonary arrests assessed using a standardized Advanced Life Support (ALS) assessment. *Resuscitation* 2008; **77**: 63-8

5 O'Dowd A. Locums make up a fifth of doctors in emergency units at weekends. *BMJ* 2013; **346**

6 Ziv A, Wolpe PR, Small SD, Glick S. Simulation-Based Medical Education: An Ethical Imperative. *Academic Medicine* 2003; **78**: 783-8

7 Phillips PS, Nolan JP. Training in basic and advanced life support in UK medical schools: questionnaire survey. *BMJ* 2001; **323**: 22-3

8 Vozenilek J, Huff JS, Reznek M, Gordon JA. See One, Do One, Teach One: Advanced Technology in Medical Education. *Academic Emergency Medicine* 2004; **11**: 1149-54

9 Beaubien JM, Baker DP. The use of simulation for training teamwork skills in health care: how low can you go? *Quality and Safety in Health Care* 2004; **13**: i51-i6

10 GMC. Tomorrow's Doctors. London: General Medical Council, 2009

11 Issenberg BS, Mcgaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review*. *Medical Teacher* 2005; **27**: 10-28

12 Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. *Acad Med* 2010; **85**: 1589-602

13 Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: A systematic review and meta-analysis. *JAMA - Journal of the American Medical Association* 2011; **306**: 978-88

14 Cooper S, Cant R, Porter J, et al. Simulation based learning in midwifery education: A systematic review. *Women and Birth* 2012; **25**: 64-78

15 Lynagh M, Burton R, Sanson-Fisher R. A systematic review of medical skills laboratory training: Where to from here? *Med Educ* 2007; **41**: 879-87

16 McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Academic Medicine* 2011; **86**: 706-11

17 Lorello GR, Cook DA, Johnson RL, Brydges R. Simulation-based training in anaesthesiology: a systematic review and meta-analysis. *British Journal of Anaesthesia* 2014; **112**: 231-45

18 Heitz C, Eyck RT, Smith M, Fitch M, Ander D. Simulation in Medical Student Education: Survey of the Clerkship Directors in Emergency Medicine. *Western Journal of Emergency Medicine* 2011; **12**: 455-60

19 Chakravarthy B, Ter Haar E, Bhat SS, McCoy CE, Denmark TK, Lotfipour S. Simulation in medical school education: review for emergency medicine. *West J Emerg Med* 2011; **12**: 461-6

20 Yuan HB, Williams BA, Fang JB, Ye QH. A systematic review of selected evidence on improving knowledge and skills through high-fidelity simulation. *Nurse Education Today* 2012; **32**: 294-8

21 Sweller J. Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction* 1994; **4**: 295-312

22 Morgan PJ, Cleave-Hogg D, Desousa S, Lam-Mcculloch J. Applying theory to practice in undergraduate education using high fidelity simulation. *Medical Teacher* 2006; **28**: e10-e5

23 Morgan P, Cleave-Hogg D. A Canadian simulation experience: faculty and student opinions of a performance evaluation study. *British Journal of Anaesthesia* 2000; **85**: 779-81

24 Gordon J, Shaffer D, Raemer D, Pawlowski J, Hurford W, Cooper J. A Randomized Controlled Trial of Simulation-Based Teaching versus Traditional Instruction in Medicine: A Pilot Study among Clinical Medical Students. *Adv Health Sci Educ* 2006; **11**: 33-9

25 Kim JH, Kim WO, Min KT, Yang JY, Nam YT. Learning by computer simulation does not lead to better test performance than textbook study in the diagnosis and treatment of dysrhythmias. *J Clin Anesth* 2002; **14**: 395-400

26 Ilgen JS, Sherbino J, Cook DA. Technology-enhanced Simulation in Emergency Medicine: A Systematic Review and Meta-Analysis. *Academic Emergency Medicine* 2013; **20**: 117-27

27 Cook DA, Brydges R, Hamstra SJ, et al. Comparative effectiveness of technology-enhanced simulation versus other instructional methods: a systematic review and meta-analysis. *Simul Healthc* 2012; **7**: 308-20

28 Beal M, Hooper L. The effectiveness of medical simulation in teaching medical students critical care medicine: a protocol for a systematic review. *PROSPERO*, 2013

29 Higgins J, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*: The Cochrane Collaboration, 2011

30 The GRADE working group. Grading quality of evidence and strength of recommendations. *BMJ* 2004; **328**: 1490

31 The Nordic Cochrane Centre. Review Manager (RevMan). Copenhagen: The Cochrane Collaboration, 2012

32 Miller GE. The assessment of clinical skills/competence/performance. *Academic Medicine* 1990; **65**: S63-7

33 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**: 177-88

34 Cavaleiro A, Guimarães H, Calheiros F. Training neonatal skills with simulators? *Acta Paediatrica, International Journal of Paediatrics* 2009; **98**: 636-9

35 Morgan PJ, Cleave-Hogg D. Comparison between medical students' experience, confidence and competence. *Med Educ* 2002; **36**: 534-9

36 Coolen EH, Draaisma JM, Hogeveen M, Antonius TA, Lommen CM, Loeffen JL. Effectiveness of high fidelity video-assisted real-time simulation: a comparison of three training methods for acute pediatric emergencies. *Int J Pediatr* 2012; **2012**: 709569

37 Lo BM, Devine AS, Evans DP, et al. Comparison of traditional versus high-fidelity simulation in the retention of ACLS knowledge. *Resuscitation* 2011; **82**: 1440-3

38 Isbye DL, Hoiby P, Rasmussen MB, et al. Voice advisory manikin versus instructor facilitated training in cardiopulmonary resuscitation. *Resuscitation* 2008; **79**: 73-81

39 Ali J, Al Ahmadi K, Williams JI, Cherry RA. The standardized live patient and mechanical patient models--their roles in trauma teaching. *J Trauma* 2009; **66**: 98-102

40 Bonnetain E, Boucheix JM, Hamet M, Freysz M. Benefits of computer screen-based simulation in learning cardiac arrest procedures. *Med Educ* 2010; **44**: 716-22

41 Cohen J. *Statistical power analysis for the behavioral sciencies*. Routledge, 1988

42 Durlak JA. How to Select, Calculate, and Interpret Effect Sizes. *Journal of Pediatric Psychology* 2009; **34**: 917-28

43 Hedges LV, Hedberg EC. Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis* 2007; **29**: 60-87

44 Van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Med Educ* 2010; **44**: 85-93

45 Guadagnoli M, Morin M-P, Dubrowski A. The application of the challenge point framework in medical education. *Med Educ* 2012; **46**: 447-53

46 McCoy CE, Menchine M, Anderson C, Kollen R, Langdorf MI, Lotfipour S. Prospective Randomized Crossover Study of Simulation vs. Didactics for Teaching Medical Students the Assessment and Management of Critically Ill Patients. *The Journal of Emergency Medicine* 2011; **40**: 448-55

47 Norman G, Dore K, Grierson L. The minimal relationship between simulation fidelity and transfer of learning. *Med Educ* 2012; **46**: 636-47

48 Maran NJ, Glavin RJ. Low- to high-fidelity simulation – a continuum of medical education? *Med Educ* 2003; **37**: 22-8

49 Ali J, Adam RU, Sammy I, Ali E, Williams JI. The Simulated Trauma Patient Teaching Module-Does it Improve Student Performance? *Journal of Trauma- Injury, Infection, and Critical Care* 2007; **62**: 1416-20

50 Hansel M, Winkelmann AM, Hardt F, et al. Impact of simulator training and crew resource management training on final-year medical students' performance in sepsis resuscitation: a randomized trial. *Minerva Anestesiol* 2012; **78**: 901-9

51 Morgan PJ, Cleave-Hogg DM. Cost and resource implications of under-graduate simulator-based education [3]. *Canadian Journal of Anesthesia* 2001; **48**: 827-8

52 Ruesseler M, Weinlich M, Müller MP, Byhahn C, Marzi I, Walcher F. Simulation training improves ability to manage medical emergencies. *Emergency Medicine Journal* 2010; **27**: 734-8

53 Steadman RH, Coates WC, Huang YM, et al. Simulation-based training is superior to problem-based learning for the acquisition of critical assessment and management skills. *Critical Care Medicine* 2006; **34**: 151-7

54 Hoepfl MC. Choosing Qualitative Research: A Primer for Technology Education Researchers. *Journal of Technology Education* 1997; **9**

55 Curran VR, Aziz K, O'Young S, Bessell C. Evaluation of the effect of a computerized training simulator (ANAKIN) on the retention of neonatal resuscitation skills. *Teaching & Learning in Medicine* 2004; **16**: 157-64

56 Gilbart MK, Hutchison CR, Cusimano MD, Regehr G. A computer-based trauma simulator for teaching trauma management skills. *The American Journal of Surgery* 2000; **179**: 223-8

57 Schwartz LR, Fernandez R, Kouyoumjian SR, Jones KA, Compton S. A Randomized Comparison Trial of Case-based Learning versus Human Patient Simulation in Medical Student Education. *Academic Emergency Medicine* 2007; **14**: 130-7

58 Tan GM, Ti LK, Tan K, Lee T. A comparison of screen-based simulation and conventional lectures for undergraduate teaching of crisis management. *Anaesthesia & Intensive Care* 2008; **36**: 565-9

59 Ten Eyck RP, Tews M, Ballester JM. Improved Medical Student Satisfaction and Test Performance With a Simulation-Based Emergency Medicine Curriculum: A Randomized Controlled Trial. *Ann Emerg Med* 2009; **54**: 684-91

60 Ten Eyck RP, Tews M, Ballester JM, Hamilton GC. Improved fourth-year medical student clinical decision-making performance as a resuscitation team leader after a simulation-based curriculum. *Simul* 2010; **5**: 139-45

61 Wenk M, Waurick R, Schotes D, et al. Simulation-based medical education is no better than problem-based discussions and induces misjudgment in self-assessment. *Adv Health Sci Educ* 2009; **14**: 159-71

62 Yang LY, Yang EJ, Ying LL, et al. The use of human patient simulator in enhancing medical students understanding of crisis recognition and resuscitation. *International Medical Journal* 2010; **17**: 209-11

FIGURE LEGENDS

Table of Legends	
Figure No.	**Caption**
Figure 1	Figure 1: Hierarchy of Outcome Measures (1 is most preferred, 11 is least preferred) *†: Sub-hierarchy for further content 1) Acute Coronary Syndrome, 2) Stroke, 3) Asthma, 4) Trauma, 5) In-hospital CPR, 6) Motor-cyclist helmet removal and stiff neck, 7) Infant CPR as first Responder, 8) ECG attachment and interpretation, 9) Intra-osseous Access, 10) Pre-hospital CPR with AED*
Figure 2	Figure 2: Study Flow Diagram. Abbreviations: RCT = Randomised Control Trial, CCM = Critical Care Medicine
Figure 3	Figure 3: Risk of bias graph: Review author's judgements across all included studies
Figure 4	Figure 4: The effectiveness of simulation on performance or knowledge scores in medical students (higher scores represent better performance or knowledge)
Figure 5	Figure 5: Funnel plot assessing risk of publication bias (Standardised Mean Difference vs. Standard Error of the Standardised Mean Difference)
Figure 6	Figure 6: The effectiveness of different types of simulation - direct and indirect analyses. (1) Right hand side of Forest Plot, (2) Left hand side of forest plot.
Table 1	Table 1: Characteristics of Included Studies
Table 2	Table 2: Sub-grouping and sensitivity analyses
Table 3	Table 3: GRADE evidence profile
SDC 1	Supplemental Digital Content 1: Table of Definitions
SDC 2	Supplemental Digital Content 2: Literature Search Terms
SDC 3	Supplemental Digital Content 3: Characteristics of excluded and ongoing Studies
SDC 4	Supplemental Digital Content 4: Risk of bias summary: Review author's judgement for each included study.
SDC 5	Supplemental Digital Content 5: Effectiveness of simulation compared to other teaching methods: Sub-grouped by time to outcome assessment
SDC 6	Supplemental Digital Content 6: Effectiveness of simulation compared to other teaching methods: Sub-grouped by type of outcome measure
SDC 7	Supplemental Digital Content 7: Effectiveness of simulation compared to other teaching methods: Sub-grouped by dose of simulation session
SDC 8	Supplemental Digital Content 8: Effectiveness of simulation compared to other teaching methods: Sub-grouped by year of study
SDC 9	Supplemental Digital Content 9: Effectiveness of simulation compared to other teaching methods: Sub-grouped by control intervention
SDC 10	Supplemental Digital Content 10: The effectiveness of simulation compared to no-teaching (higher scores represent better performance or

	knowledge)

NOT FOR PUBLICATION - Table for the purposes of automatic Endnote referencing of included texts -

Ali, 2007, Trinidad and Tobago

Ali, 2009, Canada

Bonnetain, 2010, France

Cavaleiro, 2009, Portugal

Coolen, 2012, Netherlands

Curran, 2004, Canada

Gilbart, 2000, Canada

Gordon, 2006, US

Hansel, 2012, Germany

Isbye, 2008, Denmark

Kim, 2002, South Korea

Lo, 2011, US

McCoy, 2011, US

Morgan, 2002, Canada

Ruesseler, 2010, Germany

Schwartz, 2007, US

Steadman, 2006, US

Tan, 2008, Singapore

Ten Eyck, 2009, US

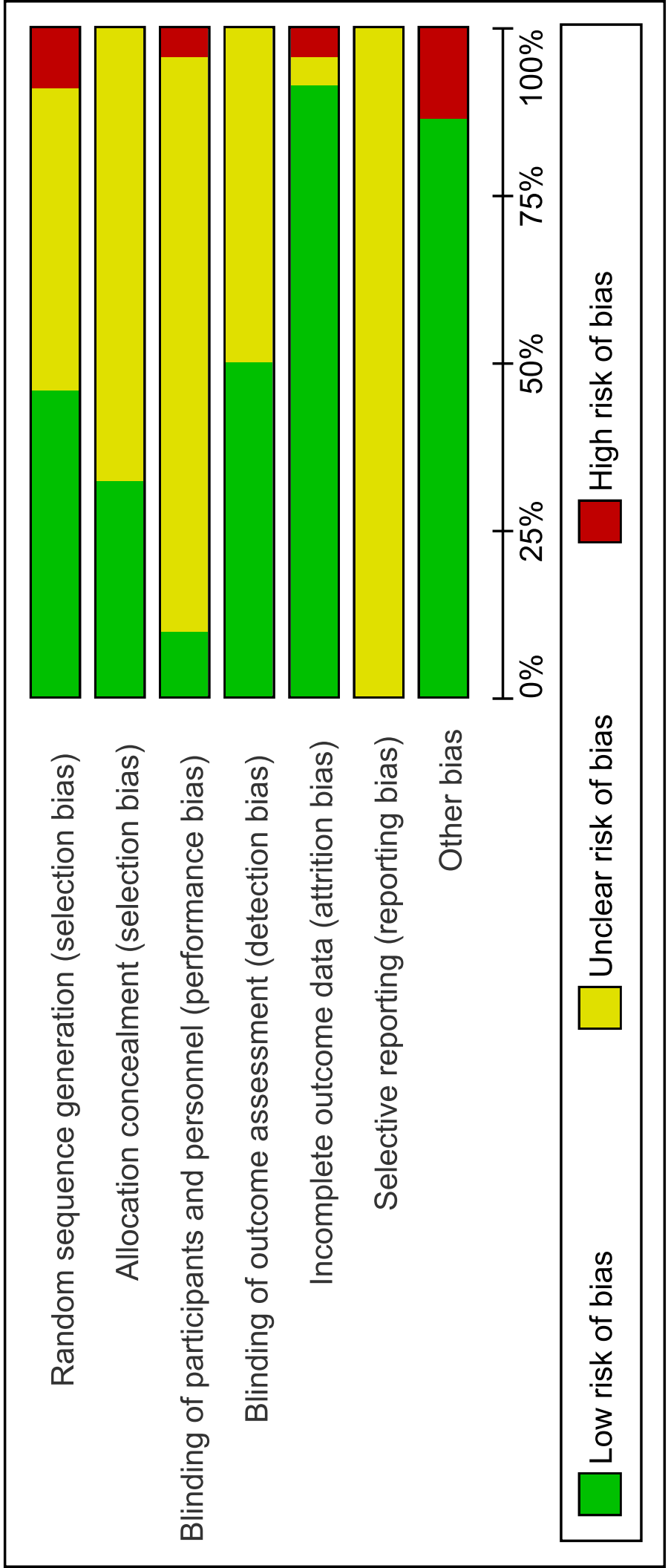
Ten Eyck, 2010, US

Wenk, 2009, Germany

Yang, 2010, Malaysia

Figure 1: Hierarchy of Outcome Measures

Figure 3: Risk of bias graph



Study or Subgroup	Simulation Mean	SD	Total	Non-Simulation Mean	SD	Total	Weight	Std. Mean Difference IV, Random, 95% CI
1.1.1 Simulation compared to other teaching modalities								
Ali 2009	85.5	3.45	48	77.5	3.8	22	6.6%	2.22 [1.59, 2.85]
Coolen 2012	68.72	13.1	29	60.64	13.4	14	6.5%	0.60 [-0.05, 1.25]
Curran 2004	210.98	22.5	16	199.89	13.81	15	6.3%	0.57 [-0.15, 1.29]
Gilbart 2000	8.41	1.31	57	8.05	4.41	50	7.3%	0.11 [-0.27, 0.49]
Gordon 2006	73	10.8	21	70.5	16.3	17	6.5%	0.18 [-0.46, 0.82]
Kim 2002	10.3	2.9	29	12.2	3	28	6.9%	-0.64 [-1.17, -0.10]
McCoy 2011	93	5.1	28	71	12.8	28	6.4%	2.23 [1.55, 2.90]
Ruessler 2010	20.7	2.5	22	14.2	2.4	22	5.9%	2.60 [1.78, 3.43]
Schwartz 2007	31.2	3.6	50	31.4	4.1	52	7.3%	-0.05 [-0.44, 0.34]
Steadman 2006	71	11.7	15	52	11.7	16	5.9%	1.58 [0.76, 2.40]
Tan 2008	46	12	33	40	12	31	7.0%	0.49 [-0.00, 0.99]
Ten Eyck 2009	89.83	8.18	90	86.44	9.81	90	7.5%	0.37 [0.08, 0.67]
Ten Eyck 2010	-28.3	53.3	34	-86	52.44	34	6.9%	1.08 [0.57, 1.59]
Wenk 2009	-483.3	85.1	16	-609.3	99.1	16	6.1%	1.33 [0.55, 2.11]
Yang 2010	35.1	4.1	37	33	4.2	40	7.1%	0.50 [0.05, 0.96]
Subtotal (95% CI)			**525**			**475**	**100.0%**	**0.84 [0.43, 1.24]**

Heterogeneity: Tau² = 0.56; Chi² = 124.60, df = 14 (P < 0.00001); I² = 89%

Test for overall effect: Z = 4.02 (P < 0.0001)

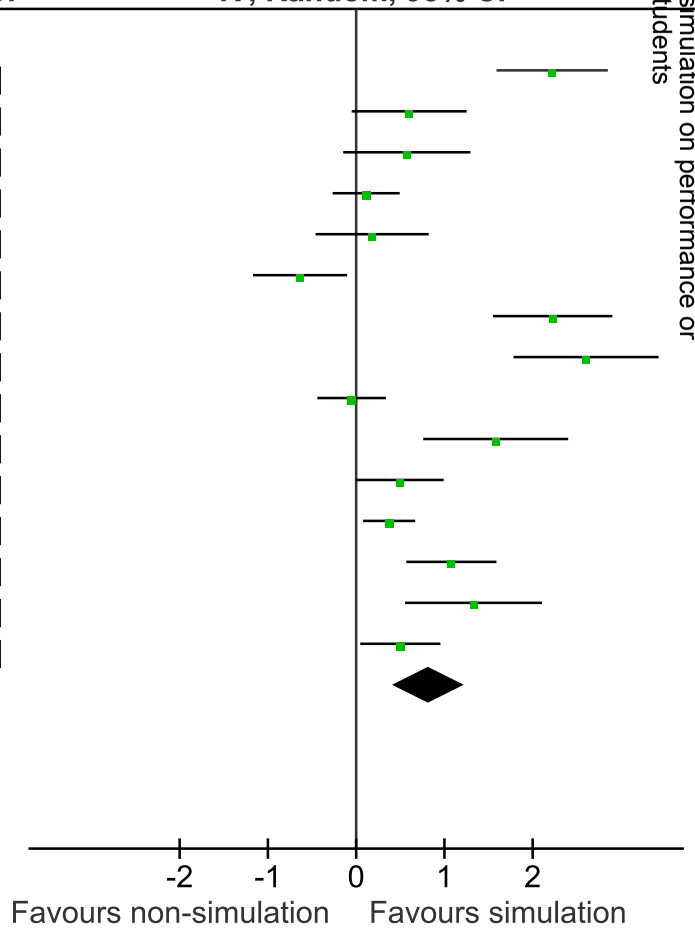


Figure 4: The effectiveness of simulation on performance or knowledge scores in medical students

Figure 5: Funnel plot assessing risk of publication bias

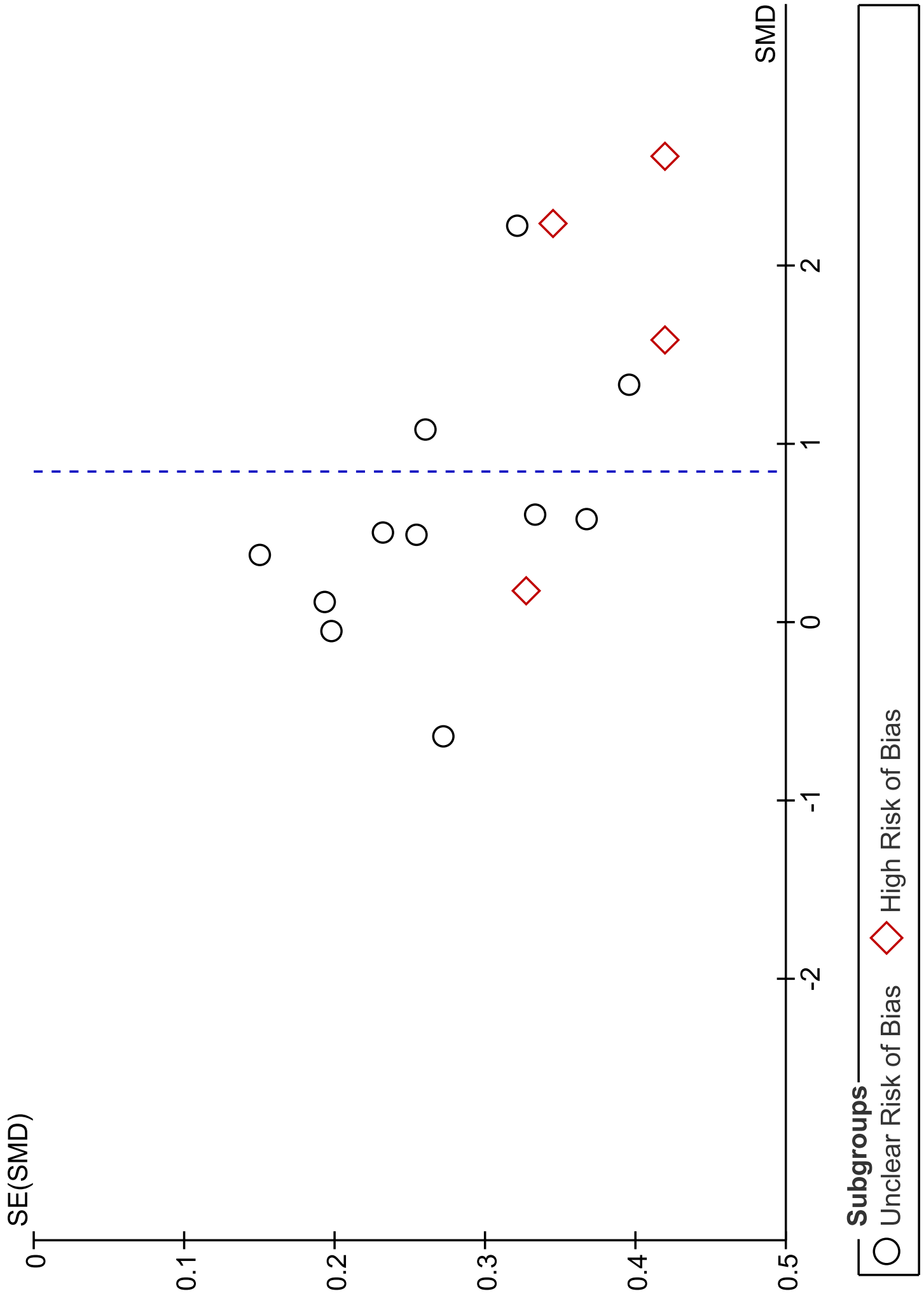


Figure 6: The effectiveness of different types of simulation - direct and indirect analyses.

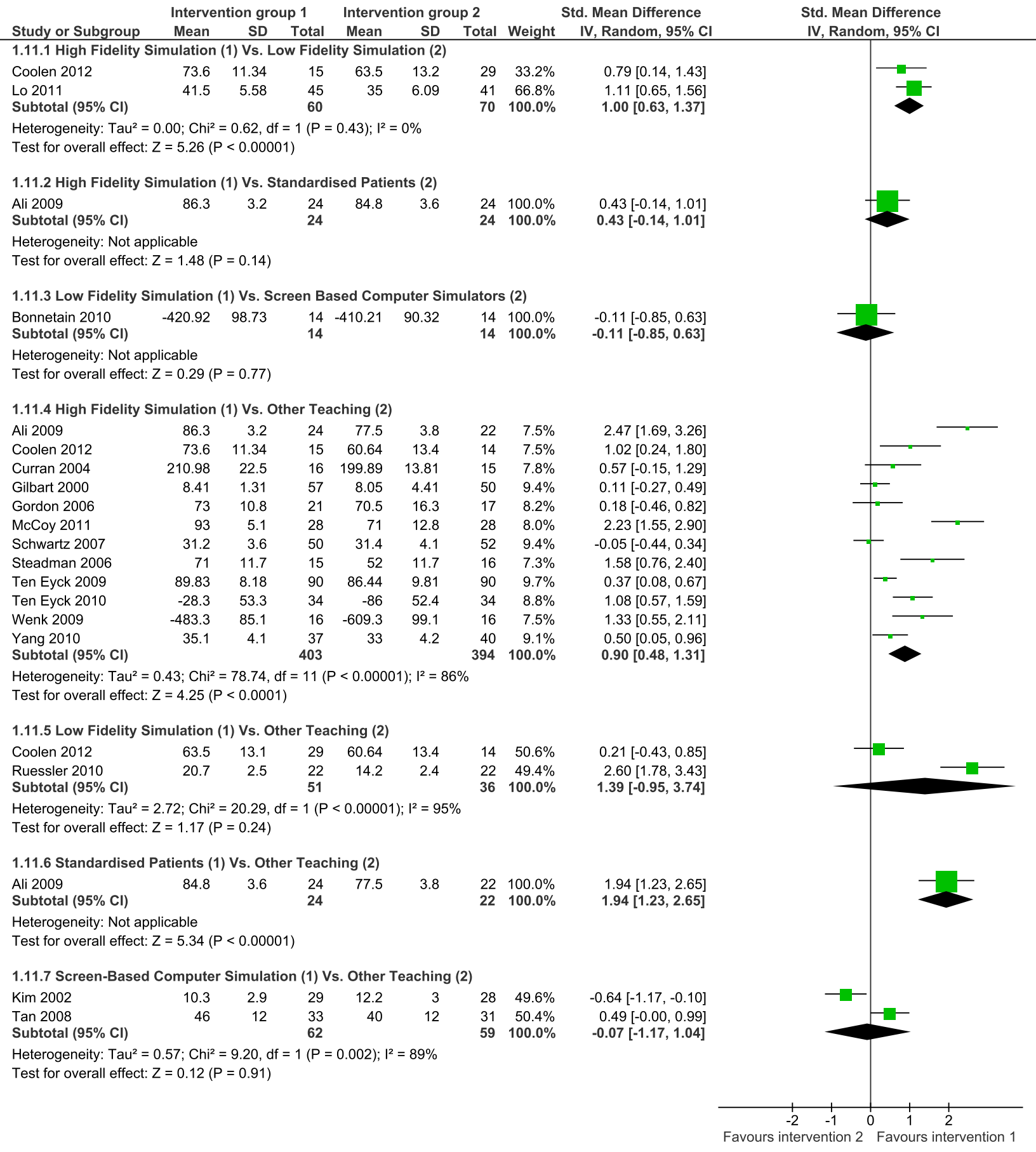


Table 3: GRADE evidence profile

TABLE 3: GRADE EVIDENCE PROFILE

Outcome	Number of Studies	Design	Limitations	Inconsistency	Indirectness	SMD (95% CI)	Quality
Effectiveness of simulation compared to other teaching methods	15	RCT	Serious [1]	Serious [2]	No serious indirectness	0.84 (0.43 to 1.24) [3]	⊕⊕⊕◯ **Moderate** [1 2 3]

SMD = Standardised Mean Difference; CI = Confidence Interval

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

1 Rating downgraded to account for serious risk of bias which may have been present in the studies that were classified as unclear.
2 Rating was downgraded to account for serious inconsistency in our results. The cause of heterogeneity (which was significant) was not detectable through sensitivity analysis or subgrouping. We decided to rate this as 'serious' rather than 'very serious' as the vast majority of studies were in agreement about the direction of effect.
3 Rating was upgraded to account for the large effect size we demonstrated.

PRISMA CHECKLIST

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	**1**
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	**2**
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	**4**
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	**5**
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	**5**
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	**5-6**
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	**6**
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	**Link to Protocol**
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	**7**
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	**7**
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions or simplifications made.	**7**
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	**7**
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	**8**
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	**9**

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	**7**
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	**10**
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	**Figure 2**
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	**Table 1**
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	**Additional Material**
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	**Figure 4**
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	**Figure 4, Table 2**
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	**Figure 3**
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	**Table 2**
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	**18**
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	**20**
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	**21**
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	**23**

Figure 2: Study Flow Diagram

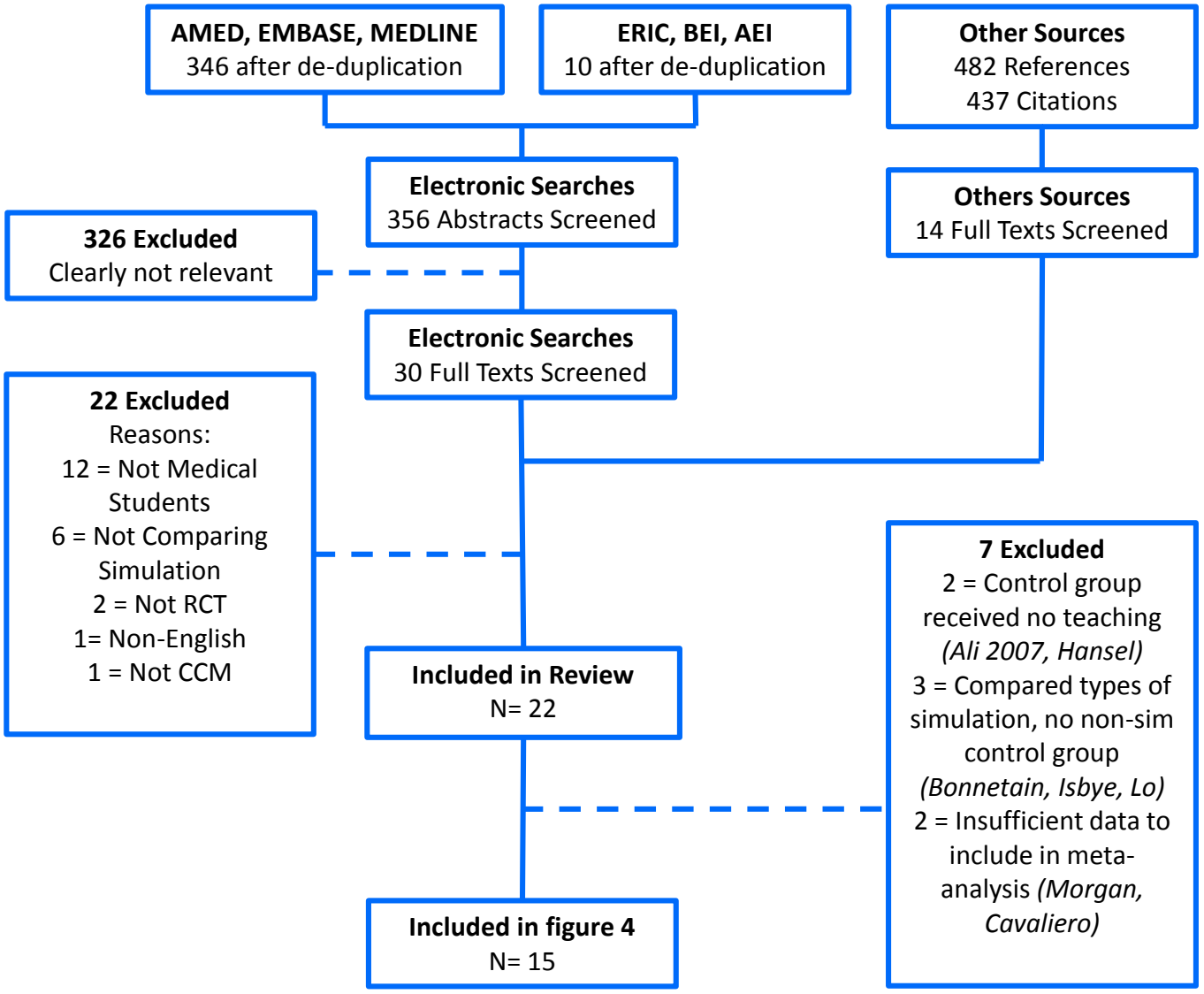


Table 1: Characteristics of Included Studies

TABLE 1: CHARACTERISTICS OF INCLUDED STUDIES

	Sample Size							Intervention			
Trial *(First Author, Year, Country)*	***Total***	*Group 1*	*Group 2*	*Control*	**Participants (Year of study, Setting)**	**Design**	**Clinical Topic**	**Experimental**	**Control**	**Duration (Hours)**	**Outcomes Assessed (Follow-up Time Points Assessed (weeks))**
Ali, 2007, Trinidad and Tobago	**38**	21	-	17	5, -	Parallel	ATLS	SPT	No Teaching	-	MCQ (0), EQ (0)
Ali, 2009, Canada	**70**	24	24	22	5, -	Parallel	ATLS	HFS, SPT	Video/Lecture	-	MCQ (0), EQ (0)
Bonnetain, 2010, France	**28**	14	14	-	2, ED	Parallel	CPR	SBCS, LFS	*No control*	-	OSCE and Time to Action (0)
Cavaleiro, 2009, Portugal	**45**	24	-	21	5, SimLab	Parallel	NR	LFS	Self-Study	0.5	MCQ (0)
Coolen, 2012, Netherlands	**43**	15	14	14	4, SimLab	Parallel	PALS	HFS, LFS	Self-Study	5.0	Sim Checklist* (1), MCQ (2), SE (2)
Curran, 2004, Canada	**31**	16	-	15	3, Campus	Parallel	NR	HFS	Video	-	Skill Checklist § (4months), MCQ (0,4months), SE (0,4months), EQ (4months)
Gilbart, 2000, Canada	**107**	57	-	50	4, -	Parallel	ATLS	HFS	Seminar	2	OSCE (2), PEP Score (2), EQ (2)
Gordon, 2006, US	**38**	38	-	38	3-4, SimLab	Crossover	MI/RAD	HFS	Lecture	1.5	SAQ (0)
Hansel, 2012, Germany	**59**	20	19	20	6, SimLab	Parallel	Sepsis	HFS, CRM‡	No Teaching	10	OSCE* (0), SAGAT* (0); EQ (0)
Isbye, 2008, Denmark	**43**	22	21	-	2, SimLab	Parallel	CPR	VAM, LFS	*No Control*	*5 (min)*	OSCE (0), (3 months)
Kim, 2002, South Korea	**57**	29	-	28	4, -	Parallel	ACLS	SBCS	Self-Study	2.5	SBA (0,1); EQ (0)
Lo, 2011, US	**86**	45	41	-	3, -	Parallel	ACLS	HFS, LFS	*No Control*	3	Skills Checklist ± (3 days and 1 yr); EQ (3 days and 1 yr)
McCoy, 2011, US	**28**	28	-	28	4, SimLab	Crossover	MI/Anaph.	HFS	Lecture	1	Skills Checklist* (0)
Morgan, 2002, Canada	**144†**	92	-	95	5, SimLab	Parallel	MI/Anaph./ Hypoxemia	HFS	Video	1.5	Skills Checklist* (Immediate); SAQ and EQ (Variable 2 to 30 days)
Ruesseler, 2010, Germany	**44**	22	-	22	5, SimLab	Parallel	ACLS/ATLS	LFS	Shadowing	3 *(Days)*	OSCE (0)
Schwartz, 2007, US	**102**	50	-	52	4, SimLab	Parallel	MI/ACLS	HFS	CBD	1	OSCE (0)
Steadman, 2006, US	**31**	15	-	16	4, -	Parallel	SOB	HFS	PBL	0.25	Skill Checklist* (0)
Tan, 2008, Singapore	**64**	33	-	31	3-4, -	Parallel	Anaph.	SBCS	Lecture	1	Skill Checklist* (1), EQ (1)
Ten Eyck, 2009, US	**90**	90	-	90	4, SimLab	Crossover	Chest*/AMS/ SOB/Trauma	HFS	CBD	6	MCQ, EQ (0)
Ten Eyck, 2010, US	**68**	34	-	34	4, SimLab	Parallel	Chest*/AMS	HFS	CBD	6	Time to Action* (0)
Wenk, 2009, Germany	**32**	16	-	16	4, SimLab	Parallel	RSI	HFS	PBL	2	Time to Action* (2), Sim Checklist* (2), MCQ (2), SE (2) EQ (2)
Yang, 2010, Malaysia	**77**	37	-	40	5, -	Parallel	ACLS	HFS	Lecture/Video	2	SBA and EMQ (2 hours)

Clinical Topic: **ATLS** = Advances Trauma Life Support, **CPR** = Cardiopulmonary Resuscitation, **NR** = Neonatal Resuscitation, **PALS** = Paediatric Advanced Life Support, **ACLS** = Advanced Cardiac Life Support, **MI** = Myocardial Infarction, **RAD** = Reactive Airways Disease, **Anaph.** = Anaphylaxis, **SOB** = Shortness of Breath of varying cause, **Chest*** = Chest Pain, **AMS** = Altered Mental State, **RSI** = Rapid Sequence Induction
***Interventions*:** **HFS** = High Fidelity Simulation, **LFS** = Low Fidelity Simulation, **SPT** = Standardized Patients, **CRM** = Crew Resource Management, **VAM** = Voice Advisory Mannequin; **SBCS** = Screen Based Computer Simulation, **PBL** = Problem Based Learning, **CBD** = Case Based Discussion
***Outcome Measures*:** **MCQ** = Knowledge based Multiple Choice Questions, **EQ** = Evaluative Questionnaire, **SE** = Self-reported Efficacy, **SAQ** = Short Answer Questions, **OSCE** = Objectively Structured Clinical Exam, **SAGAT** = Situational Awareness Global Assessment Technique, **SBA** = Single Best Answer Questions, **PEP** = Post-Encounter Probe,* = identical simulator used in test and all students orientated, § = Identical simulator used in test with no orientation, ± = Unclear if same simulator used in assessment or not.
***Other Abbreviations*:** † = Group numbers do not sum, ‡ = group not eligible for review, **ED** = Emergency Department

Table 2: Sub-grouping and sensitivity analyses

TABLE 2: SUB-GROUPING AND SENSITIVITY ANALYSES

Factor	Subgroup	Number of Studies	Number of Participants	Standardised Mean Difference† (95% CI)	p-value for heterogeneity, I^2
Main Analysis	Overall Analysis (No Sub-groups)	15	1000	0.84 (0.43 to 1.24)	<0.001, 89%
Time to Outcome Assessment*	<72 hours	11	761	0.93 (0.39 to 1.47)	<0.001, 91%
	72 hours to 3 months	5	303	0.36 (-0.15 to 0.87)	0.002, 77%
	3 months+	1	31	0.57 (-0.15 to 1.29)	NR
Types of Simulation*	High Fidelity Simulators	12	797	0.90 (0.48 to 1.31)	<0.001, 86%
	Screen Based Simulators	2	121	-0.07 (-1.17 to 1.04)	0.002, 89%
	Low Fidelity Simulators	2	87	1.39 (-0.95 to 3.74)	0.001, 95%
	Standardised Patients	1	46	1.94 (1.23 to 2.65)	NR
Type of Control Group	Lectures, Videos and Seminars	7	443	0.88 (0.26 to 1.50)	<0.001, 89%
	Problem/Case based learning techniques	5	413	0.78 (0.23 to 1.33)	<0.001, 84%
	Self-Study	2	100	-0.03 (-1.24 to 1.18)	0.004, 88%
	Clinical Shadowing	1	44	2.60 (1.78 to 3.43)	NR
Type of Outcome Assessment*	Knowledge Based Assessment	8	539	0.41 (-0.09 to 0.91)	<0.001, 86%
	Performance based Assessment	10	578	1.01 (0.49 to 1.53)	<0.001, 88%
Duration of Simulation Session **	<4 hours	9	564	0.59 (0.09 to 1.09)	<0.001, 88%
	4 to 8 hours	3	291	0.65 (0.20 to 1.11)	0.06, 64%
	8 hours+	1	44	2.60 (1.78 to 3.43)	NR
Year of study **	Junior (Years 1-3)	1	31	0.57 (-0.15 to 1.29)	NR
	Senior (Years 4+)	14	969	0.86 (0.43 to 1.29)	<0.001, 90%
Sensitivity analyses					
Excluding studies judged to have high risk of bias *(Gordon, McCoy, Ruesseler, Steadman)*	Overall analysis	11	831	0.57 (0.19 to 0.95)	<0.001, 85%
Excluding studies with industry funding *(Steadman)*	Overall analysis	14	969	0.79 (0.37 to 1.21)	<0.001, 89%
Excluding studies with crossover design *(Gordon, McCoy, Ten Eyck 2009)*	Overall analysis	12	726	0.83 (0.35 to 1.30)	<0.001, 89%
Excluding studies which used the same simulator in the outcome assessment *(Coolen, Curran, McCoy, Steadman, Tan, Ten Eyck 2010, Wenk)*	Overall Analysis	8	675	0.61 (0.05 to 1.18)	<0.001, 91%
Using p-value to impute the SD *(Steadman)*	Overall analysis	14	969	0.79 (0.37 to 1.21)	<0.001, 89%

†: Random Effects Inverse Variance Meta-analysis. *In some studies outcomes were measures at multiple time points, using multiple techniques, and some studies had multiple intervention arms, so total number of studies and participants do not sum to the total number included. ** Not all studies reported these factors and therefore do not appear in the sub-group analysis.
CI = Confidence Interval; **NR** = Not Relevant

Table 2: Sub-grouping and sensitivity analyses