But does it work with real patients? Caution needed in Health Education England's prioritisation of simulation-based training of new doctors.

Caroline Kamau-Mitchell*

*Birkbeck Business School, Birkbeck, University of London, Malet Street, London, WC1E 7HX, United Kingdom.

Citation

Kamau-Mitchell, C. (2025). But does it work with real patients? Caution needed in Health Education England's prioritisation of simulation-based training of new doctors. *Occupational Psychology Outlook, 4*(1), https://doi.org/10.53841/bpsopo.2025.4.1.10

Abstract

The United Kingdom's Health Education England (HEE) oversees training of new doctors, and it prioritises simulation-based training, but it is unclear whether such training prepares new doctors for clinical procedures involving real patients. To investigate that, this article's aim was to discuss experiments in the field, noting that a review by HEE missed at least eight studies, and misinterpreted two studies as showing positive outcomes when the results were negative or inconclusive. Occupational psychologists who work in staff development in hospitals should therefore be cautious about the HEE's review by reading the original studies themselves. This article discusses why the HEE should evaluate the impact of simulation-based training on patient outcomes, and why it should review studies which test causality (e.g., randomised-controlled trials). Looking back at Kirkpatrick's model about how to evaluate training interventions, occupational psychologists should go beyond the reaction stage (e.g., are doctors enjoying the training?) to the behaviour stage (e.g., does it help them complete clinical procedures with patients?) and the results stage (e.g., does it reduce excess patient deaths in July or August when new doctors start work in hospitals?). Some experiments found that simulation-based training can worsen performance by increasing the duration of clinical procedures, the number of attempts to success and new doctors missing vital steps. Many of the methods are still used today (e.g., in endoscopy, laparoscopy), therefore their efficacy remains questionable. When presented with newer technology such as virtual reality-based training, occupational psychologists should still ask questions about their efficacy and encourage the hospitals to supplement simulation-based training of new doctors with good quality/quantity of clinical supervision and job shadowing.

Keywords: Junior doctors; Medical doctors; Simulation-based training; Training effectiveness; Training evaluation.

Introduction

Many occupational psychologists work in staff development and training in hospitals, and they can influence the kind of training completed by new doctors who have graduated from their degrees but are not yet experienced enough to work with patients without some guidance or supervision. Organisations such as Health Education England (HEE, 2024) are influential in helping occupational psychologists understand what to prioritise in training new doctors because the HEE organises “core medical training,” which in the United Kingdom refers to pre-training phase for doctors who want to specialise in certain areas of medicine. The HEE has made simulation-based training central to its plans and it published a review which claims that this method has positive results (HEE, 2024) but the review missed out on important studies and misinterpreted some. This article presents a critical review of that review, critically considering whether simulation-based training actually makes new doctors competent enough to start working with patients on their own. Whereas the HEE’s review examined studies with a wide range of methodologies, and it did not evaluate patient-based outcomes exclusively, this article discusses randomised-controlled trials and other types of experiments which evaluated new doctors’ work with patients.

Why is it important that simulation-based training should translate to the work that new doctors do with real patients? The issue is important not just as a staff development concern for occupational psychologists to consider but because patient mortality rises after new doctors join the hospital workforce (Young et al. 2011, Jen et al. 2009). Studies show that between 4.3% and 12% more patients die in July or August when new doctors typically start work (Jen et al. 2009) because the changeover reduces average hospital efficiency compared to other months (Young et al. 2011). Doctors report that there is a sense of increased disruption within hospitals (Vaughan, McAllister & Bell, 2011). More urgent or rapid response tasks are logged by nurses during out of hours and clinical procedures take longer to complete (Blakey, Fearn & Shaw, 2013). A systematic review calculated the decline in average hospital efficiency after new doctors join the workforce in July at between -0.3% and -7.2% (Young et al. 2011). The

cumulative drain on hospital staffing increases the risks to patients and this raises the mortality risk, although there may be country differences and newer studies are needed.

Effective training has been proposed as the solution, such as using simulators (Cohen et al. 2013), but the benefits for patients are unclear. For training to be effective as a way of preventing the excess deaths, it should tackle the hospital efficiency problem: it should reduce the time it takes a new doctor to complete a clinical procedure and the time it takes other staff to provide guidance, thus keeping average staff performance nearer to optimum. The problem is the lack of evidence that simulation-based training, which is a popular training method in hospitals, is actually helpful from that point of view. A survey of 763 doctors by Vaughan et al. (2011) asked them for their views about the current changeover process and their responses suggest scepticism for non-ward training methods. Most (84.5%) expressed support for ward-based training methods such as job shadowing, suggesting that working with real patients might be better preparation.

The current article reports findings from a systematic review conducted in 2015, but which remains relevant because several simulation-based training methods used at the time are still used today. Before the review, a pilot (scoping) review using *Google Scholar* revealed that only 6.78% of studies about simulation training in medicine actually involved new doctors and none of the studies had tested patient outcomes. Most studies showed support for simulation training as an effective training method in advanced or specialised medical practice, such as in anaesthesia (e.g., Ashurst et al. 1996; Burtscher et al. 2011; Chopra et al. 1994; Holzman et al. 1995; John et al. 2007; Zausig et al. 2009) and urology (Abdelshehid et al. 2014). Even in these fields, the evidence about the non-patient outcomes of simulation training was mixed in that some studies showed benefits whereas others did not, and reviews in these fields called for studies measuring follow-up clinical data (Ross et al. 2012; Cetti et al. 2010). The studies that did exist about new doctors largely involved non-patient outcomes and they only provided modest support for simulation training, with researchers calling for more experience with real patients (Boots et al. 2009).

One of the limitations of current research in the field is that many studies which claim to show that simulation-based training works tend to rely on feedback from doctors, yet this does not mean that the training was effective, even if they say that they enjoyed or were satisfied with it. This view is consistent with Kirkpatrick's (1960 model of evaluating training, which tells us that we should evaluate it beyond how employees react to it by considering its impact on job performance in terms of behaviour and results. A randomised controlled trial by Kerr et al. (2013) in obstetric-internal medicine showed that simulation training was enjoyed by interns as an educational method, but test performance actually declined from 72.59% before the simulation training to 50.09% afterward and performance fell further to 37.59% one month later. Simulation training as an instructional method also varies in its effectiveness. A meta-analysis of 289 studies by Cook et al. (2013) spanning different health professions showed that it is helpful as an instructional method, but some simulation techniques are more effective than others. There is also the problem of conflating evidence involving experienced doctors (who integrate knowledge from simulation training with their existing knowledge) with evidence from new doctors, who need the training to give them new knowledge or a new opportunity to practice a clinical skill based on theoretical knowledge.

It remains unclear whether knowledge or skills gained from simulation-based training generalise from educational to clinical settings. While simulation training can be beneficial from a cost perspective, occupational psychologists should be asking questions about whether it is beneficial to actual learning and behaviour, and actual work with real patients. They should ask questions about whether the reported outcomes are clinically meaningful, from a patient-outcome perspective, or whether they simply tell us that new doctors enjoyed the training, which is not the same as showing that it is going to reduce errors with patients and excess deaths. There have been discussions about the need for policies and practices in hospitals to be justified by evidence of their clinical usefulness (Kamau, 2015). This article therefore discusses some randomised-controlled trials or similar types of experiments which shed light on the effects of training on new doctors' work with patients.

Methods

Protocol

Figure 1 shows the PRISMA flow diagram with the PRISMA checklist used during the process, based on Moher et al. (2009).

<Insert Figure 1 about here>

Inclusion criteria and information sources

The inclusion criteria required that eligible publications must have used an experimental method such as a randomised controlled or within-subjects design (with pre/post testing) to examine the effects of simulation training on new doctors e.g., Foundation Year 1 doctors in the UK, interns and year 1 residents in the USA. The outcome measure was a quantified indicator of the new doctor's competence while working with real patients. The publications were searched from the *Web of Science* core collection and all included databases such as MEDLINE with no restrictions in country, language, publication type or year. Searches were conducted in 2015 and updated searches were not possible because of lack of funding, although the discussion will discuss the current implications of simulation training methods which continue to be used.

Search terms and study selection

Search terms were chosen to maximise the number of publications screened concerning medical training using simulations in any country. The word 'simulation' was combined with 'training' or international synonyms (e.g. clinical + induction + simulation; clinical + training + simulation; clinical + orientation + simulation; intern + boot + camp) entered with no quotation marks and searched to appear anywhere within the full text or bibliography. All studies selected for the final stage were experimental and they measured the effects of simulation training on

new doctors working with patients.

Data collection process and data Items

Marked publications were stored in the *Web of Science* cache then exported to *EndNote* software at the end of each search session. They were stored in a structured archive and duplicate references were deleted. Abstracts and references were exported from *EndNote* to *Windows* then tabulated. Full-text publications retrieved from university libraries were stored in *Adobe* or XPS format and those from the British Library were scanned and stored in PNG format. The data extracted are summarised in table 1. The quality of each study was then assessed using the criteria summarised in table 2.

Risk of bias in individual studies and across studies

The review gauged the risk of a publication bias against null results. The review assessed whether authors used randomisation to allocate doctors to different training groups and also whether new doctors' performance was determined from objective outcomes (e.g. hospital records) or by "blind" assessors who were unaware of whether or not the doctor being assessed was simulation trained.

Summary measures, synthesis of results and analysis

The measures reviewed concerned the new doctors working with patients, e.g. the duration of procedures, the number of attempts by the new doctor before success at the procedure, the number of complications after the procedure (re-hospitalisation, infections) and the overall quality of the procedure. Each study was synthesised qualitatively and quality was assessed both qualitatively and quantitatively.

Results

Study selection

In the first stage, 7792 publication abstracts were reviewed. The most common reasons for exclusion of a publication were: the lack of data (e.g., editorial and commentary articles), using medical students or doctors who were qualified at an advanced level, using nurses or other health professionals, using qualitative methodology, using non-experimental quantitative methodology (e.g. correlation studies) and lacking patient outcome measures. Of the 7792 publications, 3960 were selected for the next stage and duplicates were deleted. A total of 672 publications underwent full text review and 10 (see table 1) met all eligibility criteria. Figure 1 is a PRISMA flow diagram showing the publication count within each stage of the review.

Study characteristics

Table 1 summarises the methodology used in each eligible study.

<Insert Table 1 about here>

Risk of bias within and across studies

The risk of bias within studies was assessed as low because most studies using randomisation or a suitable alternative and blinded assessment (table 2). Quality was normally distributed across all studies and most (8 out of 10 publications) are high in quality (≥3). The completeness of statistics was not normally distributed because the majority had complete statistics.

<Insert Table 2 about here>

Results of individual studies, synthesis of results and additional analyses

This section will summarise and compare results from studies included within the review to evaluate the effects of simulation-based training on outcomes involving real patients. Britt et al. (2009) found no significant difference between the simulator-trained residents and the control residents in 10 out of 10 performance measures, $p > .05$, including landmark identification, angulation and the ability to cannulate a patient at the first attempt. There was also no significant difference in the incidence of performance errors or complications in patients treated by the residents, including the incidence of infections or arterial puncture, $p > .05$. Giulio et al. (2004) observed no differences in the rate of success with the procedure, the number of attempts, the time taken to complete the procedure, or the need for verbal assistance when working on patients and no complication arose in either group although *p* values were not stated. Finan et al. (2011) similarly found that the residents' performance during real patient intubations (with a success rate of 67.5%) was not significantly differently from the historical control, $p = .06$. The simulator-trained residents' checklist scores for intubating the patients (64.6%) were also significantly worse than those of the historical control (82.5%), $p = .001$.

Similar results were observed by Gaies et al. (2009), who found no significant difference between the simulator-trained group and the control group in many of the skills within the simulator-based or checklist assessments e.g., bag mask skills, venipuncture, success at lumbar puncture, $p > .05$. Gaies et al. (2009) also observed that, within the venipuncture procedures involving real patients, there was no significant difference between the success rate of those who were simulator-trained compared to the control group, $p = .07$. There was additionally no significant difference in success at peripheral intravenous catheter insertion with real patients, $p = .25$, or lumbar puncture, $p = .6$. Hogle et al.'s (2009) study 1 found no significant differences between the simulator-trained and the control group residents. This was in measures of autonomy, bimanual dexterity, depth perception, efficiency or tissue handling while performing the laparoscopic cholecysectomy with real patients, $p > .05$. Hogle et al.'s (2009) study 2, a retrospective experiment, showed that simulator-trained residents presented

worse patient outcomes than residents who were not simulation-trained: they took longer in-room time, $p = .021$, and operative time, $p = .038$.

Madan et al. (1998) found that, comparing the pre-test and post-tests within-groups there were significant improvements in both the simulation-trained and control residents, $p < .02$. Miranda et al. (2007) similarly found no significant difference between residents who attended the simulation training and control group residents in 4 out 5 practices: using a large drape, $p = .14$, wearing a cap, $p = .6$, wearing a gown, $p = .9$, or wearing sterile gloves, $p = .10$. Miranda et al. found no significant change in knowledge the risk of thrombosis, infections and knowledge about complications, $p > .05$, and no significant difference in the incidence of patient complications concerning blood infections, $p = .29$. Supporting the pattern emerging from the other studies, Palter et al. (2001) found that the difference between the baseline and post-test within the simulator-trained group was stable although no p values were stated.

There was some evidence of the beneficial effects of simulation training. Britt et al. (2009) found that the assessors evaluated the simulator-trained residents as having better ability, and presenting better comfort for patients (both $p = .03$) but the limitation is that assessors were not blind to the residents' trained status. Results supportive of simulation training were also found by Palter et al. (2001), who observed that the technical skills of simulator-trained residents' in the operating room were significantly better, $p = .04$. These results were, however, limited by the fact that the new doctors were only allowed to treat real patients after they achieved proficiency on the simulator.

Giulio et al. (2004) found that residents who practiced independently on a simulator completed more diagnostic endoscopies, they required less manual assistance (e.g. retroflexion and duodenum intubation) and they were also evaluated by the instructor more positively, $p < .0001$.These results were, however, limited because the assessors were not blind to the residents' training status and because the beneficial effects observed in 3 outcomes were in the minority when compared to several other non-differing means. Gaies et al. (2009) found that the

control group performed worse than the simulation-trained group in peripheral intravenous catheter insertion, $p < .05$, and the checklist assessment of lumbar puncture, $p < .05$, but these results are limited because the checklist assessments did not involve real patients and because competence at lumbar puncture was rated by assessors who were not blinded to the residents' training status. Madan et al. (1998) similarly found a significant difference between the simulation residents and the control residents in their OSCE performance, with residents who were inducted through simulation performing significantly better. This effect replicated across 2 indicators of interpersonal skills and 2 indicators of content, $p < .05$, but within-group comparisons showed that both the simulation-trained and control group residents improved significantly from pre-test to post-test, $p < .02$.

Mayo et al. (2004) found that 91-100% of the simulation-trained performed airway management steps correctly during real clinical emergencies, suggesting a high level of performance, but this result is limited because of the lack of a control group. Miranda et al. (2007) found one beneficial change in the simulation trained residents: there was a significant improvement in knowledge about arterial puncture risk, $p < .05$, but patient outcome measures showed no significant difference after simulation training. White et al. (2012) found that residents' knowledge significantly improved from the pre-test level of 58% to the post-test level of 70%, $p < .05$. The residents' performance after the simulation training was good: 88% of steps were correctly followed and 62% of the interns performed successfully at the first attempt. However, these results are limited because there was no control group or baseline patient outcome data with which to compare these percentages.

Discussion

The 2015 review found that evidence about the benefits of simulation training on patient outcomes was mixed, contrary to the HEE's (2024) conclusions. This article highlights the fact that the HEE's review was incomplete because at least 8 out of 10 of the studies reported in this article were missing from the HEE's review. Of the two studies that it did include, the HEE

misinterpreted their outcomes. One study misinterpreted by the HEE was that by Britt et al. (2009) where new doctors' patients were followed up to gauge whether they suffered complications (e.g., an arterial puncture or blood infection) and the results showed that simulation training had no significant effect on the number of patient complications which does not justify the HEE concluding that the study showed a positive effect. The other study which the HEE should have interpreted with caution was that by White et al. (2012) where new doctors performed lumbar puncture on paediatric patients in an emergency within 5 months of the training. Although there were improvements, only 48% of the new doctors correctly performed steps involving preparing the supplies and conducting the lumbar puncture with the needle's bevel being parallel to the spinal ligament. It is unclear why the HEE missed 8 of the 10 studies, calling into question whether the HEE should revisit its review and update it to avoid misleading doctors, patients and other stakeholders into believing that simulation-based training is as effective as the HEE claims it is. Moreover, the HEE's simplification of the outcome of each study as "positive" without acknowledging negative results and risks of bias in methods used is concerning. Occupational psychologists should therefore be aware about the pitfalls of the HEE (2024) review, noting that it is not a proper systematic review because it was not reported as such (e.g., using PRISMA guidelines), did not include all the evidence, it over-simplifies the outcomes, and misinterprets some evidence. A related implication for occupational psychologists is to look at the evidence about the efficacy of training methods themselves, by going directly to the literature concerned, or by looking at literature in published, peer-reviewed journals. The HEE should have reported its review methods, databases searched, risks of bias, and given readers more information about why it concluded that a given study had a positive outcome.

Occupational psychologists need to help hospitals understand that the effects of training may not be clear-cut. Rather than saying that the training had a positive impact, it is better to be honest about the complexity of the evidence. In this review, while some experiments found that simulation training is equally or more beneficial than controls, some experiments found no significant benefits and others found that it can worsen new doctors' performance, e.g. in

lengthening procedure time and increasing the need for guidance while working with patients. This does not mean that simulation-based training should stop being used but that it should be used in conjunction with giving new doctors opportunities to work with real patients under supervision as well as opportunities to learn by observation through job shadowing. Training which helps new doctors learn new clinical procedures in “skills laboratories” (a kind of bootcamp in clinical training) can also be helpful in translating to their work with real patients (Kamau, 2014a).

Looking back to Kirkpatrick’s (1960) model about training evaluation, it is important for the HEE and occupational psychologists working on staff development in hospitals to go beyond evaluating the reaction stage (e.g., do doctors enjoy the training?), initial learning (e.g., can doctors perform well on the simulator after the training?) to the behaviour stage (e.g., after the training, do doctors complete clinical procedures on real patients correctly and in the expected time?) and the results stage (e.g., do doctors who complete the training have fewer patient complications and excess deaths, controlling for confounding factors such as patients’ medical conditions and comorbidities?). The HEE’s (2024) review did not distinguish between these types of training outcomes, and it is likely that many studies which found that simulation-based training has positive outcomes concerned the reaction stage (e.g., feedback from doctors) but not the behaviour and results stage.

Globally, occupational psychologists who influence medical training should be aware that there are country differences in the skill level of new doctors therefore training may need to concentrate on some skills more than others. A systematic review found that the average rate of inexperience with a range of clinical skills among new doctors was lowest in the UK (9.15%), followed by New Zealand (18.33%) and South Africa (19.53%) (Kamau, 2014b). I recommend that occupational psychologists collaborate with doctors in determining what good training should involve, ensuring that it covers clinical skills based on need, and promoting training which is evidence-based. Some simulation-based training technologies have changed in the past couple of years such as the advance of virtual reality-based training and 3D printing in

some areas of medicine (Bienstock & Heuer, 2022), but some of the methods in this review are still being used today, which makes this evidence relevant. For example, recent searches of what NHS hospitals use revealed that they still use the simulation training methods used by studies included in this review concerning endoscopy (Giulio et al. 2004), laparoscopy (Hogle et al. 2009), and mannequins (various authors listed in Table 1). That is also true for many hospitals worldwide, and access to new technologies using virtual reality might remain low globally because of cost. Even then, occupational psychologists working in hospitals should assess the new technologies using the same outcomes as those encouraged by this article – these being whether they train new doctors to work with real patients safely, correctly, minimising clinical errors, complications and excess mortality.

Conflicts of interest*:* None.

References

Abdelshehid, C, Quach, S, Nelson, C, Graversen, J, Lusch, A, Zarraga, J, Alipanah, R, Landman, J, and McDougall, EM. (2013). High-fidelity simulation-based team training in urology: evaluation of technical and nontechnical skills of urology residents during laparoscopic partial nephrectomy. *Journal of Surgical Education, 70*, 588-595.

Ashurst, N, Rout, CC, Rocke, DA, and Gouws, E. (1996) Use of a mechanical simulator for training in applying cricoid pressure. *British Journal of Anaesthia, 77*, 468-472.

Bienstock J and Heuer A. (2022). A review on the evolution of simulation-based training to help build a safer future. *Medicine, 101*(25), e29503

Blakey, JD, Fearn, A, and Shaw, DE. (2013). What drives the ‘August effect’? A observational study of the effect of junior doctor changeover on out of hours work. *Journal of the Royal Society of Medicine, 4*, 1-7.

Boots, RJ, Egerton, W, McKeering, H, and Winter, H. They just don't get enough! Variable intern experience in bedside procedural skills. *Internal Medicine Journal, 39*, 222-227.

Britt, RC, Novosel, TJ, Britt, LD, and Sullivan, M. (2009). The impact of central line simulation before the ICU experience. *American Journal of Surgery, 197*, 533-536.

Burtscher, MJ, Manser, T, Kolbe, M, Grote, G, Grande, B, Spahn, DR, and Wacker, J. (2011). Adaptation in anaesthesia team coordination in response to a simulated critical event and its relationship to clinical performance. *British Journal of Anaesthia, 106*, 801–6.

Cetti, RJ, Bissel, L, Singh, R, and Shaw, R. (2010). The urological foot soldier: are we equipping our foundation-year doctors? Annals of the Royal College of Surgeons of England (Supplement), 92, 284–287.

Chopra, V, Gesink, BJ, de Jong, J, Bovill, JG, Spierdijk, J, and Brand, R. Does training on an anaesthesia simulator lead to improvement in performance? *British Journal of Anaesthia, 73*, 293-297.

Cohen, E.R., Barsuk, J.H., Moazed, F., Caprio, T., Didwania, A., McGaghie, W.C., Wayne, D.B. (2013). Making July safer: Simulation-based mastery learning during intern boot camp. *Academic Medicine, 88*, 233–239

Cook, DA, Hamstra, SJ, Brydges, R, Zendejas, Szostek, JH, Wang, AT, Erwin, P, and Hatala, R. (2013). Comparative effectiveness of instructional design features in simulation-based education: Systematic review and meta-analysis. *Medical Teacher, 35*, e867–e898.

Finan, E, Bismilla, Z, Campbell, C, Leblanc, V, Jefferies, A, and Whyte, HE. (2011). Improved procedural performance following a simulation training session may not be transferable to the clinical environment. *Journal of Perinatology, 32*(7), 539-544.

Gaies, MG, Morris, SA, Hafler, JP, Graham, DA, Capraro, AJ, Zhou, J, Landrigan, CP, and Sandora, TJ. (2009). Reforming procedural skills training for paediatric residents: a randomized, interventional trial. *Pediatrics, 124*, 610-619.

Giulio, ED, Fregonese, D, Casetti, T, Cestari, R, Chilovi, F, D'Ambra, G, and Fave, GD. (2004). Training with a computer-based simulator achieves basic manual skills required for upper endoscopy: a randomized controlled trial. *Gastrointestinal Endoscopy, 60*,196-200.

HEE (2024) Workforce training and education: simulation. Health Education England, retrieved from https://www.hee.nhs.uk/our-work/simulation

Hogle, NJ, Chang, L, Strong, VEM., Welcome, AOU., Sinaan, M, Bailey, R, and Fowler, DL. Validation of laparoscopic surgical skills training outside the operating room: a long road. *Surgical Endoscopy, 23*, 1476-1482.

Holzman, RS, Cooper, JB, Gaba, DM, Philip, JH, Small, SD, and Feinstein, D. (1995). Anesthesia crisis resource management: real-life simulation training in operating room crises. *Journal of Clinical Anesthia, 7*, 675-687.

Jen, M.H. Bottle, A., Majeed, A., Bell, D., and Aylin, P. (2009). Early in-hospital mortality following trainee doctors' first day at work. *PLoS One, 4*, e7103-e7103.

John, B, Suri, I, Hillermann, C, and Mendonca, C. (2007). Comparison of cricothyroidotomy on manikin vs. simulator: a randomised cross-over study. *Anaesthesia, 62*, 1029-1032.

Kamau, C. (2014a). Effects of experimental inductions for newly qualified doctors on competence at clinical procedures. *Clinical Medicine, 14*, 380-385.

Kamau, C. (2015). Should clinical experience be a precondition for a job in NHS management? *Journal of the Royal Society of Medicine, 108*, 120–121

Kamau, C. (2014b). Systematic review of new medics' clinical task experience by country. *Journal of the Royal Society of Medicine Open, 5*, 2054270414525373.

Kerr, B, Hawkins, TL, Herman, R, Barnes, S, Kaufmann, S, Fraser, K, and Ma, IWY. (2013). Feasibility of scenario-based simulation training versus traditional workshops in continuing medical education: a randomized controlled trial. *Medical Education Online, 18*, 21312.

Kirkpatrick, D. L. (1960). Techniques for evaluating training programs: Part 3-Behavior. *Journal of ASTD, 14*(1), 13-18.

Madan, AK, Caruso, BA, Lopes, JE, and Gracely, EJ. (1998). Comparison of simulated patient and didactic methods of teaching HIV risk assessment to medical residents. *American Journal of Preventive Medicine, 15*, 114-119.

Mayo, PH, Hackney, J. E., Mueck, J. T., Ribaudo, V., and Schneider, R. F. (2004). Achieving house staff competence in emergency airway management: Results of a teaching program using a computerized patient simulator. *Critical Care Medicine, 32*, 2422-2427.

Miranda, JA, Trick, WE, Evans, AT, Charles-Damte, M, Reilly, BM, and Clarke, P. (2007). Firm-based trial to improve central venous catheter insertion practices. *Journal of Hospital Medicine, 2*, 135-142.

Moher D, Liberati A, Tetzlaff J, and Altman, DG.(2009). The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med, 6*, e1000097.

Murray, WB, Good, ML, Gravenstein, JS, van Oostrom, JH, and Brasfield, WG. (2002). Learning about new anesthetics using a model driven, full human simulator. *Journal of Clinical Monitoring and Computing, 17*, 293-300

Palter, VN, Grantcharov, T, Harvey, A, MacRae, HM. (2001). Ex vivo technical skills training transfers to the operating room and enhances cognitive learning: a randomized controlled trial. *Annals of Surgery, 253*, 886-889.

Ross, AJ, Kodate, N, Anderson, JE, Thomas, L, and Jaye, P. (2012). Review of simulation studies in anaesthesia journals, 2001–2010: mapping and content analysis. *British Journal of Anaesthia, 109*, 99–109.

Vaughan, L., McAlister, G., and Bell, D. (2011). 'August is always a nightmare': results of the Royal College of Physicians of Edinburgh and Society of Acute Medicine August transition survey. *Clinical Medicine, 11*, 322–6.

White, ML, Jones, R, Zinkan, L, and Tofil, NM. (2012). Transfer of simulated lumbar puncture training to the clinical setting. *Pediatric Emergency Care, 28*, 1009-1012.

Young, JQ, Ranji, SR, Wachter RM, Lee, CM, Niehaus, B, and Auerbach, AD. (2011). "July Effect": impact of the academic year-end changeover on patient outcomes. *Annals of Internal Medicine, 155*, 309-315.

Zausig, YA, Grube, C, Boker-Blum, T, Busch, CJ, Bayer, Y, Sinner, B, Zink W, Schaper, N, and Graf, BM. (2009). Inefficacy of simulator-based training on anaesthesiologists' non-technical skills. *Acta Anaesthia Scandinavia, 53*, 611–619.

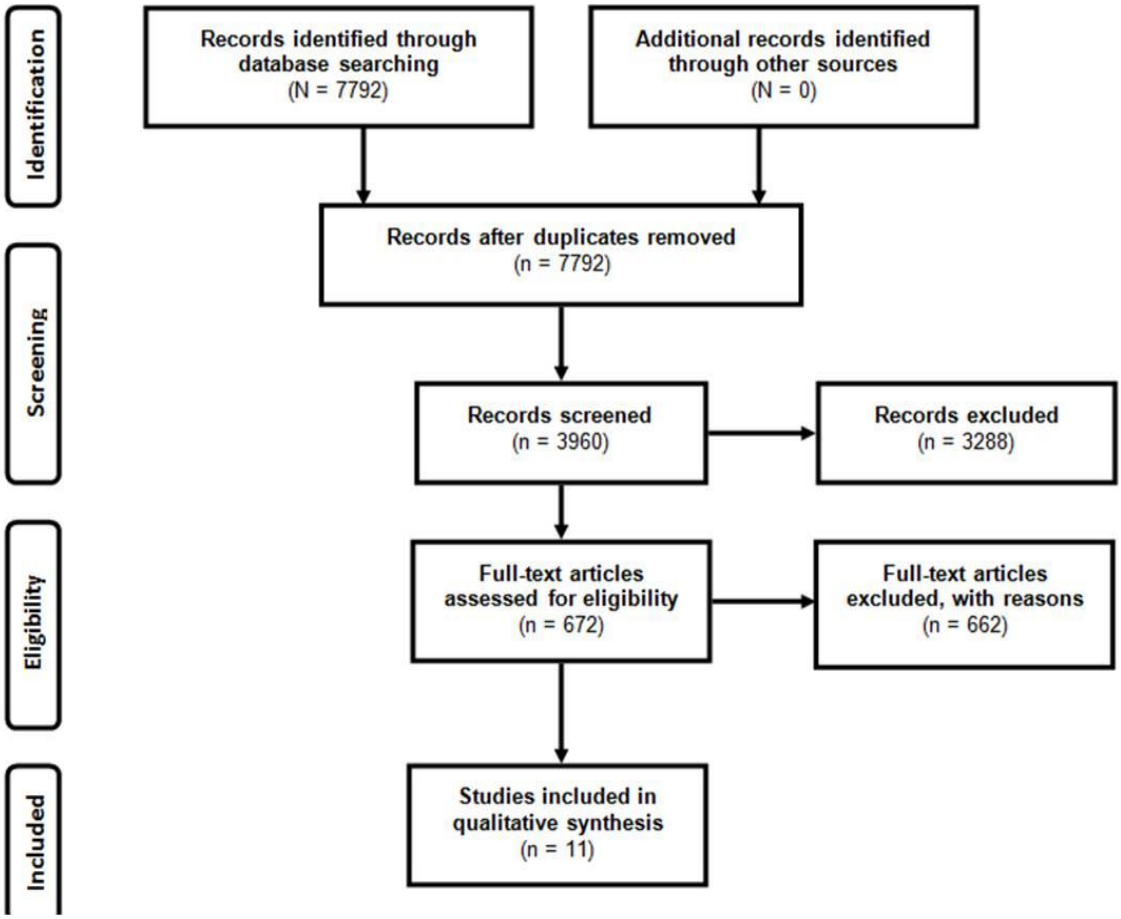

Figure 1: PRISMA flow diagram

|

Table 1: Methodological characteristics of the studies

What type of study was it?	What was the sample size?	What was the simulation training?	How long was the simulation training?	What outcome of the simulation training was measured?	How was the outcome measured?	What did the study find?
Randomised experiment by Britt et al. (2009)	34 junior residents (n=13 simulation training with didactic session; 21 control group: didactic session only).	Simulation training using *Central Line Man*	Training was for an unspecified duration	After the simulation training, a new doctor inserted a central line into a patient (e.g. positioning, needle insertion and the resident being able to place the line). After this, the new doctor's patients were followed up to gauge whether they suffered complications after this e.g., an arterial puncture or blood infection.	By recording the verdict of a senior (a fellow or critical care surgeon) about the new doctor's competence at this procedure after the simulation training. By counting the number of complications that happened among the patients treated by the new doctor after the simulation training.	Results showed that simulation training had no significant effect on number of patient complications.
Randomised experiment by Giulio et al. (2004)	22 junior residents (n=11 didactic session then simulation training; n=11 control group: didactic session only)	Simulation training using *GI Mentor* software and a mannequin	Training for 10 hours	After the simulation training, a new doctor conducted a diagnostic endoscopy on patients.	By logging the new doctor's failure or success at each procedure after the simulation training as well as the number of attempts they took before success, autonomy in successfully completing the procedure, and the procedure's duration. By recording the verdict from an instructor about the new doctor's work after simulation training.	The results were inconclusive.

What type of study was it?	What was the sample size?	What was the simulation training?	How long was the simulation training?	What outcome of the simulation training was measured?	How was the outcome measured?	What did the study find?
Non-randomised experiment (historical control) by Finan et al. (2011)	13 1st year pediatric residents simulation trained compared with a historical control group: no training.	Simulation training using a mannequin	Training for 2 hours	In the 2 months after the simulation training, data were collected when the new doctor intubated a patient's trachea.	By recording whether or not the new doctor did this successfully after no more than two attempts. By an observer assessing the new doctor's performance.	Simulation training produced significantly worse outcomes in some measures, and in other measures it had no significant effect.
Randomised experiment by Gaies et al. (2009)	38 interns (n=18 simulation trained; n=20 control group: observed seniors at work)	Simulation training using high fidelity bag masking ventilator; anatomic model emitting fluid etc.	Training varied in duration by skill type.	After the simulation training, a new doctor performed the following procedures on patients over 6 months: lumbar puncture, peripheral intravenous catheter insertion and venipuncture on pediatric patients.	By logging the number of attempts the new doctor took in each procedure and whether or not they completed the procedure successfully.	Simulation training had no significant effect.
Randomised experiment by Hogle et al. (2009)	12 residents (n=6 simulator trained; n=6 control group)	Simulation training using *LapSim*	Training for 2 sessions a week over a 1 month period.	After the simulation training, a new doctor conducted laparoscopic cholecysectomies on patients.	By a surgeon who watched a video recording of the new doctor working on the patient. The surgeon blind to the new doctor's training type.	Simulation training had no significant effect.
Retrospective; non-randomised experiment by Hogle et al. (2009)	140 residents (n=80 simulator trained; n=60 control group)	Simulation training using *LapSim*		After the simulation training, a historical control group was used to compare the incidence of complications among	By comparing the incidence of complications among patients treated and procedure duration (operative time and in-room	Simulation training produced significantly worse procedure duration outcomes.

What type of study was it?	What was the sample size?	What was the simulation training?	How long was the simulation training?	What outcome of the simulation training was measured?	How was the outcome measured?	What did the study find?
				patients treated by the new doctors after the simulation training, including mortality.	time).	
Randomised experiment by Madan et al. (1998)	12 junior residents (n=6 simulation trained; n=6 control group: didactic session)	Simulation training using HIV-positive patients	Training for 1 hour, divided into three sessions.	After the simulation training, a new doctor's ability to communicate effectively with the patient in investigating a urinary tract infection.	By a primary care physician who was blinded to the new doctor's training condition. The physician used an OSCE format to assess the new doctor by watching a video recording of their interaction with the patient.	Simulation training produced a significantly better outcome.
Randomised experiment by Mayo et al. (2004)	50 house officers (divided into immediate versus delayed simulator training groups)	Simulation training using a mannequin	Training 30-40 minutes	After the simulation training, a new doctor's ability to work with another new doctor on airway management of a cardiac arrest patient.	By an attending physician assessing how well the new doctor set up equipment and provided bag mask ventilation.	The results were not conclusive because there was no control group.
Non-randomised experiment by Miranda et al. (2007)	150 residents (n=40 simulation trained; n=110 ward trained)	Simulation training using a mannequin	Training for 15 hours divided into 6 sessions	After the simulation training, the new doctor worked on patients needing catheterisation.	By a physician and a radiologist who examined radiographs and other patient records to assess whether the patient treated by the new doctor suffered any complication as a result, e.g. These patients were assed to determine	Simulation training had no significant effect.

What type of study was it?	What was the sample size?	What was the simulation training?	How long was the simulation training?	What outcome of the simulation training was measured?	How was the outcome measured?	What did the study find?
					whether they suffered different , e.g., pneumothorax or thromboembolic complications, blood infections, re-hospitalisation of the patient or a visit to emergency within 30 days of having been treated by the new doctor.	
Randomised experiment by Palter et al.(2001)	18 residents (n=9 simulator demonstration then simulation training to proficiency; n=9 simulator demonstration only)	Simulation training using a synthetic abdominal wall model	Training for 3 hours split into two sessions less than 3 weeks apart.	After the simulation training, a new doctor surgically closed a patient's abdominal wall.	By a member of staff in the surgical theatre who was blinded to the new doctor's training group, watched, and used a global rating scale to evaluate them.	Simulation training produced significantly better outcomes.
Within-subjects experiment by White et al. (2012)	21 junior residents (all simulation trained)	Simulation training using a neonatal mannequin	Training for 1 hour and additional simulation practice time	After the simulation training, a new doctor performed lumbar puncture on a pediatric patient in an emergency within 5 months of the training.	By an assessor who used a checklist to assess the new doctor after the simulation training. The checklist items included angling the needle at 15^{o}, a bevel parallel to the spinal ligament, and success at the first attempt.	The results were not conclusive because there was no control group, and only 48% of the new doctors correctly performed steps involving preparing the supplies and

What type of study was it?	What was the sample size?	What was the simulation training?	How long was the simulation training?	What outcome of the simulation training was measured?	How was the outcome measured?	What did the study find?
						conducting the lumbar puncture with the needle's bevel being parallel to the spinal ligament.

Table 2: Study quality assessment

Study	Confound absent	Control group used	Blinding used	All P values stated	Sample size adequate	Summary of quality
Britt et al. (2009)	No	Yes	No	No	No	∑ = 1 Low
Giulio et al. (2004)	Yes	Yes	No	No	No	∑ = 2 Low
Finan et al. (2011)	Yes	Yes	No	Yes	No	∑ = 3 High
Gaies et al. (2009)	Yes	Yes	No	Yes	Yes	∑ = 4 High
Hogle et al. (2009)	Yes	Yes	Yes	Yes	No	∑ = 4 High
Hogle et al. (2009)	Yes	Yes	Yes	Yes	Yes	∑ = 4 High
Madan et al. (1998)	Yes	Yes	Yes	Yes	No	∑ = 4 High
Mayo et al. (2004)	No	No	No	Yes	Yes	∑ = 2 Low
Miranda et al. (2007)	Yes	No	Yes	Yes	Yes	∑ = 4 High
Palter et al.(2001)	No	Yes	Yes	Yes	No	∑ = 3 High
White et al. (2012)	No	No	Yes	Yes	Yes	∑ = 3 HIgh
Skewness and *SE*	-.66(.66)	-.66(.66)	-.21(.66)	-1.92(.66)	.21(.66)	-.86 (.66)
Kurtosis and *SE*	-1.96(1.28)	-1.96(1.28)	-2.44(1.28)	2.04(1.28)	-2.44(1.28)	-.26 (1.28)
Mean and *SD*	.64(.50)	.64(.50)	.55(.52)	.81(.40)	.45(.52)	3.09(1.04)