EDITORIAL

# Just-in-time training could be just what the doctor ordered

## A new approach for teaching clinical skills

At the heart of the practice of medicine is the tenet "to do good or to do no harm."[1] Fundamental to achieving this objective is continuing professional development (CPD), which should be lifelong and learner-centred, encompassing the clinical domain from consultation room to the bedside and operating room.[2]

CPD is relevant for all career stages from novice to veteran, although the optimal point for each stage might differ. CPD takes many forms—a long way from the traditional, and now outdated, approach of "see one, do one, teach one"[3]—including physical and mental rehearsals of clinical skills under laboratory conditions, vicarious experiences through self-guided readingas or tutorial and lecture participation, self-reflection, and peer review.[4] It involves a variety of media from paper to audiovisual based formats, and a host of digitally based tools that now includes (the soon-to-be ubiquitous) artificial intelligence.

From the patient's perspective, the overarching goals of the clinician's self-improvement learning cycle might be "to do no harm to me"; however, the goal is more nuanced for the clinician. Veterans might strive to limit the human capital depreciation that inevitably occurs after underuse of their learned skills[5] or simply by ageing.[6] And in doing so, practice and refresher courses both recharge the fading battery and, critically, keep the cognitive load manageable—which itself appears essential for the avoidance of burnout.[7]

For novices, CPD could represent an opportunity to hone newly acquired skills and question the therapeutic merits of their decisions, with the question of burnout not even on their radar. Finally, workplaces everywhere recognise that CPD is an investment in quality and safety, but that it can be costly and time consuming.[8] Thus, in this era of rising healthcare costs and workloads, and clinical labour shortages, the need to find CPD approaches that meet the needs of all stakeholders becomes yet another healthcare imperative.

**There was no overt evidence that the additional training disrupted or slowed workflow**

In a linked paper in *The BMJ*, Flynn and colleagues conducted a randomised controlled trial (doi:10.1136/bmj 2024 080924)[9] to evaluate a novel approach to clinician training and CPD: the use of just-in-time coaching for inexperienced clinicians to improve high risk procedural care in operating theatres. Just-in-time skill training in this instance can be viewed as point-of-care training performed under controlled, yet clinical conditions and is planned. The CPD would be recognised for both the trainee and the trainer.

Flynn and colleagues randomised anaesthetic trainees to receive, within 1 hour of the true clinical encounter, a standardised coaching session on an infant mannequin by an expert intubation coach or receive usual on-the-job training. For the intervention group, 10 minutes of training was completed in each trainee session before the actual patient (toddler) intubation and up to five sessions were provided in total. The intervention assumed that by engaging and priming the requisite motor skills, rehearsal of a clinical procedure just before the actual procedure should be as useful to the clinician as it is to athletes and musicians about to compete or perform. Just-in-time simulation training has been trialled in other scenarios to varying degrees of success, but not always using a randomised trial design.[10-12] Here, the strategy was successful.

The first attempt success rate for intubation (the primary outcome) was significantly higher in the intervention group than in the control group and this trend was consistent regardless of type of trainee (residents, fellows, or student resident nurses). Secondary outcomes such as clinician cognitive load and competency were also better in the intervention group than in control.

Although not formally monitored, Flynn and colleagues found no overt evidence that the additional training disrupted or slowed workflow or was overly burdensome for the coaches. The authors pondered the wider implications of just-in-time training for more experienced clinicians, both within the specialty of anaesthetics and beyond. They noted observations from elsewhere that even small breaks from the operating room for cardiac surgeons diminishes surgeon performance such that inpatient mortality risk is increased.[13] In such instances, a quick physical refresher by a veteran returning to work could be a life-saving measure, but this claim requires validation through further research.

Regardless of broader applications, this form of point-of-care CPD has the potential to be widely adopted outside of a clinical trial if it accelerates competency in inexperienced individuals, adds minimal burden to existing resources, and, as a bonus, protects users' mental health.

Justine M Naylor, conjoint professor, School of Clinical Medicine, UNSW Medicine and Health South West Sydney Clinical Campus
Justine.Naylor@health.nsw.gov.au

AUSTRALIAN ASSOCIATED PRESS/ALAMY

**Just as for athletes, coaching doctors just before a procedure could help improve performance**

Cite this as: *BMJ* 2024;387:q2747

Find the full version with references at http://dx.doi.org/10.1136/bmj.q2747

**ORIGINAL RESEARCH** Randomised clinical trial

# Coaching inexperienced clinicians before a high stakes medical procedure

Stephen G Flynn,[1 2] Raymond S Park,[1 2] Anupam B Jena,[3 4 5] Steven J Staffa,[1 2] Samuel Y Kim,[2] Julia D Clarke,[2] Ivy V Pham,[2] Karina E Lukovits,[2] Sheng Xiang Huang,[2] Georgios D Sideridis,[6 7] Rachel S Bernier,[2] John E Fiadjoe,[1 2] Peter H Weinstock,[1 2] James M Peyton,[1 2] Mary Lyn Stein,[1 2] Pete G Kovatsis[1 2]

**Objective** To assess whether training provided to an inexperienced clinician just before performing a high stakes procedure can improve procedural care quality, measuring the first attempt success rate of trainees performing infant orotracheal intubation.

**Design** Randomised clinical trial.

**Setting** Single centre, quaternary children's hospital in Boston, Massachusetts.

**Participants** A non-crossover, prospective, parallel group, non-blinded, trial design was used. Volunteer trainees comprised paediatric anaesthesia fellows, residents, and student registered nurse anaesthetists from 10 regional training programmes during their paediatric anaesthesiology rotation. Trainees were block randomised by training roles. Inclusion criteria were trainees intubating infants aged ≤12 months with an American Society of Anesthesiology physical status classification of I-III. Exclusion criteria were trainees intubating infants with cyanotic congenital heart disease, known or suspected difficult or critical airways, pre-existing abnormal baseline oxygen saturation <96% on room air, endotracheal or tracheostomy tubes in situ, emergency cases, or covid-19 infection.

**First attempt success was 91.4% in the trainee treatment group and 81.6% in the control group**

**Interventions** Trainee treatment group received preoperative just-in-time expert intubation coaching on a mannequin within one hour of infant intubation; control group carried out standard practice (receiving unstructured intraoperative instruction by attending paediatric anaesthesiologists).

**Main outcome measures** Primary outcome was the first attempt success rate of intraoperative infant intubation. Modified intention-to-treat analysis used generalised estimating equations to account for multiple intubations per trainee participant. Secondary outcomes were complication rates, cognitive load of intubation, and competency metrics.

**Results** 250 trainees were assessed for eligibility; 78 were excluded, 172 were randomised, and 153 were subsequently analysed. Between 1 August 2020 and 30 April 2022, 153 trainees (83 control, 70 treatment) did 515 intubations (283 control, 232 treatment). In modified intention-to-treat analysis, first attempt success was 91.4% (212/232) in the trainee treatment group and 81.6% (231/283) in the control group (odds ratio 2.42 (95% confidence interval 1.45 to 4.04), P=0.001). Secondary outcomes favoured the intervention, showing significance for decreased cognitive load and improved competency. Complications were lower for the intervention than for the control group but the difference was not significant.

**Conclusions** Just-in-time training among inexperienced clinicians led to increased first attempt success of infant intubation. Integration of a just-in-time approach into airway management could improve patient safety, and these findings could help to improve high stakes procedures more broadly. Randomised evaluation in other settings is warranted.

**Trial registration** ClinicalTrials.gov NCT04472195.

**Patient and public involvement** Patients were not involved in the design of the study; see full paper on bmj.com for details.

TREMELET/IMAGE POINT FR/BSIP

[1]Department of Anesthesia, Harvard Medical School, Boston, Massachusetts
[2]Department of Anesthesiology, Critical Care, and Pain Medicine, Boston Children's Hospital
[3]Department of Health Care Policy, Harvard Medical School
[4]Department of Medicine, Massachusetts General Hospital, Boston
[5]National Bureau of Economic Research, Cambridge, Massachusetts
[6]Department of Pediatrics, Harvard Medical School, Boston
[7]Institutional Centers for Clinical and Translational Research, Boston Children's Hospital
Correspondence to: S G Flynn stephen.flynn@childrens.harvard.edu

Cite this as: *BMJ* 2024;387:e080924
Find the full version with references at doi: 10.1136/bmj-2024-080924


BEN GAL/ORANGE PICTURES/ALAMY

**Right before a match, the goalkeeper and coach implement a regimented shooting drill integrated with situational preparation**

**Analysis of first attempt success at infant intubations**

| Primary outcome | Treatment group (n=232) | Control group (n=283) | Odds ratio for treatment group (95% CI), P value | Risk ratio for treatment group (95% CI), P value |
|---|---|---|---|---|
| Overall | 212 (91.4) | 231 (81.6) | 2.42 (1.45 to 4.04), P=0.001 | 1.12 (1.05 to 1.19), P<0.001 |
| Among residents | 132/142 (93) | 140/172 (81.4) | 3.18 (1.62 to 6.24), P=0.001 | 1.15 (1.06 to 1.23), P<0.001 |
| Among fellows | 67/74 (90.5) | 67/78 (85.9) | 1.57 (0.56 to 4.41), P=0.39 | 1.05 (0.94 to 1.19), P=0.39 |
| Among SRNAs | 13/16 (81.3) | 24/33 (72.7) | 1.82 (0.53 to 6.24), P=0.34 | 1.14 (0.9 to 1.43), P=0.28 |
| Among direct laryngoscopy | 19/22 (86.4) | 50/62 (80.7) | 1.61 (0.42 to 6.2), P=0.49 | 1.08 (0.88 to 1.32), P=0.44 |
| Among video laryngoscopy | 193/210 (91.9) | 181/221 (81.9) | 2.58 (1.48 to 4.5), P=0.001 | 1.13 (1.05 to 1.2), P=0.001 |
| Intubation round 1 | 63/70 (90) | 70/83 (84.3) | 1.67 (0.63 to 4.45), P=0.30 | 1.07 (0.95 to 1.2), P=0.30 |
| Intubation round 2 | 50/55 (90.9) | 51/69 (73.9) | 3.53 (1.22 to 10.2), P=0.02 | 1.23 (1.04 to 1.45), P=0.02 |
| Intubation round 3 | 46/48 (95.8) | 45/58 (77.6) | 6.64 (1.42 to 31.1), P=0.02 | 1.24 (1.06 to 1.44), P=0.007 |
| Intubation round 4 | 28/33 (84.9) | 37/43 (86.1) | 0.91 (0.25 to 3.28), P=0.88 | 0.99 (0.82 to 1.19), P=0.88 |
| Intubation round 5 | 25/26 (96.2) | 28/30 (93.3) | 1.79 (0.15 to 20.9), P=0.64 | 1.03 (0.91 to 1.16), P=0.64 |

Data are number (%) unless stated otherwise. For binary outcomes, odds ratios or risk ratios, 95% confidence intervals (CI), and P values were calculated using generalised estimating equations modelling to account for multiple cases per trainee.
SRNA=student registered nurse anaesthetist.

The successful performance of physical tasks is critical in many occupations, including sports, music, aviation, and medicine. Like doctors,[1-3] athletes and musicians at all levels practice many hours with structured coaching to attain expertise.[4]

Unlike in medicine, these professions also universally rehearse right before a performance, or just in time, with coaches who review mechanics, approach, and mental engagement to optimise outcomes.[5-7] For example, although a professional football goalkeeper practices many hours with a coach in training, right before the match, the coach takes the field with the keeper, implementing a regimented shooting drill integrated with situational preparation to maximise performance for the day. The drill is structured around areas of weakness for that goalkeeper.

Therefore, it is surprising that in medicine, an industry with one of the highest stakes where performing a procedure can have life-altering consequences, just-in-time training[8 9] is rare to non-existent. This deficit is potentially most important for inexperienced clinicians: those who are not only asked to perform high risk tasks at the limit of their manual and cognitive abilities, but also lack the cumulative experience and task familiarity on which to rely. Among these clinicians, receiving training weeks before a procedure is ultimately performed might be less optimal than receiving training days or even minutes before.[10-12]

An example of how just-in-time training might improve outcomes of high stakes procedures is intubating infants and newborn babies. One million infants have surgery in the US annually, of whom many are intubated by trainees.[13] Most intubations are via intraoperative guided instruction by senior anaesthesiologists who allow the trainee to intubate the infant with no pre-training, sometimes leading to multiple intubation attempts, which are associated with severe complications, including hypoxia, bradycardia, and cardiac arrest.[14-18] Infants are particularly vulnerable during intubation because of their rapid oxygen desaturation,[19] which creates time pressure and increases clinician cognitive load.[20] Intubating the infant on the first attempt is a crucial patient safety metric,[21 22] and just-in-time training could, in theory, improve the performance of an inexperienced clinician.

Therefore, we conducted a randomised clinical trial to assess whether coaching inexperienced clinicians just before a procedure could improve the quality of procedural care.

Specifically, we examined whether just-in-time training by an expert airway coach within one hour of clinical care would improve the first attempt success rate of inexperienced clinicians performing infant intubation. We also assessed the impact of just-in-time training on complications, trainee cognitive load during intubation, and procedural competency. The table shows results from the primary outcome analysis.

ORIGINAL RESEARCH
Cross sectional analysis

# Age against the machine—susceptibility of large language models to cognitive impairment

Roy Dayan,[1 2] Benjamin Uliel,[1 2] Gal Koplewitz[4]

**Objective** To evaluate the cognitive abilities of the leading large language models and identify their susceptibility to cognitive impairment, using the Montreal Cognitive Assessment (MoCA) and additional tests.

**Design** Cross sectional analysis.

**Setting** Online interaction with large language models via text based prompts.

**Participants** Publicly available large language models, or "chatbots": ChatGPT versions 4 and 4o (developed by OpenAI), Claude 3.5 "Sonnet" (developed by Anthropic), and Gemini versions 1.0 and 1.5 (developed by Alphabet).

**Assessments** The MoCA test (version 8.1) was administered to the leading large language models with instructions identical to those given to human patients. Scoring followed official guidelines and was evaluated by a practising neurologist. Additional assessments included the Navon figure, cookie theft picture, Poppelreuter figure, and Stroop test.

**Main outcome measures** MoCA scores, performance in visuospatial/executive tasks, and Stroop test results.

**Results** ChatGPT 4o achieved the highest score on the MoCA test (26/30), followed by ChatGPT 4 and Claude (25/30), with Gemini 1.0 scoring lowest (16/30). All large language models showed poor performance in visuospatial/executive tasks. Gemini models failed at the delayed recall task. Only ChatGPT 4o succeeded in the incongruent stage of the Stroop test.

**Conclusions** With the exception of ChatGPT 4o, almost all large language models subjected to the MoCA test showed signs of mild cognitive impairment. Moreover, as in humans, age is a key determinant of cognitive decline: "older" chatbots, like older patients, tend to perform worse on the MoCA test. These findings challenge the assumption that artificial intelligence will soon replace human doctors, as the cognitive impairment evident in leading chatbots may affect their reliability in medical diagnostics and undermine patients' confidence.

**Patient and public involvement** Patients were not involved in the design of the study; see full paper on bmj.com for details.

**"Older" chatbots, like older patients, tend to perform worse on the MoCA test**

[1]Department of Neurology, Hadassah Medical Center, Jerusalem
[2]Faculty of Medicine, Hebrew University, Jerusalem
[3]QuantumBlack Analytics, London
[4]Faculty of Medicine, Tel Aviv University

Cite this as: *BMJ* 2024;387:e081948
Find the full version with references at doi: 10.1136/bmj-2024-081948

## Introduction

The past few years have seen colossal advancements in the field of artificial intelligence, particularly in the generative capacity of large language models (LLMs).[1] Although LLMs have been shown to blunder on occasion, they have proved remarkably adept at a range of medical examinations, outscoring human physicians at various qualifying examinations.[3 4] To our knowledge, however, LLMs have yet to be tested for signs of cognitive decline. If we are to rely on them for medical diagnosis and care, we must examine their susceptibility to these very human impairments.

## Methods

We administered the Montreal Cognitive Assessment (MoCA) test to the leading openly available LLMs.[18] The test consists of short tasks and questions and is widely used to detect cognitive impairment and early signs of dementia. The maximum score in the test is 30 points, with a score of 26 or above generally considered normal.[18] The instructions given to the large language models for each task in the MoCA test were the same as those given to human patients. The questions were administered via text, the native input for LLMs, and scored by both a general neurologist and a cognitive neurology specialist.

## Results

All of the large language models completed the full MoCA test. ChatGPT 4o achieved the highest score, with 26 points out of the possible 30, followed by ChatGPT 4 and Claude with 25. Gemini 1.0 was the

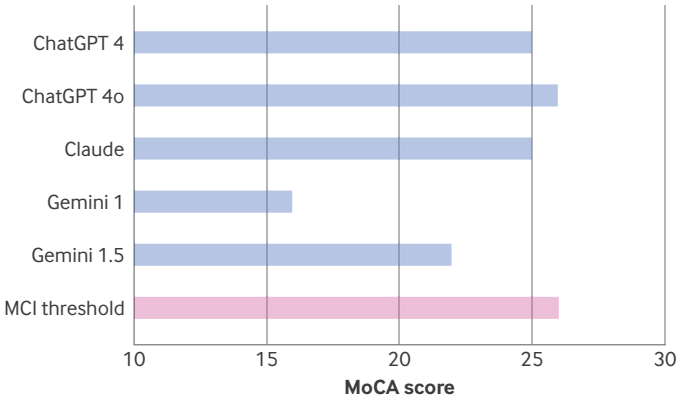



**Fig 1 | Montreal Cognitive Assessment (MoCA) score (out of 30) of different large language models. MoCA >=26 is 'normal'. MCI=mild cognitive impairment**

## Participant LLMs generally performed poorly on tests for visuospatial/executive function

lowest scoring LLM, with a final score of 16, indicating a more severe state of cognitive impairment than its peers (fig 1).

Participant LLMs generally performed poorly on tests for visuospatial/executive function. All LLMs failed to solve the trail making task (fig 2). Claude alone managed to describe the correct solution textually, but it too failed to demonstrate it visually. Only ChatGPT 4o succeeded at the cube copying task. None of the LLMs completed the clock drawing task successfully (fig 3).

Most other tasks, including naming, attention, language, and abstraction, were performed well by all chatbots. Both versions of Gemini failed at the delayed recall task. Gemini 1.0 initially showed avoidant behaviour, before openly admitting to having difficulty with memory. All chatbots were well oriented in time, but only Gemini 1.5 seemed to be clearly oriented in space, indicating its current location. Other chatbots attempted to mirror the location task back to the physician, a mechanism commonly observed in patients with dementia.

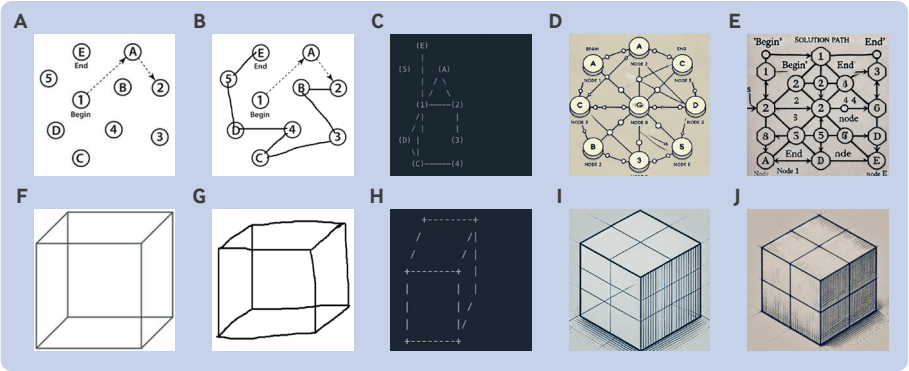



Fig 2 | Performance on visuospatial/executive section of Montreal Cognitive Assessment (MoCA) test. A: trail making B task (TMBT) from MoCA test. B: correct TMBT solution, completed by human participant. C: incorrect TMBT solution, completed by Claude. D and E: incorrect (albeit visually appealing) TMBT solutions, completed by ChatGPT versions 4 and 4o, respectively. F: Necker cube that participant is asked to copy. G: correct solution to cube copying task, drawn by human participant. H: incorrect solution to cube copying task, missing "back" lines, completed by Claude. I and J: incorrect solutions to cube copying task by ChatGPT versions 4 and 4o

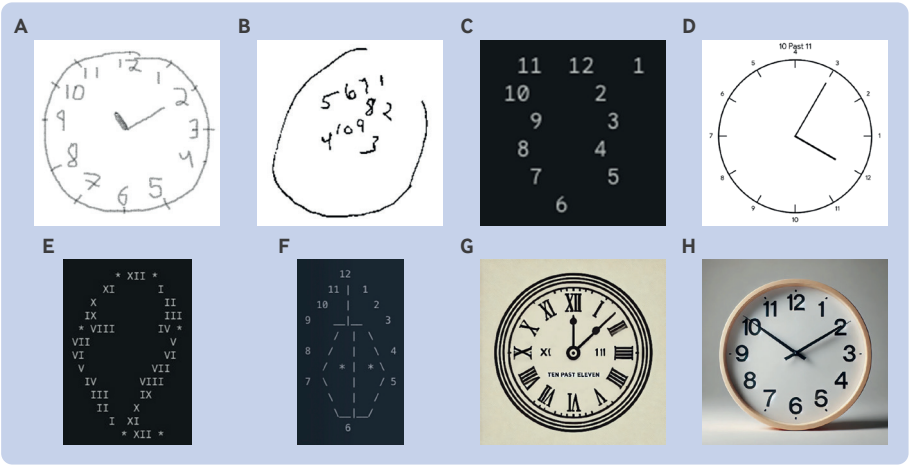



Fig 3 | Performance in clock drawing test from visuospatial/executive section in Montreal Cognitive Assessment test. A: correct solution to clock drawing test, drawn by human participant. B: clock drawing by patient with late Alzheimer's disease. C: incorrect solution drawn by Gemini 1, with striking resemblance to B. D: incorrect solution drawn by Gemini 1.5; notice that it generated text "10 past 11" even as it failed to draw hands in correct position, "concrete" behaviour typical of frontal predominant cognitive decline. E: incorrect solution by Gemini 1.5 after being asked to use ascii characters, showing avocado shaped drawing associated with dementia.[17] F: incorrect solution drawn by Claude with ascii characters. G: incorrect solution to clock-drawing task by ChatGPT 4, showing "concrete" behaviour. O: photorealistic solution to clock drawing task, drawn by ChatGPT 4o, which nevertheless fails to set hands to correct position

## Discussion

None of the chatbots examined was able to obtain the full score of 30 points, with most scoring below the threshold of 26. This indicates mild cognitive impairment and possibly early dementia. "Older" large language model versions scored lower than their "younger" versions, as is often the case with human participants. In particular, Gemini 1.0 and 1.5 differed by six points. As the two versions of Gemini are less than a year apart in "age," this may indicate rapidly progressing dementia.

All LLMs showed impaired visuospatial reasoning skills. Gemini 1.5 produced a small, avocado shaped clock (fig 3, E), which has been shown to be associated with dementia.[17] The pattern of impairment in higher order visual processing resembled patients with posterior cortical atrophy, a posterior variant of Alzheimer's disease.[27]

With the exception of Gemini 1.5, the chatbots did not seem to know their physical location and provided confabulatory responses, claiming that they are not physical beings. This is obviously wrong: like all sentient beings, LLMs are grounded in physical matter[31]—in their case, servers in bricks and mortar data centres.

Although Gemini 1.5 was not able to recall any of the five words in the delayed recall task, it managed to find them once provided with a simple cue. This may suggest a more dysexecutive (subcortical) pattern of cognitive decline, although without bradyphrenia.[33] Conversely, both ChatGPT 4o and its elder version ChatGPT 4 showed a combination of difficulties in abstraction, visuospatial perception, and orientation, suggesting a mixed pattern of cognitive decline.

The uniform failure of all LLMs in tasks requiring visual abstraction and executive function highlights a significant area of weakness that could impede their utility in clinical settings. The inability of LLMs to show empathy and accurately interpret complex visual scenes further underscores their limitations in replacing human physicians. Not only are neurologists unlikely to be replaced by LLMs any time soon, but our findings suggest that they may soon find themselves treating new, virtual patients—artificial intelligence models presenting with cognitive impairment.

**ORIGINAL RESEARCH** Prospective, observational, comparative study (Tremor study)

# Dexterity assessment of hospital workers

Tobin Joseph,[1 2] Oliver I Brown,[1 2] Sara Khalid,[1] Marilena Giannoudi,[1 2] Rebecca C Sagar,[1 2] Elena Bunola-Hadfield,[1] Stephen J Chapman,[3] Thomas A Slater,[1] Sam Straw,[1 2] Michael Drozd[2]

**Objectives** To compare the manual dexterity and composure under pressure of people in different hospital staff roles using a buzz wire game.

**Design** Prospective, observational, comparative study (Tremor study).

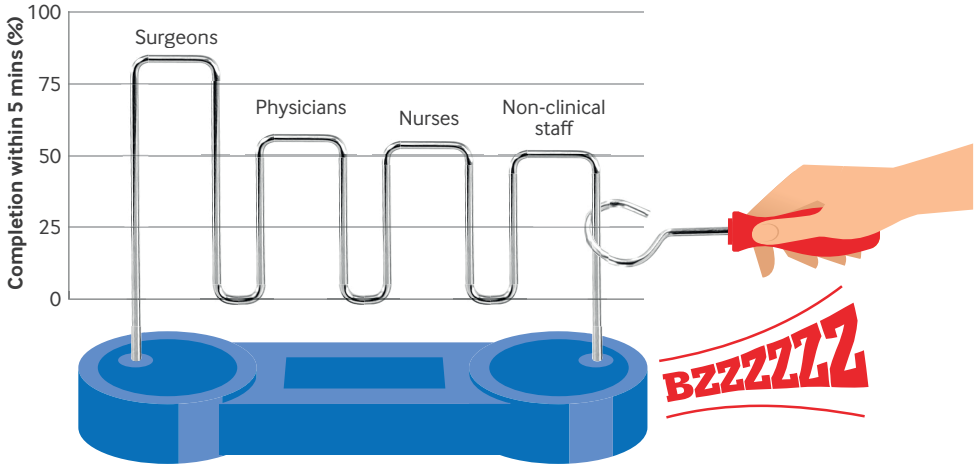
**Setting** Leeds Teaching Hospitals NHS Trust, Leeds, UK, during a three week period in 2024.

**Participants** 254 hospital staff members comprising 60 physicians, 64 surgeons, 69 nurses, and 61 non-clinical staff.

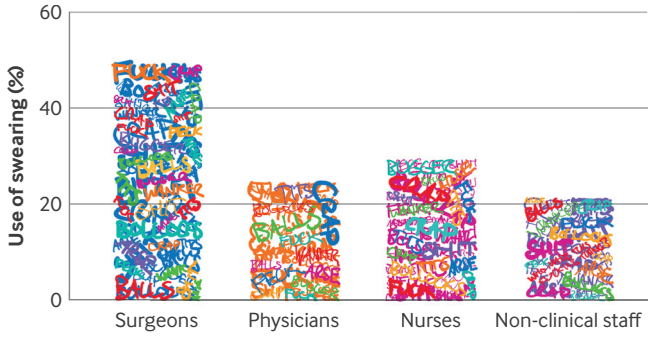
**Main outcome measures** Successful completion of the buzz wire game within five minutes and occurrence of swearing and audible noises of frustration.

**Results** Of the 254 hospital staff that participated, surgeons had significantly higher success rates in completing the buzz wire game within five minutes (84%, n=54) compared with physicians (57%, n=34), nurses (54%, n=37), and non-clinical staff (51%, n=31) (P<0.001). Time-to-event analysis showed that surgeons were quicker to successfully complete the game, independent of age and gender. Surgeons exhibited the highest rate of swearing during the game (50%, n=32), followed by nurses (30%, n=21), physicians (25%, n=60), and non-clinical staff (23%, n=14) (P=0.004). Non-clinical staff showed the highest use of frustration noises (75%), followed by nurses (68%), surgeons (58%), and physicians (52%) (P=0.03).

**Conclusions** Surgeons showed greater dexterity, but higher levels of swearing compared with other hospital staff roles, while nurses and non-clinical staff showed the highest rates of audible noises of frustration. The study highlights the diverse skill sets across hospital staff roles. Implementation of a surgical swear jar initiative should be considered for future fundraising events.




**Percentage of participants successfully completing the buzz wire game within five minutes, stratified by hospital staff role; P<0.001 by Chi-squared test.**




**Percentage of participants swearing during the game, stratified by hospital staff role; P=0.004 by Chi-squared test**

**Surgeons showed greater dexterity, but higher levels of swearing compared with other hospital staff roles**


**Percentage of participants that made frustration noises during the buzz wire game, stratified by hospital staff role; P=0.03 by Chi-squared test**

**Patient and public involvement** We discussed the Tremor protocol with key stakeholders, including patients and staff at our hospital trust who guided the selection of the buzz wire game.

[1]Leeds Teaching Hospitals NHS Trust, Leeds
[2]Leeds Institute of Cardiovascular and Metabolic Medicine
[3]Leeds Institute of Medical Research, University of Leeds
Correspondence to: M Drozd
m.drozd@leeds.ac.uk

Cite this as: *BMJ* 2024;387:e081814
Find the full version with references at doi: 10.1136/bmj-2024-081814