research

Appraisal of systematic reviews of prognostic studies p 77

Skin tone and pulse oximeter accuracy p 78

Interventions for ankle fractures p 80

Access to general practice p 82

RESEARCH METHODS AND REPORTING Quality appraisal tool

Critical appraisal tool for systematic reviews of prognostic factor studies (AMSTAR-PF)

Henry ML, O'Connell NE, Riley RD, et al

Cite this as: *BMJ* 2025;391:e085718
Find this at doi: 10.1136/bmj-2025-085718

The ability to predict the onset or natural history of an illness, or how people may respond to a treatment, guides clinical decision making. These predictions are commonly based on prognostic factors: clinical, patient, or societal variables that are identified as being predictive of a certain future outcome. Prognostic factor research has increased across fields, with a subsequent increase in the number of systematic reviews of prognostic factors studies. Understanding the quality of such prognostic factor reviews is essential for confidence in their findings, but there is no quality appraisal instrument to specifically assess systematic reviews of prognostic factor studies. A measurement tool to assess systematic reviews of prognostic factor studies, AMSTAR-PF, has been developed to fill this gap.

Summary of key differences between AMSTAR 2 and AMSTAR-PF with selected examples

Aspect of tool	Key changes in AMSTAR-PF	Critical examples
Content	Added questions or items for prognostic factor studies and review quality	Question 7c: added question on obtaining prognostic factor effect estimates
		Question 8a: added question about the process of assessing risk of bias
		Question 9a: added question detailing the approach to synthesis
		Question 14: added question on the certainty of findings
	Adapted elements of existing questions to be more relevant to prognostic factors	Question 1: substituted the PICOTS acronym for PICO
		Question 3: stipulated types of prognostic factor studies for inclusion
	Revised signalling points: additions, removals, and modifications	Question 2a: added the requirement for a publicly available protocol
		Question 4: added requirement for the full search strategy to be presented, and recommended 12 months (not 24 months) for the search time frame
		Questions 5, 7a, 8a: added signalling about the plan for resolution of disagreements
		Question 8b: modified the risk of bias assessment to be based on QUIPS
		Question 9a: added specific synthesis guidance specific to prognostic factors
Structure and style	Changed question answering options	All questions: AMSTAR-PF has yes; probably yes; probably no; no; and for some questions, not applicable; AMSTAR 2 had response options of yes, no, partial yes (for some questions), and equivalent of not applicable
	Changed signalling point answering options	All signalling points: AMSTAR-PF has yes, probably yes, probably no, no, and for some questions, not applicable; AMSTAR 2 had checkboxes only
	Reordered certain questions to assist with a logical flow	Moved excluded studies after the screening, as opposed to after data extraction (questions 5-6 in AMSTAR-PF; questions 5 and 7 in AMSTAR 2) Grouped together questions about meta-analysis and data synthesis (questions 9a, 9b, 10 in AMSTAR-PF; questions 11 and 15 in AMSTAR 2)
	Developed an auto-populating spreadsheet version of the tool, alongside traditional document forms and guidance notes	
Overall rating	Modified method for arriving at an overall judgment of quality for the review; removed the concept of critical domains	
Guidance notes	Developed de novo detailed guidance notes for all aspects of the tool	

AMSTAR 2=a measurement tool to assess systematic reviews, version 2; AMSTAR-PF=a measurement tool to assess systematic reviews of prognostic factor studies; PICOTS=population, index prognostic factor, comparator prognostic factors, outcome, timing, setting; PICO=patient, population or problem, intervention, comparison, outcome; QUIPS=quality in prognosis studies.

Pulse oximetry in people with darker skin tones

ORIGINAL RESEARCH Measurement and diagnostic accuracy study

PRIYA SUNDRAM

The impact of skin tone on performance of pulse oximeters used by NHS England COVID Oximetry @home scheme

Martin DS, Doidge JC, Gould D, et al

Cite this as: *BMJ* 2026;392:e085535
Find this at doi: 10.1136/bmj-2025-085535

Study question What is the impact of skin tone on the measurement and diagnostic accuracy of five fingertip pulse oximeters provided by the National Health Service for use at home in the NHS England COVID Oximetry @home scheme?

Methods 903 critically ill adults admitted to intensive care units who were screened for or enrolled into an ongoing trial evaluating oxygen therapy were enrolled into this study. Pulse oximetry derived peripheral oxygen saturation (SpO_2) measurements were compared with paired arterial oxygen saturation (SaO_2) measurements from arterial blood analysed by co-oximetry. Skin tone was objectively measured using a handheld spectrophotometer and presented as individual typology angle. Diagnostic accuracy for identifying SaO_2 ≤92% was assessed by false negative and false positive rates for SpO_2 using thresholds of ≤92% and ≤94% and the area under the receiver operating characteristic curve, and by the presence of occult hypoxaemia (SaO_2 <88% with SpO_2 >92%).

COMMENTARY Current devices may overestimate oxygen saturation measurement

Pulse oximetry is one of the most widely used medical technologies worldwide, yet it performs less accurately for people with darker skin.[1] This inequity requires urgent action. In their study, Martin and colleagues[2] provide strong prospective evidence that pulse oximeters overestimate oxygen saturation in people with darker skin tones.

Evidence of this bias has accumulated for decades—first noted in 1990 and rediscovered during the covid-19 pandemic.[3 4] These earlier studies were retrospective and relied on routinely collected data, leaving room for scepticism about measurement quality, timing, and whether race or skin tone explained the observed differences.

The prospective cohort study by Martin and colleagues[2] addresses these questions. This study of skin tone and pulse oximeter accuracy evaluated five pulse oximeters used in the National Health Service (NHS) England COVID Oximetry @home scheme.[5] The investigators paired simultaneous pulse oximeter readings with arterial blood gas measurements across 24 intensive care units and measured skin tone objectively using spectrophotometry. They found that oxygen saturation measurements were falsely raised in patients with darker skin tones, which could result in missed hypoxaemia. Clinicians and policymakers should now confront the implications of these findings and identify strategies to mitigate harm.

Pulse oximeters are foundational to clinical assessment from the home to the intensive care unit, informing triage, decisions to prescribe oxygen therapy, and treatment thresholds. During the covid-19 pandemic, for instance, patients with darker skin were often sicker when arriving at hospital or intensive care.[6 7] Because eligibility for dexamethasone in covid-19 treatment depended on the presence of hypoxaemia, falsely raised readings effectively increased the treatment threshold,[8] potentially restricting access to a treatment that reduces mortality by almost 20%.[9]

The implications extend far beyond covid-19. Pulse oximeters guide millions of decisions each day—from home monitoring to emergency response to anaesthesia. Inaccuracies may delay treatment across a wide range of conditions, including cardiorespiratory emergencies, sickle cell crises, and chronic respiratory failure. The problem is even more consequential in lower resource settings globally, where pulse oximetry often remains the only means to assess oxygen saturation. This study raises two questions: how did this happen, and what must change now?

Lack of diversity

How did this happen? The root cause is not clinician error but a failure of device design and regulation.[10] Early pulse oximeter studies were flawed, relying

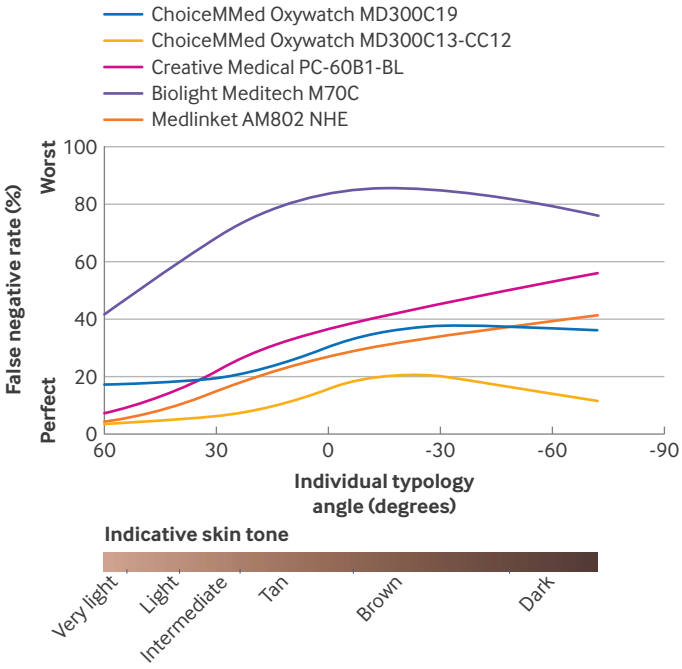
Thomas S Valley
thomas.valley@cuanschutz.edu
Andrew W Fogarty
See bmj.com for author details

Study answer and limitations A total of 11 018 paired SpO_2-SaO_2 measurements were analysed. All tested pulse oximeters overestimated at lower values and underestimated at higher values of SaO_2. On average, SpO_2 readings were 0.6-1.5 percentage points higher for patients with darker skin tone (individual typology angle −44°) than for those with lighter skin tone (46°). At both SpO_2 thresholds assessed, false negative rates increased with darker skin tones; the proportion of SpO_2 measurements ›94% despite a paired SaO_2 ≤92% ranged from 5.3 to 35.3 percentage points higher for patients with darker skin tones than for those with lighter skin tones (7.6-62.2% *v* 1.2-26.9%, rate ratio 2.3-7.1). By contrast, false positive rates decreased with darker skin tones. A key limitation was that the study was conducted in critically ill patients, which may limit the generalisability of the findings.

What this study adds Using an objective measure of skin tone to prospectively evaluate pulse oximeter accuracy in a large cohort of patients, small variations in bias translated into substantial differences in false positive and false negative rates for detecting hypoxaemia.

Funding, competing interests, and data sharing Funded by National Institute for Health and Care Research Health Technology Assessment Programme. No competing interests declared. To request access to data, please visit: https://www.icnarc.org/data-services/access-our-data/

Trial registration ClinicalTrials.gov NCT05481515.



Variation in false negative rate of SpO_2 (peripheral oxygen saturation) ≤92% to detect SaO_2 (arterial oxygen saturation) ≤92% across range of skin tones for each pulse oximeter. False negative rate: failure of pulse oximeter to correctly identify hypoxaemia

on small groups of healthy volunteers, predominantly with light skin. Regulatory standards failed to require diversity in testing or transparent reporting by skin tone. Early evidence—reported by Jubran and Tobin in 1990—was a warning that went unheeded.[3] Our collective blind faith in pulse oximeters, combined with lax oversight, allowed this inequity to persist for more than three decades.

What do we do now? Addressing the pulse oximeter problem requires three integrated steps: technological innovation, cautious interpretation, and stronger surveillance.

Firstly, innovation must accelerate. Despite manufacturers having a clear commercial and ethical incentive to design accurate devices across the full spectrum of skin tones, progress is slow. New medical technology will need to replace older equipment, which may be costly and could limit universal adoption.

Secondly, medical education and clinical practice must adapt. Even when better devices become available, millions of existing pulse oximeters will remain in use worldwide. Clinicians must recognise the limitations of current devices and interpret readings for patients with darker skin with care and caution. Despite widespread discussion, this message has not reached all frontline settings.[11 12]

Better regulation needed

Thirdly, surveillance and transparency are essential after devices reach the market. Martin and colleagues studied device accuracy in critically ill patients—a pragmatic choice given the incidence of hypoxaemia and the availability of arterial blood gases. However, their findings likely underestimate the real world implications, where device quality and user experience vary widely. Regulators should mandate real world testing of pulse oximeters, especially in community and home settings, and make those data publicly available. The US Food and Drug Administration has proposed new guidance requiring diversity in validation cohorts and reporting by skin tone.[13] Regulators should go further by mandating monitoring after devices reach the market to ensure the ongoing reliability of vital medical equipment that is frequently used.

Martin and colleagues showed that pulse oximeters perform differently depending on skin tone, and the potential clinical implications are clear. Regulation must now catch up with science: inclusive validation, transparent data, and continuous oversight should become non-negotiable standards for medical devices.

Clinicians, meanwhile, should interpret oxygen saturation within the clinical context, integrating patient symptoms, clinical trajectory, and awareness of device limitations. The goal is not to abandon pulse oximetry but to understand its limits and make it equitable, ensuring that the technology designed to measure oxygen does not itself perpetuate inequalities in those who receive it.

Cite this as: *BMJ* 2026;392:s37

Find the full version with references at http://dx.doi.org/10.1136/bmj.s37

A step forward for ankle fracture management

ORIGINAL RESEARCH Randomised non-inferiority clinical trial

Cast immobilisation versus surgery for unstable lateral malleolus fractures (SUPER-FIN)

Kortekangas T, Lehtola R, Leskelä H-V, et al

Cite this as: *BMJ* 2026;392:e085295
Find this at doi: 10.1136/bmj-2025-085295

Study question Can cast immobilisation achieve outcomes comparable to surgery for unimalleolar Weber B ankle fractures with a congruent mortise on initial radiography but deemed unstable by external rotation stress testing?

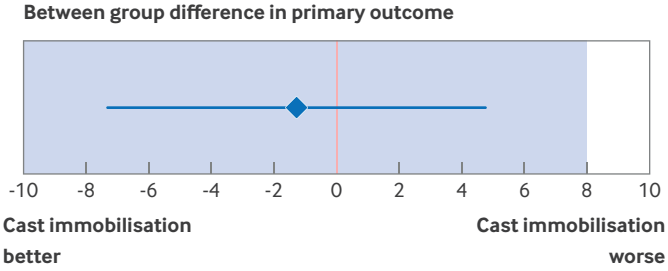
Methods This randomised, pragmatic, non-inferiority trial was conducted at a specialist university hospital trauma centre in Finland between January 2013 and July 2021. Skeletally mature patients aged 16 years and older with a unimalleolar Weber B fibula fracture with a congruent mortise on initial static radiography underwent standard fluoroscopic external rotation stress testing. Fractures showing medial clear space widening were classified as unstable. Participants with a congruent but unstable ankle mortise (n=126) were randomly assigned to receive either standard cast immobilisation for six weeks (n=62) or surgery (open reduction and internal plate fixation) followed by cast immobilisation for six weeks (n=64). The primary outcome was the Olerud-Molander Ankle Score (range 0-100, with higher scores indicating better function and fewer symptoms) at two years, with a predefined non-inferiority margin of −8 points. Secondary outcomes were ankle function, pain, health related quality of life, range of motion, radiographic findings, and adverse events.

Study answer and limitations At two years, the mean Olerud-Molander Ankle Score was 89 in the cast group and 87 in the surgery group (mean difference 1.3 points, 95% confidence interval −4.8 to 7.3), confirming non-inferiority of cast immobilisation. No statistically significant differences were observed in secondary outcomes. In the surgery group, one participant had a superficial wound infection, one had delayed wound healing, and nine underwent procedures to remove hardware, two of whom developed postoperative infections (one deep and one superficial). Limitations include the single centre design and limited statistical power for rare complications.

What this study adds Cast immobilisation was non-inferior to surgical fixation for Weber B ankle fractures with a congruent mortise on standard radiography but deemed unstable by external rotation stress testing, with fewer treatment related harms.

Funding, competing interests, and data sharing This study was supported by state funding for university level health research (Oulu University Hospital). No competing interests declared. The data underlying the primary findings in this paper and the code used to analyse the data are openly and publicly available at https://doi.org/10.23729/fd-392c45f8-500c-32e5-adf9-f32f84b4e78a. Contact the corresponding author for problems with accessing the data.

Study registration ClinicalTrials.gov NCT01758796.



Between group difference in primary outcome, Olerud-Molander Ankle Score at two year follow-up. Error bar indicates two sided 95% confidence interval. The shaded area indicates the zone of non-inferiority

The BMJ is an Open Access journal. We set no word limits on *BMJ* research articles but they are abridged for print. The full text of each *BMJ* research article is freely available on bmj.com.

The online version is published along with signed peer and patient reviews for the paper, and a statement about how the authors will share data from their study. It also includes a description of whether and how patients were included in the design or reporting of the research.

The linked commentaries in this section appear on bmj.com as editorials. Use the citation given at the end of commentaries to cite an article or find it online.

CORRECTION

Ionising radiation and cardiovascular disease: systematic review and meta-analysis

This research paper by Little and colleagues (*BMJ* 2023;380:e074224, published in print issue of 11 March 2023) has a correction notice. For more details please go to doi:10.1136/bmj-2022-072924.

Fractures of the ankle that remain anatomically aligned are usually managed non-surgically in a cast or walking boot. However, fractures considered unstable (ie, at risk of falling out of alignment) are often treated with surgery. About 20 000 people are admitted to hospitals in England each year because of an ankle fracture.[1] In some patients, ankle instability is clear—for example, those with a fracture dislocation of the ankle. In other patients, instability is less obvious, such as those with an isolated fracture of the fibula at the level of the syndesmosis (Weber B fractures). For this common type of ankle fracture, stability depends on the extent of ligamentous injury, which cannot be easily determined clinically.

Several approaches are used to test the stability of the ankle joint after such a fracture. Many clinicians still rely on clinical assessment. Others use weightbearing radiography. Another approach to assessing stability is to undertake external rotation stress testing.[2] This test involves stressing the ankle to see if misalignment of the ankle is evident on radiographs while forces are applied by an assessor. If the test indicates misalignment, clinicians may consider this as an indication for internal fixation surgery. High quality data about whether surgery or non-surgical management should be offered in this specific situation is, however, limited. The SUPER-FIN trial addresses this uncertainty.[3]

In the SUPER-FIN trial, Kortekangas and colleagues compared cast immobilisation with open reduction and internal fixation surgery in adults with ankle fractures with isolated lateral malleolar (Weber B) injuries that were aligned in standard radiographs but determined to be unstable from external rotation stress testing.[3] In this randomised, pragmatic, non-inferiority trial, 126 participants were allocated 1:1 to cast immobilsation or surgery and then followed-up at two years post-randomisation. The primary outcome was the Olerud-Molander Ankle Score (OMAS, 0-100, higher scores better), which assesses patient reported symptoms and function.[4] The study achieved 96% follow-up at two years. The mean OMAS score was 89 in the cast immobilisation group and 87 in the surgery group, with a between group mean difference of 1.3 points (95% confidence interval −4.8 to 7.3). As the 95% confidence interval was within the prespecified non-inferiority margin of 8 points, the authors concluded cast immobilisation was non-inferior to surgery. Furthermore, the surgery group experienced more complications than the cast immobilisation group.

CORDELIA MOLLOY/SPL

The SUPER-FIN trial will support treatment decisions and updates to clinical guidelines

Study limitations

The SUPER-FIN team are to be commended for conducting a robust study that addresses an important clinical question. The investigators acknowledge that their trial was from a single university hospital, which may limit generalisability. Also, partial weight bearing (ie, patients to only put some weight on the ankle) was allowed for the first four weeks in both arms of the trial. This conservative approach to rehabilitation may reflect the fact that SUPER-FIN recruited participants over several years. More recent evidence suggests that unrestricted weight bearing after surgical management has functional advantages,[5] and guidelines now recommend against the use of the term partial weight bearing.[6]

One other potential limitation is that the trial had no outcome assessments before two years. Therefore, some uncertainty exists about potential differences between cast immobilisation and surgery in earlier recovery, which is important in circumstances where speed of recovery is a critical factor for patients.

Assessment differences

The investigators noted that clinicians in different settings and in different healthcare systems have different approaches to assessing potential instability in ankle fractures. These trial results will clearly have most impact in settings where external rotation stress testing is currently used routinely. Some authors have, however, noted these stress tests are not easy to standardise in terms of the technique and force used and, in the period immediately after the injury, can be limited by the patient's pain.[7] In the Netherlands, only 8% of the 178 surgeons from 68 hospitals responding to a nationwide survey reported using this procedure in the assessment of isolated Weber B injuries.[8] External rotation stress imaging was not included in the ankle fracture management guidelines of the British Orthopaedic Association Standards for Trauma.[9] This guidance recommends patients are reviewed within two weeks with radiography, with the patient weight bearing when possible, to check if ankle alignment is acceptable.

In the past decade, a growing number of randomised controlled trials have assessed the effectiveness of interventions for ankle fracture management, including but not limited to early versus late weight bearing after surgery for unstable fractures,[5] cast versus walking boot,[10] and surgery versus casting for unstable ankle fractures in adults aged 60 years or older.[11 12] Collectively these randomised controlled trials are supporting much needed advances in the evidence base for ankle fracture management and are a testament to the collaborative network of trauma and orthopaedic health professionals, researchers, and, most importantly, patient participants. The SUPER-FIN trial provides additional robust evidence for ankle fracture management and will support treatment decisions and updates to clinical guidelines.

David J Keene
david.keene@ndorms.ox.ac.uk
Matthew L Costa
See bmj.com for author details

Cite this as: *BMJ* 2026;392:s56
Find the full version with references at http://dx.doi.org/10.1136/bmj.s56

Experiences of access to general practice in England

Sinnott C, Ansari A, Price E, et al

Cite this as: *BMJ* 2026;392:e087367
Find this at doi: 10.1136/bmj-2025-087367

Study question What are the experiences and views of patients, carers, and staff on access to general practice in the context of major government plans to reform NHS primary care in England?

Methods Between July and October 2023, 70 qualitative interviews were conducted with 41 patients and carers in Devon, Medway, Blackpool, Luton, and Lancashire, and 29 staff at NHS general practices in the East of England. Analysis was based on the constant comparative method, with themes mapped to the three shifts—to digital, to community, and to prevention—proposed in the 10 year health plan for England.

Study answer and limitations This qualitative study suggests that patients, carers, and staff see the proposed changes in the NHS in England as potentially offering some benefits to patients but also introducing new risks and forms of disadvantage. The shift to greater digitisation in general practice may improve convenience for some people, but it could result in new inequities and do little to resolve the fundamental scarcity of appointments with a general practitioner. The shift to community based services was regarded as offering increased capacity, but it raised practical concerns and challenges for coordination and

continuity, especially with services spanning larger areas. Prevention initiatives were seen as important but risked fragmenting care, prioritising single disease models, and increasing workload for practice staff, with implications for patient initiated access to care. One limitation of this study is that it does not offer a full range of insights on the multiple influences on access to general practice nor of experiences of access, as it focuses only on aspects of the study relevant to the 10 year plan.

What this study adds While improving access to general practice is an aim of the reforms in the NHS, some proposals could lead to other unintended consequences, such as worsening dissatisfaction or increasing inequities. More clarity on the benefits to patients and what is good enough in terms of access is needed, along with careful co-design and evaluation of the detail of the three shifts.

Funding, competing interests, and data sharing
See full paper on bmj.com for funding and competing interests. Data are available upon reasonable request subject to development of a data sharing agreement and additional ethical approval.

WHAT IS ALREADY KNOWN ON THIS TOPIC

- Access to general practice is currently a priority for patients in the NHS in England
- Despite an increase in number of appointments, public satisfaction with access to NHS general practice has decreased over recent years
- Improving access to general practice is a key element of the government's new 10 year health plan

WHAT THIS STUDY ADDS

- None of the three shifts proposed by government—to digital, to community, and to prevention—is likely to meaningfully affect public dissatisfaction with access to general practice
- Some aspects of the three shifts are likely to increase inequities and create other unintended consequences, such as increased workload in general practice, new demand, and burden of treatment
- More clarity is needed on what the benefits are to patients and what is sufficient in terms of access, along with careful codesign and evaluation of the three shifts

MORE RESEARCH ON BMJ.COM

Knee bracing for osteoarthritis
Holden MA, et al. Provision of knee bracing for knee osteoarthritis (PROP OA): multicentre randomised controlled trial
BMJ 2026;392:e086005. http://dx.doi.org/10.1136/bmj-2025-086005

Impact of honour model on blood donation
Liu Y, et al. Impact of shifting blood donation policy from gift to honour model: staggered difference-in-differences analysis in China
BMJ 2026;392:e084999. http://dx.doi.org/10.1136/bmj-2025-084999

PPIs and stomach cancer
Duru O, et al. Long term use of proton pump inhibitors and risk of stomach cancer: population based case-control study in five Nordic countries
BMJ 2026;392:e086384. http://dx.doi.org/10.1136/bmj-2025-086384

Rates of newly recorded diagnoses in relation to the covid-19 pandemic
Russell MD, et al. Time trends in newly recorded diagnoses of 19 long term conditions before, during, and after the covid-19 pandemic: population based cohort study in England using OpenSAFELY
BMJ 2026;392:e086393. http://dx.doi.org/10.1136/bmj-2025-086393

Ovulation regimens and neonatal outcomes
Wei D, et al. Natural ovulation versus programmed regimens before frozen embryo transfer in ovulatory women
BMJ 2026;392:e087045. http://dx.doi.org/10.1136/bmj-2025-087045