

# Real time self-rating of decision certainty by clinicians: a systematic review

Authors: Myura Nagendran,<sup>A</sup> Yang Chen<sup>B</sup> and Anthony C Gordon<sup>C</sup>

## ABSTRACT

### Background

We sought to establish to what extent decision certainty has been measured in real time and whether high or low levels of certainty correlate with clinical outcomes.

### Methods

Our pre-specified study protocol is published on PROSPERO, CRD42019128112. We identified prospective studies from Medline, Embase and PsycINFO up to February 2019 that measured real time self-rating of the certainty of a medical decision by a clinician.

### Findings

Nine studies were included and all were generally at high risk of bias. Only one study assessed long-term clinical outcomes: patients rated with high diagnostic uncertainty for heart failure had longer length of stay, increased mortality and higher readmission rates at 1 year than those rated with diagnostic certainty. One other study demonstrated the danger of extreme diagnostic confidence – 7% of cases (24/341) labelled as having either 0% or 100% diagnostic likelihood of heart failure were made in error.

### Conclusions

The literature on real time self-rated certainty of clinician decisions is sparse and only relates to diagnostic decisions. Further prospective research with a view to generating hypotheses for testable interventions that can better calibrate clinician certainty with accuracy of decision making could be valuable in reducing diagnostic error and improving outcomes.

**KEYWORDS:** Decision certainty, decision confidence, uncertainty tolerance, clinical decision making, systematic review

## Introduction

Good decision making represents a key step in providing high-quality healthcare to patients.<sup>1</sup> However, clinicians typically have to make decisions with incomplete data in an environment

beset with uncertainty.<sup>2</sup> A 2015 Institute of Medicine report stated ‘nearly all patients will experience a diagnostic error in their lifetime, sometimes with devastating consequences’.<sup>3</sup> The act of diagnostic calibration is the process by which a clinician’s confidence in the accuracy of their diagnosis aligns with their actual accuracy.<sup>4–6</sup> This alignment of confidence and accuracy is applicable to decision making in every sphere of medicine, provided that both confidence and accuracy are measured. The importance of such alignment is in ensuring that both under- and over-confidence do not adversely affect decision making.

Unfortunately, the explicit measurement of confidence or certainty for a given medical decision is rare and generally suppressed; being uncertain can lead to criticism and feelings of vulnerability as well as possible medico-legal ramifications.<sup>2,7</sup> There is also an implicit bias of conflating high confidence with high competence. Indeed, a recent essay highlighted that speaking with confidence or assertiveness is sometimes prioritised in medical school, perpetuating over-confidence ahead of humility and appreciation of diagnostic uncertainty.<sup>8</sup>

Most studies which examine the relationship between self-rated confidence and outcomes such as diagnostic accuracy tend to administer retrospective questionnaires that suffer from recall bias or assess confidence in structured vignettes away from the ‘real-life’ clinical situations.<sup>5,9–11</sup> While vignettes may be a reasonable proxy for a clinician’s performance in real life, they cannot tell us about the association between decision certainty and outcomes such as length of stay, readmission and mortality. There have been several recent systematic reviews on uncertainty with regard to how we define or measure it, the tolerance of uncertainty and its impact on staff.<sup>12–15</sup> However, there have been no reviews to date focusing specifically on measuring decision certainty in real time during actual clinical practice. We therefore sought to establish to what extent decision certainty has been measured in real time and whether high or low levels of certainty correlate with clinical outcomes.

## Methods

This review was prepared according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement and registered in the online PROSPERO database (CRD 42019128112) prior to data extraction.<sup>16</sup>

## Review question

Our primary question was how many studies have been published that measure the self-rated certainty for a real time medical

**Authors:** <sup>A</sup>NIHR academic clinical fellow in intensive care medicine, Imperial College London, UK; <sup>B</sup>NIHR academic clinical fellow in cardiology, Institute of Cardiovascular Science, University College London, UK; <sup>C</sup>professor of intensive care medicine, Imperial College London, UK and Centre for Perioperative and Critical Care Research, London, UK

decision by a clinician (of any grade) in real clinical practice. Secondary questions for studies identified by the primary research question included: what scale this certainty is measured on, what is the distribution of certainty ratings and whether the certainty level is associated with any effect on clinical processes or outcomes.

### Data sources and searches

We performed a comprehensive electronic database search using Medical Subject Headings (MeSH) and free-text terms for various forms of the following three terms: decision making (anywhere in the full study record), certainty (in the title) and clinician (in the title). These three terms (and their synonyms) were combined using 'AND' so that all three needed to be present for a study to be retrieved. The exact search strategy is listed in supplementary material S1. The following electronic databases were searched from inception to 01 February 2019: MEDLINE, Embase and PsycINFO.

### Eligibility criteria

Publications were selected for review if they satisfied the following inclusion criteria: English language and prospective study that included real time self-rating of the certainty of a medical decision by a clinician (of any grade). Real time was defined as occurring at the time of documentation of the clinical encounter. Synonyms for certainty such as confidence were also accepted. Exclusion criteria included articles in languages other than English, non-study records (eg letters, editorials, case reports, reviews), studies using vignettes or questionnaires (even if the vignette was of a real patient

scenario). Any study in which a clinician made a prediction but did not additionally self-rate their certainty in that prediction was also excluded eg doctor predicting likelihood of sepsis as 70% in a patient but not explicitly rating their certainty in that prediction.

### Data extraction and quality assessment

Two authors (Myura Nagendran (MN), Yang Chen (YC)) independently screened titles and abstracts for potentially eligible studies. After removal of clearly irrelevant records, full text reports were reviewed for eligibility. Final decisions on inclusion were made by consensus between MN and YC. Study risk of bias was assessed by MN and YC using the Critical Appraisal Skills Programme (CASP) checklist for cohort studies or the Cochrane Risk of Bias tool for randomised trials.<sup>17,18</sup> Discrepancies were resolved by discussion between MN and YC to arrive at a consensus decision. No formal quantitative synthesis was planned.

### Results

The electronic search was performed on 01 February 2019. Flow of records is detailed in Fig 1. After screening, 17 full-text articles were assessed of which eight were excluded immediately. One full-text article could not be retrieved despite numerous efforts (the journal ceased publication in 1999) but we included the abstract that is still available.<sup>19</sup> Therefore, nine studies remained for inclusion after full-text review.<sup>19-27</sup> Overall risk of bias was high in all studies, including the only randomised trial we included (see supplementary material S2).

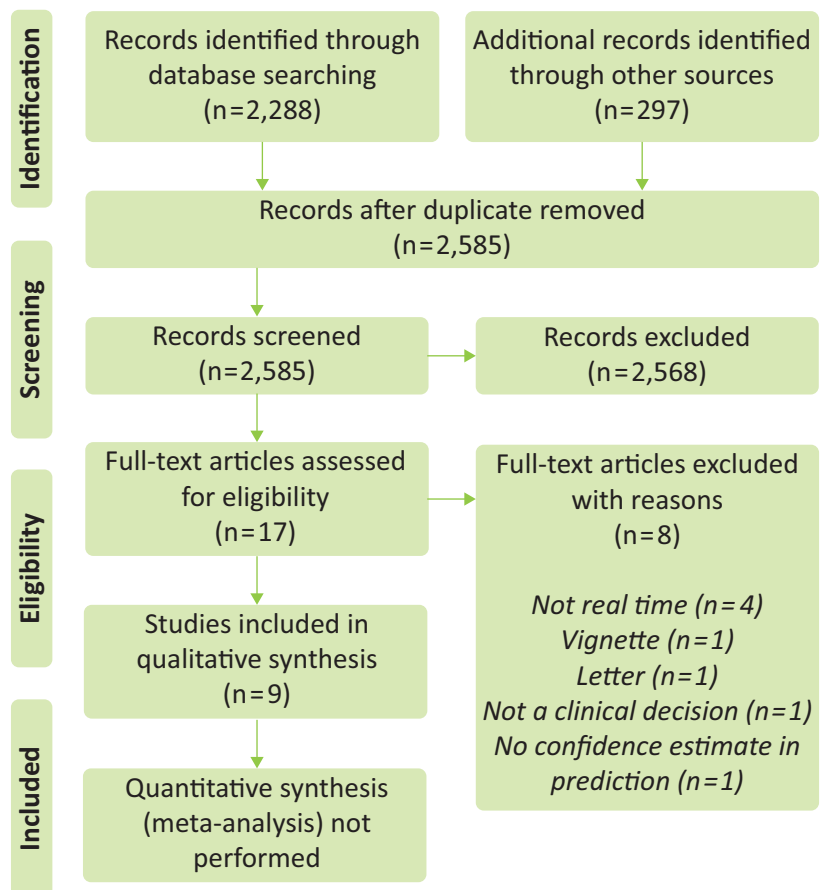


Fig 1. Flow of study records (PRISMA diagram).

**Table 1. Study characteristics**

Study	Year	Design	Specialty	Sample size	Duration	Decision category
AAP	2016	Multicentre prospective cohort study	General surgery	126	3 years	Diagnosis
Boots	2005	Multicentre prospective cohort study	Intensive care	470	7 months	Diagnosis
Bruyninckx	2009	Multicentre prospective cohort study	General practice	1,996	1 year	Diagnosis
Buntinx	1991	Prospective cohort study	General practice	318	2 months	Diagnosis
Davis	2005	Single centre prospective cohort study	Emergency medicine	276	1 year	Diagnosis
Green	2008	Single centre prospective cohort study	Emergency medicine	592	4 months	Diagnosis
Lave	1997	Single centre prospective cohort study	General medicine	1,778	Unclear	Diagnosis
McCullough	2002	Multicentre international prospective cohort study	Emergency medicine	1,538	21 months	Diagnosis
Robaei	2011	Randomised clinical trial	Emergency medicine	68	Unclear	Diagnosis

### Study characteristics

Study characteristics appear in Table 1. Effective sample size ranged from 68 to 1,996 (median 470, interquartile range (IQR) 276 to 1,538). The decision to which a certainty self-rating was applied related to diagnosis in every study. In four studies, the certainty scale applied to a specific diagnosis (heart failure<sup>21,23,26</sup> and pneumonia<sup>25</sup>).

### Certainty scales

Every study except one had a scale ranging from uncertain to certain (the exception ranged from high certainty of not heart failure to high certainty of heart failure).<sup>21</sup> Two studies used a four-part qualitative scale (either certain, rather certain, rather uncertain or uncertain; and certain, probable, suspected or unknown).<sup>22,27</sup> One studied provided a numerical adjunct to a three-part confidence rating on the likelihood of diagnosis (high >50%, medium 20–50% or low <20%).<sup>25</sup> Two studies used a one to 10 scale,<sup>20,24</sup> one used a one to seven scale<sup>21</sup> and three studies allowed a self-rating anywhere from 0–100.<sup>19,23,26</sup> Certainty was dichotomised from the original scale prior to analysis in three studies.<sup>22,23,26</sup>

### General findings

Study findings appear in Table 2. Extreme certainty ratings were reported in two studies.<sup>23,26</sup> McCullough *et al* found that clinicians rated the likelihood of acute heart failure as 0% in 232 patients (19 of whom were eventually diagnosed with heart failure) and 100% in 109 patients (5 of whom were eventually diagnosed without heart failure).<sup>26</sup> Green *et al* reported that over a quarter of all patients (26%) with dyspnoea were rated as having 0% likelihood of acute heart failure.<sup>23</sup> The difference between certainty ratings of junior trainees and senior doctors was examined by only one study;<sup>20</sup> the acute abdominal pain (AAP) study group found that diagnostic accuracy between the two groups was comparable at 44 and 43% respectively, but trainees were less certain than seniors for urgent vs non-urgent diagnosis (trainee median 7 (IQR 6 to 8) versus senior median 8 (IQR 7 to 9)). Both groups recommended imaging in 77% of patients despite seniors being more certain about their diagnoses.

Buntinx *et al* studied chest pain in general practice with a follow-up period occurring between 2 weeks and 2 months after

the index clinical encounter.<sup>27</sup> The proportion of diagnoses rated certain or probable increased from 74 to 88% with the benefit of follow-up information. The only randomised trial we identified focused on assessing the value of providing brain natriuretic peptide (BNP) measurements to clinicians in the intervention group once they had rated their certainty in a diagnosis of heart failure.<sup>21</sup> The provision of BNP reduced diagnostic uncertainty and improved diagnostic accuracy.

### Associations between certainty self-rating and process or outcome measures

Green *et al* found that patients who had diagnoses rated with certainty had shorter median length of stay than those rated with uncertainty (5.4 vs 6.6 days,  $p=0.02$ ).<sup>23</sup> Age-adjusted Cox proportional hazards analyses showed uncertainty to be an independent predictor of death (hazard ratio (HR) 1.9, 95% confidence interval (CI) 1.0–2.3,  $p=0.05$ ) as well as death or rehospitalisation (HR 2.2, 95% CI 1.7–2.5,  $p=0.01$ ) by 1 year. In contrast, Lave *et al* found that certainty rating was not associated with vital status at discharge but did account for variation in all measures of resource utilisation assessed except adjusted pharmacy charges.<sup>19</sup> Bruyninckx *et al* found that in general practice, the proportion of urgent referrals for chest pain was similar between certain (14%) and uncertain (18%) ratings.<sup>22</sup> The main difference observed was in decisions not to refer (occurring 65% of the time when certain vs 29% of the time when uncertain) and non-urgent referrals (occurring 20% of the time when certain or 54% of the time when uncertain).

### Discussion

This is, to our knowledge, the first review of studies measuring the real time self-rated certainty of decisions made by clinicians and there are several prominent findings. First, all studies were generally at high risk of bias making applicability to contemporary clinical practice difficult. Second, the decision to which a certainty self-rating was applied related to diagnosis in every case and was often not the main focus of the study. Third, certainty tended to be measured by a quantitative rather than qualitative scale though visualisation of the distribution of certainty ratings was only present in two of nine studies. Fourth, extreme certainty ratings (0 or 100% likelihood of heart failure) were reported in two studies,

Table 2. Study findings.

Study	Distribution of self-rated certainty	Scale for self-rating	Association between self-rating and process or outcome measures
AAP	Varied depending on diagnosis	VAS which ranged from 1 (total uncertainty) to 10 (total certainty)	Although surgeons were more certain about their diagnoses, they did not propose to use less imaging than trainees. Imaging recommended by both groups in 77% of patients, though it actually occurred in 84%.
Boots	High confidence in 57%, medium in 25% and low in 18%	Clinical diagnostic confidence rated to one of three levels: high >50% likelihood, medium 20–50% likelihood and low <20% likelihood	High clinical confidence in pneumonia diagnosis was reported as an independent predictor for bronchoscopy (OR 2.19, CI 95% 1.29–3.72, p=0.004) although many factors were assessed in logistic regression.
Bruyninckx	Dichotomised to binary scale (certain 1,673 vs uncertain 323)	Four-point categorical scale (certain, rather certain, rather uncertain or uncertain)	When certain, 65% of cases not referred. When uncertain, 29% of cases not referred. Urgent referral similar. Main difference in non-urgent referrals.
Buntinx	Certain (33%), probable (41%), suspected (20%) and unknown (6%)	Four-point categorical scale (certain, probable, suspected or unknown)	An ECG was performed more frequently when the diagnosis was uncertain: 20% of patients with a certain diagnosis compared with 34% of patients whose diagnosis was only suspected.
Davis	41% of ratings were 9 or 10 while 12% of ratings 5 or less	Confidence scale from 1 to 10	Not reported
Green	69% rated certainty and 31% rated uncertainty. For full diagram of distribution see Green SM, Martinez-Rumayor A, Gregory SA <i>et al.</i> Clinical uncertainty, diagnostic accuracy, and outcomes in emergency department patients presenting with dyspnea. <i>Arch Intern Med</i> 2008;168:741–8.	0 to 100% scale of likelihood. Certainty estimate of either <20% or >80% was classified as clinical certainty, while estimates 21% to 79% were defined as clinical uncertainty	Patients judged uncertainly with respect to the presence or absence of acute heart failure were more likely to be admitted to the hospital, had a longer index hospital LOS, and had higher rates of 1-year morbidity and mortality, especially in those ultimately diagnosed as having acute heart failure.
Lave	Unclear	VAS with hash marks at 0, 25, 50, 75 and 100	Certainty rating was a significant factor accounting for variation in all measures of resource utilisation except adjusted pharmacy charges.
McCullough	47% rated low probability (0–20% likelihood), 28% rated intermediate probability (21–79% likelihood) and 25% rated high probability (80–100% likelihood)	0 to 100% scale of likelihood. Certainty estimate of either <20% or >80% was classified as clinical certainty, while estimates 21% to 79% were defined as clinical uncertainty	Not reported
Robaei	14 ratings for score 1–2, 42 ratings for score 3–5, 12 ratings for score 6–7. For full diagram of distribution see Robaei D, Koe L, Bais R <i>et al.</i> Effect of NT-proBNP testing on diagnostic certainty in patients admitted to the emergency department with possible heart failure. <i>Ann Clin Biochem</i> 2011;48:212–7.	1 to 7 scale (1 being high certainty diagnosis not heart failure and 7 being high certainty diagnosis is heart failure)	Not reported

CI = confidence interval; ECG = electrocardiogram; LOS = length of stay; OR = odds ratio; VAS = visual analogue scale.

one of which found a 7% diagnostic error rate among these cases. Fifth, very few studies reported clinical outcomes, such as length of stay, readmission or long-term mortality, but in the one study that did, there was an association between uncertainty and worse outcomes.

### Comparison with the literature

Comparing our findings with the existing non-real time literature requires caution. Traditionally, these have taken the form of vignette studies that provide experimental rigour. For example,

Meyer *et al* investigated diagnostic calibration (the relationship between diagnostic accuracy and the confidence in that accuracy).<sup>5</sup> They were able to adjust both the difficulty of the diagnostic case and the stage of evolution of the case (ie early with little information or later on with more test results). This ability to adjust is a major advantage of vignette studies. They reported that diagnostic calibration was worse with more difficult cases (accuracy dropped as expected but confidence stayed at a reasonably similar level). They also noted that higher confidence was associated with decreased requests for additional diagnostic tests. However, one study in our review reported that imaging requests were similar between junior and senior doctors despite increased confidence recorded by the seniors.<sup>20</sup> This difference may exemplify the problems inherent in comparing vignettes studies to real clinical practice.

More recently, a vignette study from Lawton *et al* found that more experienced emergency medicine doctors were more tolerant of uncertainty and that this explained about a quarter of the relationship between experience and lower risk aversion.<sup>10</sup> However, the authors stated that they could not conclude from their design that less risk-averse strategies improve patient safety. A major weakness of vignette studies is that by their nature they cannot provide information on real patient outcomes.

Studies using questionnaires or surveys can address this issue by categorising a cohort of clinicians into various groups based on their responses and then looking at their outcomes. For example, Baldwin *et al* measured risk aversion and diagnostic uncertainty among 46 doctors by administering a structured questionnaire and then looked to see if the results were associated with their decisions to admit infants with bronchiolitis.<sup>9</sup> When adjusted for severity of illness, physicians' risk attitudes were not associated with admission rates. However, data from our review suggests that certainty was associated with outcomes such as length of stay and readmission.<sup>23</sup> One reason for differences between questionnaire and real time studies may be that the former take an aggregate level overview while the latter present more granular data on each individual decision. A clinician rating their personality as risk averse and afraid of uncertainty may nevertheless have a spectrum of certainty self-ratings depending on the particulars of an individual case and other factors such as time of day, presence of senior support and volume of diagnostic results available. Such granularity cannot be captured in a questionnaire study.

### Implications for clinicians and researchers

Our findings have several important implications. First, all of the studies we identified focused on diagnosis. While diagnostic error has become a major target for patient safety efforts, there are many other decisions within healthcare that also have a large impact on patient outcomes such as treatment choice and discharge decisions. These could also make valuable targets for future research in this area.<sup>28</sup>

For clinicians, there may be educational value, both for trainees and senior doctors in seeing how their self-rated certainty relates to longer-term outcomes and indeed how they compare to their peers. Future studies could couple this information into an audit and feedback system that facilitates reflective practice.<sup>29</sup> The mere act of quantifying a certainty rating might form a cognitive brake and thus an important debiasing strategy that mitigates against avoidable medical error occurring from under- or over-confidence.<sup>30–33</sup>

For researchers, there is very weak existing evidence to suggest a relationship between certainty self-ratings and outcomes such as mortality, readmission and length of stay. At the very least, this may imply that uncertainty could act as a surveillance metric for patients at higher risk of an adverse outcome during their admission (or even after discharge) and therefore a potential target for quality improvement efforts. It is possible that improvements in clinical processes such as reduced waste from over-investigation or better patient flow to appropriate environments could arise as a result of highlighting patients where decisions are made with high or low certainty. However, more prospective work focusing on a broader medical context and identifying the many factors that influence clinician certainty will be required before interventions can be tested.

### Limitations

Our findings should be considered in light of several limitations. First, although comprehensive, our search may nonetheless have missed some potentially includable studies. As with any review, a degree of pragmatism is required to balance the identification of studies with logistical constraints. This is especially important in this review as we were effectively looking for medical decisions in any sphere of clinical practice and therefore the search terms could not realistically have been any broader than they were. Second, we elected *a priori* not to perform quantitative synthesis as we anticipated that there would be extensive heterogeneity. While this makes sense from a methodological point of view, it nonetheless means that we cannot, from this review, give quantitative estimates of the benefits of measuring decision certainty.

### Conclusions

The literature on self-rated real time certainty of clinician decisions focuses exclusively on diagnosis and is notably sparse. All studies were generally at high risk of bias making applicability to contemporary clinical practice difficult. The next challenge will be to generate hypotheses for testable interventions that can better calibrate clinician certainty with accuracy of decision making. This could be valuable in reducing diagnostic error and improving outcomes. ■

### Funding

There was no specific funding for this study. MN and YC are supported by National Institute for Health Research (NIHR) academic clinical fellowships. Anthony C Gordon (ACG) is funded by a UK NIHR research professor award (RP-2015-06-018). MN and ACG are both supported by the NIHR Imperial Biomedical Research Centre.

### Supplementary material

Additional supplementary material may be found in the online version of this article at [www.clinmed.rcpjournals.org](http://www.clinmed.rcpjournals.org):

- S1 – Search strategy.
- S2 – Risk of bias.

### References

- 1 Saposnik G, Redelmeier D, Ruff CC *et al*. Cognitive biases associated with medical decisions: a systematic review. *BMC Med Inform Decis Mak* 2016;16:138.

- 2 Simpkin AL, Schwartzstein RM. Tolerating uncertainty — the next medical revolution? *N Engl J Med* 2016;375:1713–5.
- 3 McGlynn EA, McDonald KM, Cassel CK. Measurement is essential for improving diagnosis and reducing diagnostic error: a report from the institute of medicine. *JAMA* 2015;314:2501–2.
- 4 Meyer A, Krishnamurthy P, Sur M *et al*. Calibration of diagnostic accuracy and confidence in physicians working in academic and non-academic settings. *Diagnosis* 2015;2:eA10.
- 5 Meyer AND, Payne VL, Meeks DW *et al*. Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA Intern Med* 2013;173:1952–8.
- 6 Meyer AND, Singh H. The path to diagnostic excellence includes feedback to calibrate how clinicians think. *JAMA* 2019;321:737–8.
- 7 Logan R. Concealing clinical uncertainty—the language of clinicians. *N Z Med J* 2000;113:471–3.
- 8 Treadway N. A student reflection on doctoring with confidence: mind the gap. *Ann Intern Med* 2018;169:564.
- 9 Baldwin RL, Green JW, Shaw JL *et al*. Physician risk attitudes and hospitalization of infants with bronchiolitis. *Acad Emerg Med* 2005;12:142–6.
- 10 Lawton R, Robinson O, Harrison R *et al*. Are more experienced clinicians better able to tolerate uncertainty and manage risks? A vignette study of doctors in three NHS emergency departments in England. *BMJ Qual Saf* 2019;28:382–8.
- 11 Yee LM, Liu LY, Grobman WA. The relationship between obstetricians' cognitive and affective traits and their patients' delivery outcomes. *Am J Obstet Gynecol* 2014;211:692.e1–6.
- 12 Alam R, Cheraghi-Sohi S, Panagioti M *et al*. Managing diagnostic uncertainty in primary care: a systematic critical review. *BMC Fam Pract* 2017;18:79.
- 13 Bhise V, Rajan SS, Sittig DF *et al*. Defining and measuring diagnostic uncertainty in medicine: a systematic review. *J Gen Intern Med* 2018;33:103–15.
- 14 Cranley L, Doran DM, Tourangeau AE *et al*. Nurses' uncertainty in decision-making: a literature review. *Worldviews Evid Based Nurs* 2009;6:3–15.
- 15 Strout TD, Hillen M, Gutheil C *et al*. Tolerance of uncertainty: A systematic review of health and healthcare-related outcomes. *Patient Educ Couns* 2018;101:1518–37.
- 16 Moher D, Liberati A, Tetzlaff J *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.
- 17 CASP. *CASP Checklists*. Oxford: CASP. <http://casp-uk.net/casp-tools-checklists> [Accessed 29 April 2019].
- 18 Higgins JPT, Altman DG, Gøtzsche PC *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928.
- 19 Lave JR, Bankowitz RA, Hughes-Cromwick P *et al*. Diagnostic certainty and hospital resource use. *Cost Qual Q J* 1997;3:26–32.
- 20 Acute Abdominal Pain Study Group. Diagnostic accuracy of surgeons and trainees in assessment of patients with acute abdominal pain. *Br J Surg* 2016;103:1343–9.
- 21 Robaei D, Koe L, Bais R *et al*. Effect of NT-proBNP testing on diagnostic certainty in patients admitted to the emergency department with possible heart failure. *Ann Clin Biochem* 2011;48:212–7.
- 22 Bruyninckx R, Van den Bruel A, Aertgeerts B *et al*. Why does the general practitioner refer patients with chest pain not-urgently to the specialist or urgently to the emergency department? Influence of the certainty of the initial diagnosis. *Acta cardiologica* 2009;64:259–65.
- 23 Green SM, Martinez-Rumayor A, Gregory SA *et al*. Clinical uncertainty, diagnostic accuracy, and outcomes in emergency department patients presenting with dyspnea. *Arch Intern Med* 2008;168:741–8.
- 24 Davis DP, Campbell CJ, Poste JC *et al*. The association between operator confidence and accuracy of ultrasonography performed by novice emergency physicians. *J Emerg Med* 2005;29:259–64.
- 25 Boots RJ, Lipman J, Bellomo R *et al*. Predictors of physician confidence to diagnose pneumonia and determine illness severity in ventilated patients. Australian and New Zealand practice in intensive care (ANZPIC II). *Anaesth Intensive Care* 2005;33:112–9.
- 26 McCullough Peter A, Nowak Richard M, McCord J *et al*. B-type natriuretic peptide and clinical judgment in emergency diagnosis of heart failure. *Circulation* 2002;106:416–22.
- 27 Buntinx F, Truyen J, Embrechts P *et al*. Chest pain: an evaluation of the initial diagnosis made by 25 Flemish general practitioners. *Fam Pract* 1991;8:121–4.
- 28 Prescott HC, Iwashyna TJ. Improving sepsis treatment by embracing diagnostic uncertainty. *Ann Am Thorac Soc* 2019;16:426–9.
- 29 Patel S, Rajkomar A, Harrison JD *et al*. Next-generation audit and feedback for inpatient quality improvement using electronic health record data: a cluster randomised controlled trial. *BMJ Qual Saf* 2018;27:691–9.
- 30 Croskerry P, Singhal G, Mamede S. Cognitive debiasing 1: origins of bias and theory of debiasing. *BMJ Qual Saf* 2013;22(Suppl 2):ii58–64.
- 31 Croskerry P, Singhal G, Mamede S. Cognitive debiasing 2: impediments to and strategies for change. *BMJ Qual Saf* 2013;22(Suppl 2):ii65–72.
- 32 Mamede S, Schmidt HG, Rikers RMJP *et al*. Conscious thought beats deliberation without attention in diagnostic decision making: at least when you are an expert. *Psychol Res* 2010;74:586–92.
- 33 Mamede S, van Gog T, van den Berge K *et al*. Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *JAMA* 2010;304:1198–1203.

**Address for correspondence: Dr Myura Nagendran, NIHR academic clinical fellow in intensive care medicine, St Mary's Campus, Imperial College London, Praed Street, London W2 1NY, UK.**  
**Email: myura.nagendran@imperial.ac.uk**