BMJ Open

Evaluating AI-based comprehensive clinical decision support for sepsis and ARDS: protocol for a Clinician Turing Test

Antonia Angeli Gazola ,[1] Nicholas S Bishop ,[1] Benjamin E Schmid,[1] Romain Pirracchio ,[2] Thomas S Valley,[3,4,5,6] Sivasubramanium V Bhavani,[7,8,9] Dustin C Krutsinger,[10,11] Heather M Giannini,[12,13] Yingying Lu,[1] Lyle H Ungar,[14] Nuala J Meyer,[12,15] Meeta Prasad Kerlin,[1,12,16] Gary E Weissman [1,12,16,17]

To cite: Angeli Gazola A, Bishop NS, Schmid BE, *et al*. Evaluating AI-based comprehensive clinical decision support for sepsis and ARDS: protocol for a Clinician Turing Test. *BMJ Open* 2025;**15**:e106757. doi:10.1136/bmjopen-2025-106757

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2025-106757).

Received 18 June 2025
Accepted 08 December 2025

Check for updates

© Author(s) (or their employer(s)) 2025. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

For numbered affiliations see end of article.

Correspondence to
Dr Gary E Weissman; gary.weissman@pennmedicine.upenn.edu

ABSTRACT

Introduction Few artificial intelligence (AI) clinical decision support systems (CDSSs) are ever evaluated in practice. Although some signal of clinical effectiveness may be needed to justify AI deployment and testing, such data are typically unavailable in early-stage research. This conundrum is especially relevant in the intensive care unit (ICU), where conditions like sepsis and acute respiratory distress syndrome (ARDS) require high-stakes decisions. Our group developed the AI ventilator assistant (AVA), a novel AI CDSS for patients with sepsis ARDS receiving invasive mechanical ventilation. But the promising results of predictive performance estimates are not sufficient to assess AVA's clinical safety and appropriateness prior to future evaluation and deployment. Therefore, we propose a Clinician Turing Test as a novel validation approach to determine whether clinicians can distinguish AVA-generated treatment recommendations from those enacted by real human clinicians. If AVA's recommendations are consistently indistinguishable from those of real clinicians, thereby 'passing' this Turing test, this would provide a strong preclinical signal of safety and appropriateness.

Methods and analysis This multisite, randomised, electronic, vignette-based Phase 1b study will use a Clinician Turing Test design. We aim to recruit 350 critical care clinicians, including physicians and advanced practice providers from six US hospitals. Participants will review nine clinical vignettes of patients with sepsis and ARDS derived from the Molecular Epidemiology of Severe Sepsis in the ICU cohort and an associated profile of a suggested treatment plan. For each participant–vignette combination, the source of the treatment profile will be randomly assigned (AI-generated by AVA vs the actually enacted treatment from real human clinicians) in a 1:1 allocation. The primary endpoint is the participants' accuracy in identifying whether a treatment profile was AI-generated or human-generated, assessed using equivalence testing through a mixed-effects logistic regression model with random effects for participants and vignettes. Secondarily, a fitted binary classifier will assess discrimination ability using the C-statistic. Secondary endpoints include clinicians' perceptions of the safety and appropriateness of the treatment profiles, confidence in distinguishing AI-generated and human-generated recommendations, interest in AI CDSSs for sepsis and ventilator management and the time to complete the survey. This novel Phase 1b design provides preliminary but essential information about an AI CDSS's clinical appropriateness without the risk or cost of actual deployment, thereby informing decisions about future clinical implementation and evaluation in real clinical environments.

Ethics and dissemination This protocol was approved by the Institutional Review Board of the University of Pennsylvania (Protocol #858201). Results are expected in 2026 and will be submitted for publication in peer-reviewed journals and presented at scientific conferences.

Trial registration number NCT07025096.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ The Clinician Turing Test is a study design that provides clinically relevant data about an artificial intelligence (AI) model's recommendations without the costs or risks of deployment in a clinical environment.
⇒ The Clinician Turing Test provides important clinical information about an AI model's recommendations that are unobtainable from traditional Phase 0 studies that only report predictive performance.
⇒ The Clinician Turing Tests rely on randomisation and blinding to enhance the rigour of the design and reduce bias in inferences during the analysis.
⇒ The design relies on clinicians' responses to electronic vignettes that may not fully reflect workflows or cognitive processes in a real clinical environment.
⇒ The clinical vignettes are based on real patient cases but are drawn from a single, quaternary referral centre and may not reflect the typical cases encountered in most hospitals.

INTRODUCTION

Background and rationale

Few artificial intelligence (AI)-based and machine learning (ML)-based clinical decision support systems (CDSSs) are ever evaluated in practice.[1 2] This critical lack of data about the clinical effectiveness of AI/ML CDSSs has so far limited their widespread impact despite significant promise.[3] Predictive performance measures themselves, often

BMJ Open: first published as 10.1136/bmjopen-2025-106757 on 24 December 2025. Downloaded from http://bmjopen.bmj.com/ on March 21, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

derived from Phase 0 and Phase 1 studies, are unreliable signals of clinical effectiveness. Real-world clinical testing in Phase 2 or Phase 3 studies, such as clinical trials, is needed for robust and valid assessments of AI CDSSs' effects at the bedside. But the costs, complexity and potential safety concerns associated with real-world clinical evaluations of untested AI systems are substantial. Thus, any early signals from preclinical studies supporting the safety and effectiveness of clinical AI/ML systems would be highly valuable in planning for future clinical evaluations. But researchers may be caught between having insufficient clinical data to justify Phase 2 or 3 studies and being simultaneously unable to obtain clinically meaningful data using traditional Phase 0 and 1 study designs. This conundrum is particularly relevant in the intensive care unit (ICU) where life-threatening conditions such as sepsis and acute respiratory distress syndrome (ARDS) are common, and AI/ML CDSS would be particularly helpful for such complex decision making. At the same time, real-world clinical deployment and evaluation of such a system poses logistical and safety concerns that are not addressed by typical early-phase research evaluating predictive performance alone.

Therefore, we propose a novel Phase 1b study design to overcome these knowledge gaps and serve as a bridge between preclinical and clinical studies of AI/ML CDSSs. The proposed design is a clinical modification to the test proposed by Alan Turing in 1950 to determine if a machine exhibits human-like intelligence.[4] In Turing's original proposal, if a human was unable to distinguish between a human and a machine as the source of text in a transcribed conversation, then the machine would have 'passed' the test. Analogues of Turing's test have been applied to generative AI systems to assess their apparent behavioural characteristics,[5] proposed as a model for human-AI collaboration in cancer care,[6] used to identify bias in AI radiology models[7] and used to evaluate the quality of answers from a generative AI chatbot about diabetes mellitus.[8] However, none of these studies have used Turing's approach to establish the equivalence of human and AI-generated treatment recommendations.

The Clinician Turing Test is a special case of Turing's original test in which the goal is to determine if an expert clinician can reliably distinguish between treatment suggestions that originate from another real clinician or from an AI/ML CDSS. If such an expert cannot distinguish those sources, then the output of an AI/ML CDSS might reasonably be considered equivalent to real, human clinicians, with a strong, preliminary signal of equipoise in advance of future clinical deployment and testing. We illustrate and detail several important features of the design and analysis of a Clinician Turing Test using an AI/ML CDSS used to guide treatment decisions for patients with sepsis and ARDS in the ICU. This design relies on an electronic, vignette-based survey that is feasible to conduct without the substantial investments in data and computational infrastructure needed to deploy and integrate AI/ML CDSS into real clinical workflows. This design also provides essential information about the clinical appropriateness of an AI/ML CDSS's output while ensuring patient safety through testing outside of a real clinical environment.

Study aim

To address knowledge gaps that limit the implementation and evaluation of AI/ML CDSSs, in general, and that of the AI ventilator assistant (AVA) CDSS, in particular, we propose a novel, preclinical study design. We will conduct a Clinician Turing Test in this Phase 1b study to assess AVA, an AI/ML CDSS developed by our team to support the management of patients with sepsis and ARDS in the ICU. AVA will generate recommendations for a comprehensive range of simultaneous and inter-related treatment decisions required in caring for people with sepsis and ARDS. The development and validation of a preliminary version of the AVA model in a Phase 0 study has been reported separately.[9] This study will compare AVA's clinical recommendations to treatment decisions made by real clinicians using an equivalence testing design.

Primary objective

- To assess the ability of expert clinicians to distinguish between AVA's treatment recommendations and actual treatments enacted by real clinicians for patients with sepsis and ARDS.

Secondary objectives

- To assess clinicians' determination of the safety and appropriateness of AVA's treatment recommendations and assess the equivalence of those assessments with respect to treatments enacted by real clinicians.
- To measure clinicians' confidence in distinguishing AI-generated treatment recommendations from human-enacted treatments.
- To identify specific treatment decisions within the larger treatment profiles that may require further refinement prior to future clinical deployment.
- To assess clinicians' interest in using an AI/ML CDSS for sepsis and ventilator management in clinical practice.
- To evaluate the time spent completing each vignette and assess differences among vignettes with AI-recommended and human-enacted treatments.

The overarching goal of this work is to evaluate AVA's equivalence to real human practice while introducing a novel study design for accelerating and improving the development, testing and safe implementation of AI/ML CDSS in clinical practice.

METHODS AND ANALYSIS

Study design

This is a multisite, randomised, electronic, vignette-based, cross-sectional study (figure 1). This Clinician Turing Test is a Phase 1b study. It is not a clinical trial as there are

BMJ Open: first published as 10.1136/bmjopen-2025-106757 on 24 December 2025. Downloaded from http://bmjopen.bmj.com/ on March 21, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

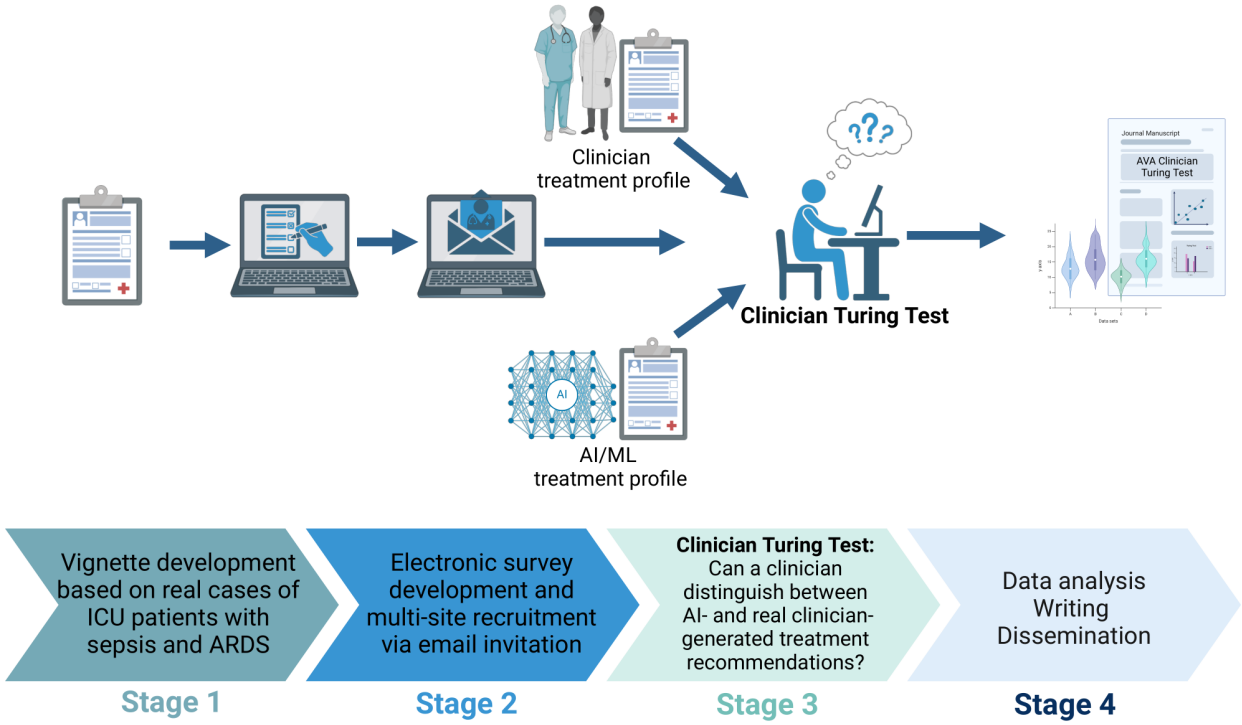


Figure 1 Clinician Turing Test study overview. Study overview diagram. This figure was created in BioRender. AG, A. (2025) https://BioRender.com/t88k194. AI, artificial intelligence; ARDS, acute respiratory distress syndrome; AVA, AI ventilator assistant; ICU, intensive care unit; ML, machine learning.

no actual patients, treatment decisions or patient care involved.

Study population and recruitment

We will recruit participants with experience in caring for patients with sepsis and acute respiratory failure requiring invasive mechanical ventilation (IMV). To be eligible to participate in this study, individuals will meet the following criteria: (1) be at least 18 years old; (2) work as a physician (eg, MD or DO; attending or fellow) or advanced practice provider (eg, nurse practitioner or physician assistant); (3) work in medical critical care, anaesthesia critical care, surgical critical care or emergency medicine in a US-based hospital or medical centre. Medical residents, interns or physicians who work in other clinical specialties or who have not completed residency training will be excluded from the study.

The target recruitment goal will be 350 clinicians who provide complete survey responses (see sample size calculation below). Partial responses will be included in the analysis but will not be counted toward the recruitment goal. Our study team will develop a curated contact list of 700 eligible clinicians based on the above criteria with their email addresses from participating study sites. We estimate a 50% response rate based on prior studies recruiting from similar populations with comparable participant incentives. Study team members from each site will review departmental websites, clinician directories and other publicly available and local resources to identify candidate participants whose eligibility will be verified by the study team and site investigators.

We will send invitations to participate in the study via email using Qualtrics (Provo, Utah, USA). The survey invitation will include a description of the study, compensation and contact information for the study team. Follow-up emails will be sent to non-responders or to those providing partial responses until the target recruitment goal is met (figure 2). Individuals who decline consent will not receive follow-up emails.

Participants will provide an electronic, signed consent through the survey instrument. A sample copy of the informed consent document is available in online supplemental material 1. On receiving a recruitment invitation email, clinicians will access the survey through a personalised, non-transferable link that will allow multiple entries to facilitate survey completion. The survey itself will take approximately 30–45 min to complete. Internet access and a computer, tablet or smartphone will be needed to participate. All recruitment and survey instrument materials will be in English. Each participant will receive a US$75 gift card as an incentive following completion of the survey instrument.

Patient and public involvement

Neither study participants nor the public were involved in the study design, recruitment, conduct or planned reporting and dissemination of the results.

Instrument development and randomisation

We will develop and administer the survey using the Qualtrics platform. The survey will collect demographic information and details on participants' professional

BMJ Open: first published as 10.1136/bmjopen-2025-106757 on 24 December 2025. Downloaded from http://bmjopen.bmj.com/ on March 21, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

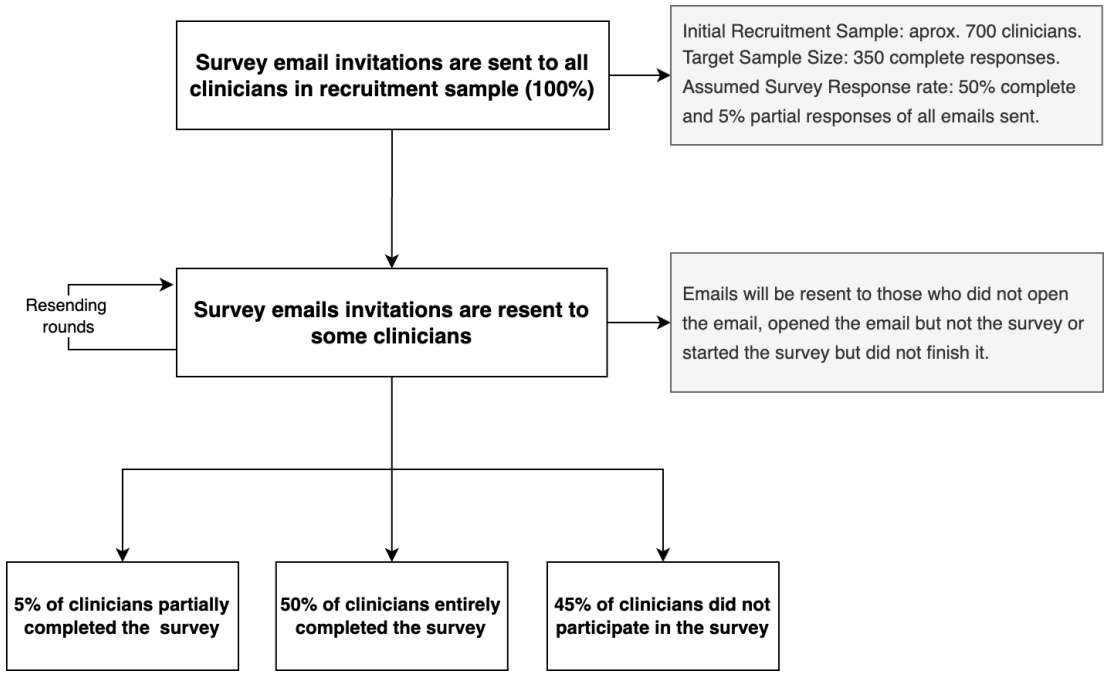


Figure 2 Survey recruitment and response flowchart. The percentages are based on the total number of emails sent (100%). We anticipate that 50% of candidate participants will complete the survey, 5% will partially complete the survey and 45% will not participate in the survey (did not open the email, opened the email but did not start the survey or did not consent). This process will be repeated as needed until the required number of complete responses is obtained.

clinical experience and training. Then participants will be presented with a brief visual tutorial on how to interpret the vignette information and how to complete the survey, where a sample vignette will be present (online supplemental material 3).

Participants will then be presented with nine clinical vignettes in random order to prevent ordering effects. Each vignette will contain a concise summary of relevant clinical details for a patient who meets sepsis and ARDS criteria and was just initiated on IMV. The case summary will include the most recent chest radiograph image prior to intubation, and laboratory test results and vital signs from the preceding 24 hours. Following the vignette, participants will be presented with a comprehensive treatment profile for that patient that includes ventilator settings, medication infusion doses and administration rates, and antibiotic treatments, among others. Participants will be asked to evaluate the safety and clinical appropriateness of the proposed treatment profile, identify whether they think the treatment profile was generated by an AI/ML CDSS or by a human clinician, and be asked to rate their confidence on their guess of the treatment profile source. Additionally, participants will rate the safety and appropriateness of the treatment recommendations using a 7-point Likert scale and provide feedback on any specific elements of the treatment profile they consider to be unsafe or inappropriate.

For eight of the study vignettes, participants will be randomised to view either the AI-generated treatment profile or a profile reflecting what a real human clinician did in practice. The randomisation will ensure a 1:1 allocation so that each participant sees four treatment profiles from each source. A ninth vignette, in which the treatment profile is obviously derived from an AI system through the use of non-standard numbers and additional decimal places in suggested medication doses and ventilator settings, will also be included as an internal control. Participants will not be informed of the study arm allocation beforehand and will not be able to return to or modify responses to previously completed vignettes. There will be no time limit to complete the survey. Although participants are expected to respond based on their own clinical knowledge, the online design prevents direct control over the use of external aids. However, participants will be instructed not to use large language models or other electronic resources to assist in guessing vignette sources.

Survey data responses will be stored in Qualtrics and later exported in tabular format for analysis. The study team will iteratively test and refine the survey instrument to ensure clear communication of relevant information in the vignettes and reliable functionality and data capture prior to study launch.

Clinical vignette development

We extracted clinical data for the vignettes from the Complete Inpatient Record using Comprehensive Electronic Data project, which contains electronic health record (EHR) for all hospital-based encounters at Penn Medicine from 1 July 2017 through the present.[10] We identified candidate cases of patients with sepsis and ARDS from the Molecular Epidemiology of SepsiS in the Intensive Care Unit (MESSI) study.[11] The MESSI study prospectively identifies patients at Penn Medicine with sepsis and each enrolled patient undergoes rigorous chart

BMJ Open: first published as 10.1136/bmjopen-2025-106757 on 24 December 2025. Downloaded from http://bmjopen.bmj.com/ on March 21, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Table 1 Study endpoints and outcomes

Primary study endpoint	Primary outcome measure
Accuracy of participants' guess on treatment profile source (AI vs Human) for each clinical vignette	AVA's potential clinical performance by assessing equivalence to treatment decisions made by real expert clinicians
Secondary study endpoints	**Secondary outcome measures**
Participants' assessments of the safety and appropriateness of a treatment profile	Binary assessment (yes/no) of whether AVA-generated treatment profiles are considered clinically safe and appropriate
Identification of unsafe or inappropriate treatments in AVA-generated treatment profiles	Qualitative open-text responses detailing why a treatment profile was considered unsafe or inappropriate
Participants' confidence in their treatment profile source guesses (AI vs Human) and in their other clinical assessments	Likert-scale rating of confidence in the treatment profile source, guess (AI vs Human) and in their other clinical assessments
Clinician's interest in using an AI/ML CDSS for sepsis and ventilator management in clinical practice	Likert-scale rating of interest in using an AI/ML CDSS for sepsis and ventilator management in clinical practice
Time spent answering questions on each vignette and completing the survey	Explore associations between survey completion time, accuracy and confidence

Table 1 contains the primary and secondary endpoints and outcome measures of this study.
AI, artificial intelligence; AVA, AI ventilator assistant; CDSS, clinical decision support system; ML, machine learning.

review by a team of clinical experts to identify sepsis and ARDS and the date on which ARDS criteria were met. To identify candidate clinical cases that will serve as the basis for the vignettes, we identified encounters of patients requiring IMV who were (1) not used in the training or validation of the AVA model (to avoid using cases already seen by the model during the training process) and (2) are in the MESSI cohort to ensure they meet sepsis and ARDS criteria.

The study team reviewed eligible cases from the MESSI cohort to identify a range of clinical scenarios representing sepsis and ARDS. We initially filtered encounters to include patients whose admission to the ICU, start of IMV and onset of ARDS occurred on the same date, those without tracheostomy at IMV initiation and those admitted in 2022 (representing the most recent MESSI cases and occurring after the three primary COVID-19 waves). 17 cases met these initial criteria. Each case was then screened against the model's initial inclusion and exclusion criteria to ensure consistency. To be eligible, patients had to be ≥18 years old at admission, intubated ≥24 hours after admission and to have received IMV for at least 12 hours.

After manual review, 11 cases met all criteria. Two cases were excluded due to atypical presentations of ARDS and sepsis. This resulted in the nine most clinically straightforward cases, although they remained medically complex. Among these, one case was selected as the internal control vignette because of the patient's atypical hospital course characterised by a prolonged hospitalisation. The model was still being developed when these cases were selected, and the model outputs were not used in the vignette selection process.

Each vignette will be presented using a standardised structure and will include key elements such as the history of present illness, medical history, medical assessment findings, vital signs, laboratory results, medications, non-invasive respiratory support parameters, chest imaging and other significant or complementary findings relevant to the case and that could inform clinical decisions presented in the treatment profile. For each vignette, a human treatment profile will be developed based on the actual management decisions made by the treating clinical team recorded in the EHR at the time. The AI-generated treatment profiles will be provided by AVA for the same cases. The model outputs will remain identical for each vignette and participant. AVA is a fully deterministic model, rather than a stochastic one, and produces the same output every time given the same inputs. Both human and AI-generated treatment profiles will include ventilator management settings (eg, fractional inspired oxygen, tidal volume, respiratory rate, positive end-expiratory pressure (PEEP)), haemodynamic support drugs (eg, norepinephrine, vasopressin), sedatives and analgesics (eg, fentanyl, midazolam), antimicrobial medication categories (eg, MRSA coverage, gram-negative coverage) and other therapies (eg, neuromuscular blocking agents, high-dose steroids, inhaled pulmonary vasodilators).

Study outcomes

The primary study outcome is the accuracy of participants' guesses of the treatment profile source (AI vs human) for each clinical vignette. Thus, we will estimate participants' odds of guessing the source correctly conditional on the study arm assignment for each vignette using an equivalence design. Secondarily, we will assess participants' perceptions of the safety and appropriateness of the presented treatment profiles and compare them between study arms using the same equivalence design. Each treatment profile will be evaluated using a binary (yes/no) measure, determining whether the approach is safe and appropriate for clinical practice. Moreover, participants will be asked to identify and explain any elements of the

BMJ Open: first published as 10.1136/bmjopen-2025-106757 on 24 December 2025. Downloaded from http://bmjopen.bmj.com/ on March 21, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

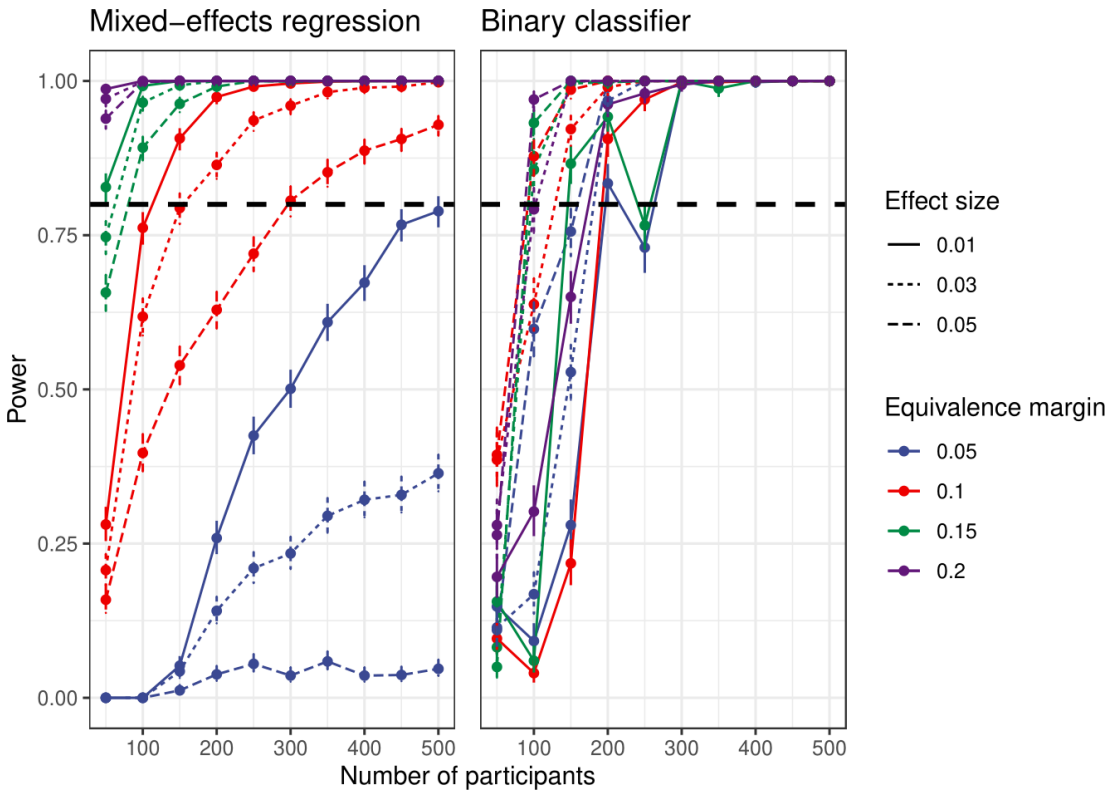


Figure 3 Power curves. Selected estimates of the study power (y-axis) for the primary and secondary statistical analyses as a function of the number of participants (x-axis). The left panel presents power curves for the primary analysis using a mixed-effects logistic regression model with an equivalence design, while the right panel shows power curves for the secondary analysis to fit a binary classifier. The effect size is the expected change in probability of a correct guess when assigned to the intervention arm. The equivalence margin is the range of probabilities (left) or C-statistic (right) above and below the baseline. Each point estimate was obtained through 1000 Monte Carlo simulations assuming each participant reviewed eight vignettes and a baseline probability of a correct guess p=0.05 (left), equivalent to a C-statistic of 0.50 (right), aligned with the Turing test design. The black, dashed line indicates the minimum desired power of 0.80.

treatment profiles they consider unsafe or inappropriate, thereby providing qualitative insights into specific areas requiring refinement before future clinical deployment. Another secondary endpoint will examine participants' confidence in distinguishing AI-generated treatment profiles from those created by human clinicians, as well as in their assessment of safety and appropriateness of the treatment recommendations. This will be measured using a 7-point Likert scale, allowing a more nuanced understanding of certainty in each guess. We will also assess clinicians' interest in using an AI/ML CDSS for sepsis and ventilator management in clinical practice using a 7-point Likert scale. This measure will help understand clinicians' openness to AI integration and their perceptions of its potential utility in real-world clinical workflows. Finally, the study will assess the time participants spend answering questions on each vignette and completing the survey, to permit exploration of associations between survey completion time, study arm and confidence (table 1).

Statistical analysis

Sample size calculations

We estimated at least 350 participants with completed surveys will be needed to achieve the study objectives. This estimate was based on two distinct but complementary sample size calculations for the primary and secondary statistical analyses (figure 3). We did not specifically power the study based on anticipated subgroup analyses, which will be exploratory.

Primary power calculation

Based on the primary statistical analysis described below, we used simulation methods to estimate the needed sample size.[12] We ran 1000 simulations of each combination of study characteristics and calculated the proportion of simulations in which $p<0.05$ and estimated binomial CIs around that proportion to estimate the expected power of that design. We simulated study characteristics including the number of participants (range 50–500), the number of study vignettes presented to each participant (6, 8 or 10; this count does not include the internal control vignette which is not associated with a randomised treatment profile), the expected increase in the probability of a correct guess when randomised to the intervention arm (ie, receiving a treatment profile from an AI compared with that from a human; range 0.01–0.10) and the equivalence margin for the hypothesis test (range 0.05–0.20). Vignette-level and participant-level random effects, and a residual error term, were each drawn from a normal distribution with $\mu=0$ and $\sigma=0.03$. A final guess of

BMJ Open: first published as 10.1136/bmjopen-2025-106757 on 24 December 2025. Downloaded from http://bmjopen.bmj.com/ on March 21, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

human or AI was sampled from a Bernoulli distribution using the total probability for each participant–vignette combination. For all simulations, we assumed a baseline probability of a correct guess as the source of a treatment profile of 0.5 aligned with the assumptions of the traditional Turing test. A logistic regression model was fit to each simulated data set such that

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + a_i + b_j + \sigma_{ij}$$

where y_{ij} is the binary outcome of a correct guess, x_{ij} is the treatment assignment for each participant–vignette, a_i and b_j are the crossed random effects for participant and vignette, respectively, and σ_{ij} is the residual error term. Based on the estimated coefficient $\beta_1$for each model, we used two one-sided t-tests (which converge with the Wald test for large N) to evaluate if that coefficient fell within the equivalence margin.[13] Using this equivalence design, the null hypothesis is that the difference in the effect D between treatment and control is:

$$H_0: D = |p_T - p_C| \geq \delta$$

Where p_C is the baseline probability of a correct guess in the human arm, here 0.5, p_T is the probability of a correct guess in the AI intervention arm and δ is the equivalence margin. Using this equivalence design, the alternative hypothesis is stated as

$$H_A: D = |p_T - p_C| < \delta$$

For each pair of one-sided tests, we used the larger of the two p values as is convention for the two, one-sided test approach.[14]

Based on this approach, we estimated that recruiting 350 participants to review eight vignettes would provide 85% (95% CI 83% to 87%) power to detect a 5% change in the probability of a correct guess from a baseline of 50% with an equivalence margin of ±10%. The study team determined that these parameters were aligned with a clinically meaningful margin and also consistent with previous Turing tests in the clinical domain.[8] We chose to include eight vignettes to balance participant burden with statistical power. In our preliminary power calculations, using fewer than eight did not provide sufficient statistical power. However, using more than eight vignettes would be uncommon for this type of study, given the complexity of the vignettes and would likely reduce the response rate and quality of responses. These sample size calculations also suggest sufficient power to reject equivalent hypothesis tests for the secondary outcomes of safety and acceptability.

Secondary power calculation

Alternatively, a Clinician Turing Test might be thought of as fitting a binary classifier to detect the source of a treatment suggestion. In this case, if the classifier (participant) is unable to adequately discriminate between treatment recommendations from a human and AI, then the AI would have 'passed' the test. Therefore, in a secondary power calculation, we estimated the sample size needed to fit a binary classifier defined as a logistic regression model with a single predictor variable (the participant's guess of human or AI) to predict the source of the treatment suggestions (a binary outcome variable indicating AI or human origin) with a baseline C-statistic of 0.5 to reflect the design of the Turing test. As above, we ran simulations across combinations of study characteristics to estimate the power, including number of participants, number of vignettes, effect size (change in the probability of correctly identifying AI recommendations), baseline C-statistics (representing performance under random guessing) and equivalence margin. We estimated that for a baseline C-statistic of 0.5, and a 5% increase in the probability of guessing correctly when presented with output from an AI system, recruiting 350 participants to review eight vignettes would yield 100% (95% CI 99% to 100%) power to reject the null hypothesis with an equivalence margin of ±0.10.

Statistical analysis plan

Given the lack of an established statistical framework for analysing any Turing test, we propose a two-tiered approach that is appropriate for this Clinician Turing Test. Statistical analyses will be conducted using R language for statistical computing[15] once data collection is complete, with the final analysis performed after reaching the target 350 complete clinician responses, expected within 6 months of recruitment period. We will use descriptive statistics to summarise categorical variables using counts and proportions, while continuous variables will be reported as means, medians, SD, ranges and IQRs, as appropriate. Where appropriate, χ^2 tests for categorical data and t-tests for normally distributed continuous variables will assess significant differences between groups. Analyses at the vignette level will include responses to the eight randomised vignettes.

The primary endpoint analysis will assess whether the true source of either AI-generated or human-generated treatment profiles is associated with the odds of participants' guesses using an equivalence design with a predefined margin of 50%±10%. Mixed-effects logistic regression models, adjusting for random effects across participants and vignettes, will estimate the odds of correct classification. The secondary analysis will estimate the performance of a binary classification model to make the same distinction, which will be evaluated using the area under the curve of the receiver operating characteristic (AUC-ROC), with an equivalence margin of ±0.10 for the C-statistic.

Secondary endpoint analyses will test the equivalence of participants' assessments of safety and appropriateness of AI-generated profiles compared with those of human origin, also using mixed-effects logistic regression. Additional evaluations will include participant confidence on answering, interest in AI CDSSs, qualitative feedback for model refinement and analysis of response time patterns. For all analyses involving 7-point Likert-scale responses,

BMJ Open: first published as 10.1136/bmjopen-2025-106757 on 24 December 2025. Downloaded from http://bmjopen.bmj.com/ on March 21, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

scores of 5 or higher will be considered high, while scores of 4 or lower will be considered low.

For primary and secondary endpoints (table 1), sensitivity analyses will exclude partially completed surveys and responses from participants who did not correctly identify AI as the source of the treatment profile in the internal control vignette, assessing any potential differences in the primary effect estimate. In an additional sensitivity analysis, we will explore whether the order of the presentation of AI recommendations in the internal control vignette affects the odds of correctly identifying the treatment recommendation source in subsequent vignettes. In this analysis, we will repeat the primary analysis but with an indicator variable for each participant–vignette observation identifying whether it occurred before or after the presentation of the control vignette. If the estimated coefficient on this vignette is significant, we will conduct additional sensitivity analyses stratified by the placement of the internal control vignette.

The final study design and statistical plan described will be reviewed and discussed with the research team before execution, any modifications will be documented and reported, and no interim changes will be made to the prespecified analysis plan once the data have been examined.

Data security and confidentiality

Clinical data used in the vignettes are based on real patient cases but are completely deidentified in their presentation in the survey instruments. All survey responses will remain anonymous during the analysis and reporting of results. Only an email identifier will be maintained on secure servers to send participant incentives on survey completion. Data management will be conducted by the research team, and a data monitoring committee is not required due to the minimal risk associated with this study.

DISCUSSION

This study protocol outlines the design and analysis plan of a Clinician Turing Test aimed at evaluating whether clinicians can distinguish between treatment recommendations generated by humans versus those generated by an AI CDSS. We detail the study protocol for a Clinician Turing Test of AVA, an AI/ML CDSS for managing sepsis and ARDS, across size geographically diverse health systems in the USA. The results from this proposed study will provide essential information for planning future deployment and evaluation of AVA in real clinical workflows and clinical trials. More broadly, this protocol for a novel Phase 1b study design provides a roadmap for the preclinical evaluation of AI/ML CDSSs in a safe, efficient and relatively low-cost environment. Thus, the Clinician Turing Test fills a much-needed gap between Phase 0 and Phase 2 studies of AI/ML CDSSs that yield important data about the clinical content of the CDSS and go beyond traditional measures of predictive performance.

Our previously published AVA model can potentially fill an important gap in clinical decision support in the ICU. It was designed specifically to provide comprehensive decision support for patients at risk of sepsis and ARDS because their management requires making many simultaneous and inter-related treatment decisions. By contrast, traditional decision support systems, even those based on AI or ML methods, typically provide a recommendation for only a single binary treatment. Using a collective intelligence and imitation learning approach,[16] AVA provides recommended treatments for ventilator settings, medication dosages and infusion rates and antibiotic coverage, among other treatments. Thus, the results of this study will provide a valid assessment of the equivalence of AVA's treatment recommendations with those treatments enacted by real clinicians. This critical information about the equivalence of AVA's recommendations, and their safety and appropriateness, based on expert clinician assessments, will be essential in planning for future clinical implementation that would otherwise not be obtainable through traditional Phase 0 studies of AI/ML CDSSs.

Prior work on Turing tests in the clinical domain has used a non-inferiority design.[8 17] Although a non-inferiority design may answer an important question about an AI system, it is less aligned with the spirit of the original Turing test, which is instead better evaluated using an equivalence design. For example, many AI-based chess systems today can easily defeat human players (and thus pass a non-inferiority test) but would quickly fail a Turing test because of their non-human-like playing style.[16 18] An AI model could fail a Turing test on either side of the equivalence margin, for example, 'much worse' than human or 'much better' than human, each with different implications for use in clinical care. Thus, we propose that a Clinician Turing Test is best evaluated by assessing the ability of an expert clinician to distinguish the source of a clinical recommendation (human vs AI) using an equivalence design.

Previous studies have used variations of the Turing test to evaluate AI tools in other contexts, including medical imaging,[7 19–21] patient perception and trust,[22] and statistical and mathematical comparisons between AI and human performance.[5] While these studies highlight the relevance of Turing tests in AI evaluation, they do not focus on clinician assessments of AI-provided recommendations. To our knowledge, no Clinician Turing Test has been conducted to evaluate an AI/ML CDSS for treatment recommendations aimed at guiding clinical reasoning in a medical context in general, or in critical care, in particular. Furthermore, many previous Turing tests of AI systems have lacked assessments of the safety and appropriateness of AI outputs and comprehensive methodological details, including clear sample size calculations, which are essential for robust AI validation.

To address this gap, we adopt a stepwise approach in which equivalence is the primary endpoint and appropriateness and safety are complementary secondary

BMJ Open: first published as 10.1136/bmjopen-2025-106757 on 24 December 2025. Downloaded from http://bmjopen.bmj.com/ on March 21, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

outcomes that build on it. Indistinguishability reflects clinical reasoning and judgement and is reinforced by assessments of safety and appropriateness that offer a more detailed validation. This sequential evaluation increases our confidence that AVA's recommendations are not just similar in appearance or tone, but truly aligned with clinical practices that practising experts would expect from other clinical experts. It also allows us to explore divergences, identifying where AVA's recommendations differ from human ones and whether those differences are meaningful with respect to safety or appropriateness. These insights are critical for establishing equipoise between AVA and current clinical practice.

Our study protocol has three key strengths. First, it bridges the evaluation gap between preclinical and clinical AI/ML by providing early evidence of AI safety and clinical performance through an equivalence test. If AVA's recommendations prove indistinguishable from those of human clinicians, this would provide strong justification for advancing an AI/ML CDSS to a Phase 2 or 3 clinical trials. Second, the design ensures patient safety by simulating clinical evaluation in a risk-free environment, offering insight into potential benefits and harms without influencing actual care. Third, this approach is resource efficient. Traditional clinical trials for AI/ML validation are costly and logistically challenging, whereas our electronic, vignette-based survey enables early evaluation at relatively low cost and high feasibility.

This work also has limitations. First, it will not account for certain human factors, such as variations in clinician decision-making styles or the impact of an ICU's high-stress workflow. Because the study will be conducted in a simulated environment, it may not fully replicate the complexities of real-world clinical settings. Second, this study is limited to cases from a single institution which may affect the generalisability of findings. Future studies should include a more diverse patient population and a broader range of evaluators to enhance applicability. Third, while clinicians will be blinded to whether recommendations are AI-generated or human-generated, unmeasured biases or differences in practice patterns may influence participant assessments, though this limitation is mitigated through randomisation. Future research should explore these effects further in more varied clinical environments.

In summary, this protocol for a Clinician Turing Test offers essential information about the AVA CDSS in a novel Phase 1b study design. This equivalence test also serves as a roadmap for safe, efficient and cost-effective evaluations of other AI/ML CDSSs prior to clinical deployment.

ETHICS AND DISSEMINATION

This study is sponsored by the University of Pennsylvania and funded by the National Institute of General Medical Sciences (NIH R35GM155262). It has been approved by the University of Pennsylvania Institutional Review Board (IRB) (Protocol #858201) and registered in ClinicalTrials.gov (NCT07025096). All participants will provide a signed informed consent to participate and for their contact information to be stored in a secure directory for potential future research before starting the survey (online supplemental material 2). Any protocol amendments will be reported directly to the study's IRB. Study findings will be published in peer-reviewed journals and presented at scientific conferences, with no identifying information disclosed. Results are expected in the spring of 2026. The full protocol is available in online supplemental material 1. Additional study material will be added to our public GitHub repository (https://github.com/weissman-lab/ClinicianTuringTest), such as our sample size calculations code and, once results are available, a summary of the survey responses at the vignette-level.

Author affiliations
[1]Palliative and Advanced Illness Research (PAIR) Center, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA
[2]Department of Anesthesia and Perioperative Care, University of California San Francisco, San Francisco, California, USA
[3]Division of Pulmonary and Critical Care, Department of Medicine, University of Michigan, Ann Arbor, Michigan, USA
[4]Center for Bioethics and Social Sciences in Medicine, University of Michigan, Ann Arbor, Michigan, USA
[5]Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, Michigan, USA
[6]Department of Veterans Affairs, VA Ann Arbor Healthcare System, Ann Arbor, Michigan, USA
[7]Department of Medicine, Emory University, Atlanta, Georgia, USA
[8]Emory Critical Care Center, Atlanta, Georgia, USA
[9]Division of Pulmonary, Allergy, Critical Care and Sleep Medicine, Emory University School of Medicine, Atlanta, Georgia, USA
[10]Divisions of Pulmonary, Critical Care, and Sleep Medicine, University of Nebraska Medical Center, Omaha, Nebraska, USA
[11]Department of Anesthesiology, University of Nebraska Medical Center, Omaha, Nebraska, USA
[12]Department of Medicine, University of Pennsylvania Division of Pulmonary Allergy and Critical Care, Philadelphia, Pennsylvania, USA
[13]Virtua Pulmonology and Sleep Medicine, Virtua Health Inc, Marlton, New Jersey, USA
[14]Department of Computer and Information Science (CIS), University of Pennsylvania, Philadelphia, Pennsylvania, USA
[15]Translational Lung Biology, Lung Biology Institute, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA
[16]Leonard Davis Institute of Health Economics, Philadelphia, Pennsylvania, USA
[17]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Acknowledgements We are grateful to the busy clinicians who took time to participate in this study.

Contributors GEW is the principal investigator, guarantor and obtained funding for the study. AAG, NSB and GEW conceptualised and wrote the protocol and protocol paper text, participating in all parts of the process. AAG, NSB, BES and GEW contributed to the study design. AAG, NSB and GEW developed the study survey on Qualtrics and manually curated the recruitment list. AAG, NSB, BES and GEW reviewed the clinical cases, wrote and structured the deidentified clinical vignettes and treatment profiles for the survey. BES and GEW contributed to the creation and training of the model (AVA). NJM provided the MESSI cohort criteria. GEW did the sample size calculations. RP, TSV, SVB, DCK, HMG and GEW identified and reviewed potential study participants for their respective sites based on inclusion criteria. BES, YL, LHU, NJM and MPK provided supplementary input to the study design, protocol conceptualisation and protocol manuscript. GEW supervised and reviewed all aspects of the work. All authors reviewed and approved the final manuscript.

BMJ Open: first published as 10.1136/bmjopen-2025-106757 on 24 December 2025. Downloaded from http://bmjopen.bmj.com/ on March 21, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

Funding The study is sponsored by the University of Pennsylvania and funded by the National Institute of General Medical Sciences (Grant: Improving sepsis care with AI-based clinical decision support (R35GM155262)). The funder had no role in the design of the study and will have no role in data collection, analysis, interpretation or manuscript writing. National Institute of General Medical Sciences (NIGMS), National Institutes of Health (NIH), 45 Center Drive, Bethesda, Maryland 20892, USA. Email: info@nigms.nih.gov

Competing interests RP reports having a research contract (AOP Health A144424) and having the following grants or contracts unrelated to this work: PCORI P0562155, NIH R01EB035028 and DOD P04575392. He also reports giving lectures for AOP Heath and being an associate editor of JAMA and Intensive Care Medicine. SVB reports having the following source of funding: R01HL175626. HMG reports receiving consulting fees from Inogen and CalmWave. MPK reports having NIH/NHLBI grants unrelated to this work. GEW reports support related to this work from NIH R35GM155262, National Academy of Sciences (SCON-10001137), Gordon and Betty Moore Foundation and John A Hartford Foundation. He also reports having grants unrelated to this work (NIH R01HL162354 and NIH R03171424), receiving a speaker honourarium from the Institute for Healthcare Improvement and getting travel support for scientific meetings by the Institute for Healthcare Improvement and National Academy of Medicine. None of the other authors declare any competing interests.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: https://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Antonia Angeli Gazola https://orcid.org/0000-0003-2470-8209
Nicholas S Bishop https://orcid.org/0000-0001-6324-8140
Romain Pirracchio https://orcid.org/0000-0002-7007-4908
Gary E Weissman https://orcid.org/0000-0001-9588-3819

REFERENCES

1 Fleuren LM, Thoral P, Shillan D, *et al*. Machine learning in intensive care medicine: ready for take-off? *Intensive Care Med* 2020;46:1486–8.
2 Plana D, Shung DL, Grimshaw AA, *et al*. Randomized Clinical Trials of Machine Learning Interventions in Health Care: A Systematic Review. *JAMA Netw Open* 2022;5:e2233946.
3 Longhurst CA, Singh K, Chopra A, *et al*. A Call for Artificial Intelligence Implementation Science Centers to Evaluate Clinical Effectiveness. *NEJM AI* 2024;1:AIp2400223.
4 Turing AM. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind* 1950;LIX:433–60.
5 Mei Q, Xie Y, Yuan W, *et al*. A Turing test of whether AI chatbots are behaviorally similar to humans. *Proc Natl Acad Sci U S A* 2024;121:e2313925121.
6 A Turing test for artificial intelligence in cancer. *Nat Cancer* 2020;1:137–8.
7 Tripathi S, Augustin A, Dako F, *et al*. Turing test-inspired method for analysis of biases prevalent in artificial intelligence-based medical imaging. *AI Ethics* 2022;2:1–9.
8 Hulman A, Dollerup OL, Mortensen JF, *et al*. ChatGPT- versus human-generated answers to frequently asked questions about diabetes: A Turing test-inspired survey among employees of a Danish diabetes center. *PLoS One* 2023;18:e0290773.
9 Weissman GE, Bishop NS, Ungar LH, *et al*. Training and validation of a pandemic clinical decision support model for ventilators using collective intelligence and imitation learning. American Thoracic Society 2024 International Conference, May 17-22, 2024 - San Diego, CA; May 2024:A5057. 10.1164/ajrccm-conference.2024.209.1_MeetingAbstracts.A5057 Available: https://www.atsjournals.org/doi/book/10.1164/ajrccm-conference.2024.C22
10 Schneider ALC, Ginestra JC, Kerlin MP, *et al*. The Complete Inpatient Record Using Comprehensive Electronic Data (CIRCE) project: A team-based approach to clinically validated, research-ready electronic health record data. *Learn Health Syst* 2025;9:e10439.
11 Palakshappa JA, Anderson BJ, Reilly JP, *et al*. Low Plasma Levels of Adiponectin Do Not Explain Acute Respiratory Distress Syndrome Risk: a Prospective Cohort Study of Patients with Severe Sepsis. *Crit Care* 2016;20:71.
12 Arnold BF, Hogan DR, Colford JM Jr, *et al*. Simulation methods to estimate design power: an overview for applied research. *BMC Med Res Methodol* 2011;11:94.
13 Tunes da Silva G, Logan BR, Klein JP. Methods for Equivalence and Noninferiority Testing. *Biol Blood Marrow Transplant* 2009;15:120–7.
14 Lakens D, Scheel AM, Isager PM. Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science* 2018;1:259–69.
15 R Core Team. R Foundation for Statistical Computing Vienna (Austria); R: A language and environment for statistical computing, 2021. Available: https://www.R-project.org/
16 Weissman GE, Ungar LH, Halpern SD. Chess Lessons: Harnessing Collective Human Intelligence and Imitation Learning to Support Clinical Decisions. *Ann Intern Med* 2023;176:274–5.
17 Hulman A, Dollerup OL, Mortensen JF, *et al*. Study protocol: ChatGPT-generated versus expert-written answers to frequently asked questions about diabetes - an e-survey among all employees of a Danish diabetes center. *Figshare* 2023.
18 Eisma YB, Koerts R, de Winter J. Turing tests in chess: An experiment revealing the role of human subjectivity. *Computers in Human Behavior Reports* 2024;16:100496.
19 Myong Y, Yoon D, Kim BS, *et al*. Evaluating diagnostic content of AI-generated chest radiography: A multi-center visual Turing test. *PLoS One* 2023;18:e0279349.
20 Wang Z, Gao H, Wang X, *et al*. A multi-Task Learning based applicable AI model simultaneously predicts stage, histology, grade and LNM for cervical cancer before surgery. *BMC Womens Health* 2023;24:345.
21 Wu Y, Kang K, Han C, *et al*. A blind randomized validated convolutional neural network for auto-segmentation of clinical target volume in rectal cancer patients receiving neoadjuvant radiotherapy. *Cancer Med* 2023;12:e16978.
22 Nov O, Singh N, Mann D. Putting ChatGPT's Medical Advice to the (Turing) Test: Survey Study. *JMIR Med Educ* 2023;9:e46939.

BMJ Open: first published as 10.1136/bmjopen-2025-106757 on 24 December 2025. Downloaded from http://bmjopen.bmj.com/ on March 21, 2026 by guest. Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.