Evaluation of latent health risk prediction models: A protocol for a mixed-methods study using clinical triage as a vehicle for comparison and discussion

Morgan Roberts, Otso Pelkonen, Diana Shamsutdinova, Saskia, C Sanderson, Pawel Renc, Hugh Logan Ellis

Submitted to: JMIR Research Protocols
on: October 07, 2025

Table of Contents

Evaluation of latent health risk prediction models: A protocol for a mixed-methods study using clinical triage as a vehicle for comparison and discussion

Morgan Roberts[1*]; Otso Pelkonen[2*]; Diana Shamsutdinova[3]; Saskia, C Sanderson[3, 4]; Pawel Renc[5, 6, 4]; Hugh Logan Ellis[3]

[1] Somerset NHS Foundation Trust Taunton GB
[2] n/a n/a FI
[3]Department of Biostatistics and Health Informatics Institute of Psychiatry, Psychology and Neuroscience King's College London London GB
[4] AGH University of Krakow Kraków PL
[5] Massachusetts General Hospital Boston US
[*]these authors contributed equally

Corresponding Author:
Morgan Roberts

Somerset NHS Foundation Trust
Parkfield drive
Taunton
Taunton
GB

Abstract

Background: Determining clinical urgency and resource allocation within acute patient populations is complex. Tools are being developed to capture global assessments of patient health beyond disease-specific scores, aiming to provide dynamic assessments incorporating both baseline physiological reserve and immediate illness severity

Objective: This study evaluates two contrasting approaches to latent health measurement: FI-lab, a transparent algorithmic tool using bottom-up aggregation of laboratory abnormalities, and ETHOS-ARES, a transformer-based model using top-down learning from electronic health records.

Methods: This two-phase mixed-methods study uses clinical triage scenarios with ?30 clinicians sampled across hospital roles (physicians, surgeons, critical care nurses, advanced practitioners). Phase 1 compares unaided clinician judgments of severity and clinical urgency against model outputs using Spearman's rank correlation (? ? 0.70 indicates good agreement as primary outcome). A novel "clinical Turing test" assesses whether model rankings are statistically distinguishable from clinician assessments. Phase 2 allows clinicians to incorporate model outputs into identical tasks, measuring anchoring effects through within-person pre/post comparison. Semi-structured interviews explore clinical utility, trust, and perceived limitations. Case materials derive from MIMIC-IV-ED, presented as slide decks containing emergency department documentation, examination findings, laboratory results, and imaging reports. Qualitative analysis follows the Framework Method with dual independent coding.

Results: Data collection is planned for October-November 2025, with analysis to follow in December 2025.

Conclusions: We anticipate findings will quantify agreement between model outputs and clinician consensus, measure any anchoring effects from model exposure, and generate insights from qualitative data regarding clinical utility, feasibility, and factors influencing clinician trust and adoption of different approaches to latent health measurement. Clinical Trial: n/a

(JMIR Preprints 07/10/2025:85437)
DOI: https://doi.org/10.2196/preprints.85437

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Evaluation of latent health risk prediction models: A protocol for a mixed-methods study using clinical triage as a vehicle for comparison and discussion

Morgan Roberts[1]*
Email: morganroberts@doctors.org.uk
ORCID: https://orcid.org/0009-0006-0351-7410
Affiliation: Somerset NHS foundation trust

Otso Pelkonen[2]*
Email: otso.pelkonen@gmail.com
ORCID: 0009-0003-8460-0133
Affiliation: N/A

Diana Shamsutdinova[3]
Email: diana.shamsutdinova@kcl.ac.uk
ORCID: https://orcid.org/0000-0003-2434-3641
Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

Saskia C. Sanderson[3,4]
Email: saskia.sanderson@kcl.ac.uk
ORCID: https://orcid.org/0000-0001-8427-724X
Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom
Honorary Associate Professor, Department of Behavioural Science and Health, University College London, London, United Kingdom

Pawel Renc[5,6,7]

email: prenc@mgh.harvard.edu
ORCID: https://orcid.org/0000-0002-0487-7454
Massachusetts General Hospital, Boston, MA, USA
Harvard Medical School, Boston, MA, USA
AGH University of Krakow, Kraków, Poland

Hugh Logan Ellis[3]†
Email: hugh.logan_ellis@kcl.ac.uk
ORCID: https://orcid.org/0000-0002-6428-0158
Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

*Joint first authors (contributed equally)

Abstract

Background: Determining clinical urgency and resource allocation within acute patient populations is complex. Tools are being developed to capture global assessments of patient health beyond disease-specific scores, aiming to provide dynamic assessments incorporating both baseline physiological reserve and immediate illness severity.

Objective: This study evaluates two contrasting approaches to latent health measurement: FI-lab, a transparent algorithmic tool using bottom-up aggregation of laboratory abnormalities, and ETHOS-ARES, a transformer-based model using top-down learning from electronic health records.

Methods: This two-phase mixed-methods study uses clinical triage scenarios with ≥30 clinicians sampled across hospital roles (physicians, surgeons, critical care nurses, advanced practitioners). Phase 1 compares unaided clinician judgments of severity and clinical urgency against model outputs using Spearman's rank correlation ($\rho \geq 0.70$ indicates good agreement as primary outcome). A novel "clinical Turing test" assesses whether model rankings are statistically distinguishable from clinician assessments. Phase 2 allows clinicians to incorporate model outputs into identical tasks, measuring anchoring effects through within-person pre/post comparison. Semi-structured interviews explore clinical utility, trust, and perceived limitations. Case materials derive from MIMIC-IV-ED, presented as slide decks containing emergency department documentation, examination findings, laboratory results, and imaging reports. Qualitative analysis follows the Framework Method with dual independent coding.

Results: Data collection is planned for October-November 2025, with analysis to follow in December 2025.

Conclusions: We anticipate findings will quantify agreement between model outputs and clinician consensus, measure any anchoring effects from model exposure, and generate insights from qualitative data regarding clinical utility, feasibility, and factors influencing clinician trust and adoption of different approaches to latent health measurement.

Keywords

artificial intelligence, clinical decision support systems, triage, machine learning, electronic health records, deep learning, frailty, clinical validation, mixed methods research, human-computer interaction

Section* : Qualitative methods?

Article Type*: Protocol

Introduction

Hospital clinicians working on-call frequently face the challenge of deciding which patient to assess next from among multiple competing demands. A key component of this triage decision involves identifying the sickest patient; those most at risk of deterioration or in greatest need of urgent intervention. However, clinical urgency is not determined by illness severity alone. Clinicians must also weigh disease trajectories, available resources, the acuity of other patients, and the likelihood that early intervention will alter outcomes [1][2]. These multifaceted judgments, though central to clinical practice, are difficult to quantify and communicate.

To support such decisions, numerous tools have been developed to measure patient condition, ranging from organ-specific scores like the Glasgow Coma Scale (GCS) for head injury [3] and the HEART score for chest pain [4], to more general assessments of physiological disturbance. The National Early Warning Score (NEWS) exemplifies the latter, using bedside vital signs to provide a standardised measure of acute illness severity that has been widely adopted in the UK and internationally [8]. While NEWS offers transparency and allows clinicians to factor in clinical context, it relies solely on vital signs and cannot access the broader information available in electronic health records (EHR), where early markers of physiological disruption may be evident before vital signs deteriorate. Williams's 2022 review of NEWS anticipated the rise of more complex, data-driven scoring systems while emphasising the need for such tools to inform clinical decisions rather than replace them [8].

With EHRs now ubiquitous in healthcare settings, there is growing interest in whether richer data sources can provide more comprehensive assessments of patient health. We sought to evaluate how clinicians perceive and use two contrasting approaches to this challenge. Bottom-up methods aggregate simple indicators across multiple domains to build composite pictures of health [5]. The Frailty Index-laboratory (FI-lab) exemplifies this approach, calculating the proportion of abnormal laboratory results and offering transparency grounded in established frailty index methodology [6]. Top-down methods employ advanced computational techniques to learn complex health representations from large datasets [5]. ETHOS-ARES (Enhanced Transformer for Health Outcome Simulation, Adaptive Risk Estimation System) uses transformer architecture to tokenise health records, create multidimensional patient representations, and generate chronological timelines [7], identifying patterns across large patient cohorts to predict future outcomes and highlight risk states [2].

Evaluating such tools presents distinct challenges. No single validated reference standard exists for 'latent health status', and successful clinical implementation requires not only predictive accuracy but also interpretability, usability, and clinician trust. Our previous work identified key factors clinicians consider important, including the ability to compare current health against premorbid baseline and tools that guide rather than replace clinical judgment [9].

This study therefore uses clinical triage, a scenario familiar to on-call clinicians, as a vehicle to address two questions.

1. Do these tools' assessments of patient health align with those of experienced clinicians, whom we treat as our reference standard in this context?

2. Do clinicians find the tools useful when incorporated into triage decisions?

Primary objectives: Determine whether each tool's severity scores and rankings align with clinician consensus on (a) resolved severity ranking and (b) clinical urgency (order in which patients should be assessed).

Secondary objectives: Assess whether tool exposure anchors clinicians toward tool outputs; test whether tool rankings are distinguishable from human rankings (clinical Turing test); and explore through qualitative interviews how clinicians interpret and use the tools, their points of agreement or disagreement, and their perceptions of utility and limitations.

Methods

Study design

Participants

Clinicians: ≥30 currently practicing clinicians. We will aim to sample across roles: hospital physicians (internal medicine, emergency medicine), surgeons, across a range of grades as well as, ICU outreach/critical care liaison nurses and advanced clinical practitioners.

Inclusion: participates in on- call triage decisions; English- speaking; consent.

Exclusion: prior involvement in case selection.

Clinical cases

Cases come from the MIMIC-IV-ED test split used during ETHOS-ARES development (held-out at training). MIMIC-IV-ED (Medical Information Mart for Intensive Care IV - Emergency Department) is a freely accessible, de-identified database of emergency department encounters from Beth Israel Deaconess Medical Center, a large academic medical center in Boston, Massachusetts. The database contains comprehensive clinical data including patient demographics, triage information, vital signs, laboratory results, imaging reports, and clinical notes collected during routine clinical care.

Materials are delivered as a slide deck rather than a full electronic patient record (EPR). Each case slide set includes: ED clerking; a structured history and physical examination captured before ward arrival (sourced via the discharge summary but restricted to history and examination content only to avoid outcome leakage); laboratory results; and radiology reports where applicable. Each opening slide contains a brief handover-style note highlighting pertinent aspects of medical history and investigation results. Inpatient progress notes and discharge outcomes are not shown. Separate instruction slides explain the tasks, and brief explanatory slides introduce FI-lab or ETHOS-ARES as applicable depending on participant randomisation.

Tools Under Evaluation

- **FI- lab:** non- AI algorithmic index; the proportion of available routine lab tests outside reference range aggregated at the time point. Transparent, setting- agnostic.
- **ETHOS-ARES:** transformer- based foundation model producing a composite severity/acuity score from tokenised EHR data corresponding to the patient timeline relevant to the task.

Interview structure

Phase 1: No tool exposure

The interview is divided into three phases. For an overview, see table 1. In phase 1 we propose model evaluation by comparison against the assessments of working clinicians, what we consider in this situation to be our 'gold standard'. These clinicians will be given the information for 10 patients to consider in the context of a realistic clinical triage scenario. During the first task they will be asked to score patients using a modified form of the ASA physical status classification. Here we will introduce a decimalised version of this well-known score to allow clinicians to more accurately make their assessment of a patient's health status. This will be compared directly with the model outputs

The second task of phase 1 will require our clinicians to rank the same patients in order of clinical urgency. In effect, listing them in order of 'most urgently in need of care' to 'safe to wait'. We will compare this ranking to that of the previous task, the average highest perceived severity to lowest and the ranked scores from our models. This task more accurately reflects the work of a clinician and through it we hope to see if our models differ from our 'gold standard'.

Phase 2: Tool exposure and integration

For phase 2 of our study we seek to learn how clinicians interpret model output and incorporate it into their decision making. Clinicians will receive information for a further 10 patients, the tasks will be identical, however, in this instance the clinicians will be introduced to one of the two models before being shown model outputs for each patient. Clinicians will be randomised into receiving either the FI-lab output or the ETHOS-ARES output for their second 10 patients.

Phase 3: Cross- exposure and interviews

After clinicians have submitted their scores and rankings for the second 10 patients, they will be shown the alternate model output for comparison alongside both model outputs for the first 10 patients for their own review. Following this, we will follow a semi-structured interview to gather qualitative data about the clinicians' perceptions regarding the utility, benefits and perceived risks of the tool in question.

Introduction	Randomisation	Phase 1	Phase 2	Phase 3
Participating clinicians are given a series of introductory slides that explain our study and the models under evaluation. They will be familiarised with the 'handover' scenario.	Clinicians are randomised to one of four groupings that decide the order in which they will see patient groups and model outputs.	Task 1a: Overall severity score (1–5) per case (ASA- style anchors provided). Task 1b': Resolved severity ranking of the same 10 cases (participants break ties to produce a strict 1–10 order). Task 2: clinical urgency ranking (1 = see first … 10 = see last).	Short slide- deck explainer on the assigned tool. Display tool outputs (FI- lab score or ETHOS-ARES severity score) alongside each case. Repeat: Task 1a, Task 1b and Task 2 allowing model outputs to be incorporated to clinical reasoning.	Brief view of the *other* tool's outputs for 2–3 exemplar cases; short survey comparing perceived utility. Semi- structured interview (15–25 min) covering their perception of the tool, what was good, did they trust it, what was bad

Table 1: Overview of interview structure and methodology.

Quantitative methods

Phases 1 and 2 of our study yield quantitative data as illustrated in Figure 1 showing the participant groups. A corresponding dataset is collected for each participant's severity scores, severity rankings and clinical urgency rankings. For example, a clinician randomised into the AB-FI group would score Set A cases unexposed (without seeing the models' scores), and Set B while seeing the output of the FI-lab algorithm.

<table>
<tr><th></th><th></th><th colspan="20">Patient cases</th></tr>
<tr><th></th><th></th><th colspan="10">Set A</th><th colspan="10">Set B</th></tr>
<tr><th>Participant (clinician)</th><th>Group</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th><th>15</th><th>16</th><th>17</th><th>18</th><th>19</th><th>20</th></tr>
<tr><td>1
2
3
4
5</td><td>AB-FI</td><td colspan="10" rowspan="2">Unexposed task</td><td colspan="10">Exposed task (FI-lab)</td></tr>
<tr><td>6
7
8
9
10</td><td>AB-ET</td><td colspan="10">Exposed task (ETHOS-ARES)</td></tr>
<tr><td>11
12
13
14
15</td><td>BA-FI</td><td colspan="10">Exposed task (FI-lab)</td><td colspan="10" rowspan="2">Unexposed task</td></tr>
<tr><td>16
17
18
19
20</td><td>BA-ET</td><td colspan="10">Exposed task (ETHOS-ARES)</td></tr>
</table>

Figure 1. Illustration of participant grouping. For illustrative purposes, this example only shows 20 clinicians. Here, the first 10 clinicians are those who were randomized into unexposed (unassisted) scoring of the Set A cases, and exposed to one of the models for Set B (AB arm). Similarly, the latter 10 clinicians shown would first work through Set B cases without model exposure before moving onto Set A with model exposure (BA arm). Unexposed scorings, shown on a white background, are done first. Exposed scorings shown on grey background are from clinicians exposed to FI-lab output; blue background similarly for ETHOS-ARES.

Primary and secondary outcomes

For tabular summaries of primary and secondary outcomes and their corresponding statistical methodology, see tables 2 and 3 respectively.

Primary outcomes: agreement between tool- generated **r**ankings and polled clinician consensus for:

1A: Clinical urgency ranking (the order in which the patients should be seen)

1B: Resolved severity ranking (derived from severity scores with ties broken by the clinician)

Statistical analyses:

Main hypothesis: model scores are similar to the clinicians consensus. In particular, there is a statistically significant correlation between the clinicians consensus (mean scores) and model scores.

Method:

We will compute the pooled clinical consensus by averaging raw unexposed severity scores provided by the clinicians randomised into the AB arm for Set A, and across the BA arm for Set B. These averages will be used to create a consensus ranking of severity. A Spearman correlation test will be used to estimate the correlation between the consensus and ETHOS and FI-lab model rankings with corresponding confidence intervals, where subsequently a p-value < 0.05 would indicate that we reject the hypothesis of no correlation between the consensus and model scores. Similarly, we will conduct correlation analysis for clinical urgency by comparing the model-provided rankings to the clinician-provided consensus clinical urgency rankings.

In total, this will result in 4 correlation tests: ETHOS vs consensus clinical urgency, ETHOS vs consensus severity ranking, FI-lab vs consensus clinical urgency, and FI-lab vs consensus severity ranking.

Power calculations. Power calculations conducted using the "pwr" and "pwrss" R packages [10, 14], indicate that assessing 20 patients gives 80% power to detect correlations > 0.6 for Spearman correlation. This power is reasonable, given that 1) the pilot scores indicate correlations 0.62-0.90 (varying between the models and severity and urgency marks), 2) a higher bar for the detectable correlation will ensure a strong indication that the Fi-Lab or ETHOS scorings align with the practitioners' consensus.

Secondary outcomes

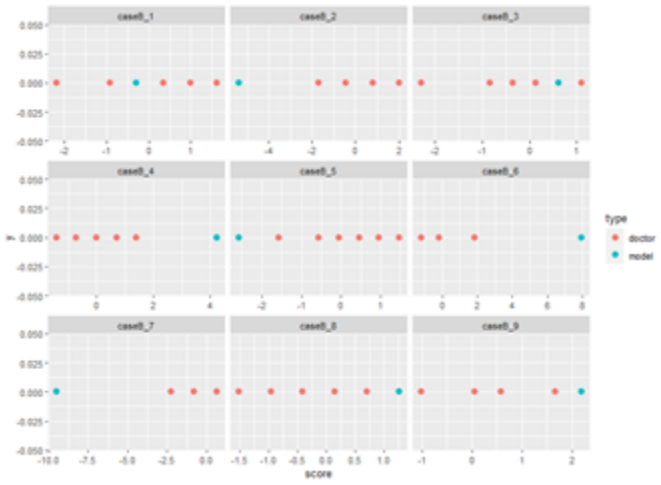
1. **Turing test: model rankings are not distinguishable from clinicians rankings**

· Hypothesis 1. Model rankings are *not* the outliers in the observed distribution of the clinician ratings.

· Method: For each participant and for the model, we compute z-scores of how far their estimates are from the participant's unexposed consensus (excluding the model), expressed in standard deviations. We will then visually inspect the model and participants' z-scores to investigate whether the models' z-score are the outliers (see Figure 2).

· Permutation test with a classifier regression (e.g. logistic regression). Here, we test whether a classifier regression can predict the label "clinican" or "model" using z-score as predictor. First, a logistic regression is fitted, and we compute an estimation for the probability that the model's scores belong to the label "model", p_{model} . As we only have 1 vector of scores for the model, we cannot perform a standard train/test evaluation and instead utilise a permutation test. In this, we will choose a random clinician to become a "model", and compute their separability from the rest of the dataset in a similar way, yielding comparator statistics p_i for all participating clinicians. We will conclude that the model is statistically distinguishable from the participants if $p_{model} > p_i$ for more than 90% of the participants.

Figure 2. Illustrative visual inspection of the z-scores of the unexposed participant and model scores.



2. Anchoring effect of the models

· Hypothesis 2A (group-level impact): Exposing participants to the model rankings makes their consensus closer to the model's scores. That is, the mean absolute difference between the consensus and the model scores is lower for the exposed consensus than for the unexposed.

· Method: As in the test for the main hypothesis, we first compute the pooled consensus rank for each of the cases, averaging unexposed rankings across the participants, e.g. (u_1, …, u_{20}), for each of the rated cases. Similarly, the exposed consensus is computed, (e_1, …, e_{20}). The absolute differences between the model scores (m_1, …, m_{20}) and the exposed and unexposed consensus will represent the closeness to the model in these scenarios, $d_{unexposed}$ = ($abs(u_1 - m_1)$, …, $abs(u_{20} - m_{20})$), $d_{exposed}$ = ($abs(e_1 - m_1)$, …, $abs(e_{20} - m_{20})$).

· Finally, we perform a one-sided paired t-test for $d_{exposed}$ and $d_{unexposed}$ to investigate whether the difference is smaller for the exposed consensus than for the unexposed, $Mean(d_{exposed}) < Mean(d_{unexposed})$.

· Power calculations show that with 30 clinicians (15 per arm) and 20 cases, we will be able to detect (a rather large) difference of over 2 rank points with 80% power. That is, if this test does not reach statistical significance, that could be due to insufficient power to detect a smaller impact of the models on consensus.

· Hypothesis 2B (individual-level impact): Clinician-level correlation with the model rankings is stronger for the exposed rankings than for the unexposed.

· Method: For each participant, we first compute correlation of their ranks with the model scores. Participants in AB arm will have "unexposed" correlation with model scores for Set A and "exposed" with Set B, and vice versa for BA, therefore each participant will have one "unexposed" and one "exposed" correlation. Second, we perform a paired one-sided t-test to investigate if the mean correlation has become higher in the exposed scenario. P-value < 0.05 will imply a statistically significant difference.

3. Correlation between the models

- Hypothesis 3: There is a significant correlation between the models' rankings.

- Method: Spearman's correlation test between the ETHOS and Fi-LAB scoring for the 20 cases. P-value < 0.05 will indicate that we can reject the hypothesis that the correlation is insignificant (zero). Similar power calculations as for the main hypothesis imply that assessing 20 patients gives 80% power to detect correlations > 0.6 as statistically significant.

4. Correlation between illness severity and clinical urgency as rated by clinicians

- Hypothesis 4: There is a statistically significant correlation between the clinicans consensus for severity and urgency.

- Method: Spearman's correlation test between the consensus severity and consensus urgency scores. Both consensus scores will be computed using the unexposed results. Similar power calculations as for the main hypothesis imply that assessing 20 patients gives 80% power to detect correlations > 0.6 as statistically significant.

Primary outcome	**Corresponding hypothesis**	**Methodology**
1A: Agreement between tool- generated **r**ankings and polled clinician consensus for clinical urgency (the order in which the patients should be seen)	Hypothesis 1A: There is no significant correlation between the models' clinical urgency rankings and the clinicians' consensus	Spearman correlation for both ETHOS-ARES vs consensus clinical urgency and FI-lab vs consensus clinical urgency
1B: Agreement between tool- generated **r**ankings and polled clinician consensus for resolved severity ranking	Hypothesis 1B: There is no significant correlation between the models' resolved severity rankings and the clinicians' consensus	Spearman correlation for both ETHOS-ARES vs consensus severity ranking and FI-lab vs consensus severity ranking

Table 2: Summary of primary outcomes and statistical methodology

Secondary outcome	Corresponding hypothesis	Methodology
2A: Clinical Turing test i.e. the degree to which model output is distinguishable from clinicians' assessment	Hypothesis 2A: Model rankings are not the outliers in the observed distribution of the clinicians' ratings	Z-score analysis: compute distance of model and participant rankings from unexposed consensus. Permutation test with logistic regression to test if a classifier can distinguish "clinician" vs "model" label using z-scores. Model is distinguishable if $p_{model} > p_i$ for >90% of participants.
3A: Anchoring effect of model on consensus level	Hypothesis 3A: Correlation of the consensus rankings with the model rankings is stronger for the exposed rankings than for unexposed	Compute absolute differences between model scores and both unexposed and exposed consensus rankings; one-sided paired t-test to test if exposed consensus is closer to model than unexposed consensus
3B: Anchoring effect of model on individual level	Hypothesis 3B: Participant-level correlation with the model rankings is stronger for the exposed rankings than for unexposed	Calculate each participant's correlation with model scores when unexposed vs exposed; paired one-sided t-test to test if mean correlation increases in the exposed scenario ($p < 0.05$ indicates significant anchoring)
4A: Correlation between models	Hypothesis 4A: There is a significant correlation between the models' rankings	Spearman correlation test between the ETHOS-ARES and FI-lab scores for the 20 cases
5A: Correlation between clinicians' assessment of severity and clinical urgency	Hypothesis 5A: There is a significant correlation between clinicians' assessment of a patient's severity and clinical urgency	Spearman correlation test between the consensus severity and consensus urgency scores; both consensus scores will be computed using unexposed scores

Table 3: Summary of secondary outcomes and statistical methodology

Qualitative

Interviews will be audio-recorded. The audio recordings will be transcribed verbatim. The transcripts will be analysed by the research team using thematic analysis. Published guidance, such as Braun & Clarke's steps to thematic analysis [11], will be followed when analysing the data to maintain scientific rigour. NVivo[12] (QSR International, Australia) will be used to manage the data and facilitate coding.

We will use the Framework Method. A short a priori codebook will be drafted from the aims (integration of tool outputs; agreement/disagreement reasons; perceived anchoring; confidence/trust; workflow impact) and refined on pilot transcripts. Two researchers will (i) independently review all transcripts, (ii) joint- code ~20% to calibrate the codebook and resolve differences by discussion, then (iii) code the remaining data independently. Data will be charted into a matrix (cases × codes) to compare across clinicians and cases. We will present exemplar quotations, conduct deviant- case analysis, and maintain an audit trail (codebook changes and decisions). We will create joint displays aligning quantitative disagreement with qualitative explanations. Reporting will follow COREQ[13].

Data Management & Security

- De- identified case materials; secure, access- controlled storage.
- Audio recordings transcribed; identifiers removed; transcripts stored securely

Ethics: Approval granted by the King's College London (KCL) Research Ethics Office, Minimal Risk Registration MRSP- 24/25- 48707.

Consent & confidentiality: Written informed consent from clinician participants; case materials are in MIMIC have already been de- identified; no patient- identifiable data will be shown.

Data Sharing

Aggregate results and analysis scripts will be shared via Open Science Framework.

Study Timeline

Phase	Timeline	Status
Study design and development	April- September 2025	Completed
Ethics approval	June 2025	Obtained
Protocol publication	October 2025	In submission

Data collection	October-November 2025	Planned
Data analysis	December 2025	Planned
Results manuscript	December 2025	Planned

Declaration of Sources of Funding:
This work was supported by a Dalhousie Department of Medicine Research Fellowship and King's College London Centre for Doctoral Studies. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Acknowledgements
The authors wish to thank Dr Timothy Bonnici for suggesting this study design to test latent health measurement tools. We are grateful to Dr Zina Ibrahim for her ongoing supervision and methodological expertise.
We acknowledge the use of large language model (LLM) tools (Claude, Anthropic) for assistance with proofreading, formatting, and manuscript preparation. These tools were not used to generate novel scientific content, analysis, or interpretation of results.

Author Contributions

HLE conceived the original study concept. MR, OP, and HLE collaboratively designed the study methodology and refined the protocol. DS provided expert statistical guidance and developed the quantitative analysis plan. SCS provided expert qualitative methodological input and developed the qualitative analysis framework. PR contributed the ETHOS-ARES model outputs and expertise on transformer-based health prediction. All authors contributed to manuscript preparation, reviewed and approved the final version, and agree to be accountable for all aspects of the work.

Bibliography

1. McDermid, Robert C, and Sean M Bagshaw, 'Physiological Reserve and Frailty in Critical Illness', in Robert D. Stevens, Nicholas Hart, and Margaret S. Herridge (eds), *Textbook of Post-ICU Medicine: The Legacy of Critical Care* (Oxford, 2014; online edn, Oxford Academic, 1 July 2014), https://doi.org/10.1093/med/9780199653461.003.0028, accessed 12 Aug. 2025.
2. Logan Ellis, H., Palmer, E., Teo, J.T. *et al.* The early warning paradox. *npj Digit. Med.* 8, 81 (2025). https://doi.org/10.1038/s41746-024-01408-x
3. Jain S, Margetis K, Iverson LM. Glasgow Coma Scale. [Updated 2025 Jun 23]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK513298/

4. Smith LM, Ashburn NP, Snavely AC, Stopyra JP, Lenoir KM, Wells BJ, Hiestand BC, Herrington DM, Miller CD, Mahler SA. Identification of very low-risk acute chest pain patients without troponin testing. Emerg Med J. 2020 Nov;37(11):690-695. doi: 10.1136/emermed-2020-209698. Epub 2020 Aug 4. PMID: 32753395; PMCID: PMC7952041.
5. Michelle Wang, Madhumita Sushil, Brenda Y Miao, Atul J Butte, Bottom-up and top-down paradigms of artificial intelligence research approaches to healthcare data science using growing real-world big data, Journal of the American Medical Informatics Association, Volume 30, Issue 7, July 2023, Pages 1323–1332, https://doi.org/10.1093/jamia/ocad085
6. Sapp DG, Cormier BM, Rockwood K, Howlett SE, Heinze SS. The frailty index based on laboratory test data as a tool to investigate the impact of frailty on health outcomes: a systematic review and meta-analysis. Age Ageing. 2023 Jan 8;52(1):afac309. doi: 10.1093/ageing/afac309. PMID: 36626319; PMCID: PMC9831271.
7. Pawel Renc, Michal K Grzeszczyk, Nassim Oufattole, Deirdre Goode, Yugang Jia, Szymon Bieganski, Matthew B A McDermott, Jaroslaw Was, Anthony E Samir, Jonathan W Cunningham, David W Bates, Arkadiusz Sitek, Foundation model of electronic medical records for adaptive risk estimation, GigaScience, Volume 14, 2025, giaf107, https://doi.org/10.1093/gigascience/giaf107
8. Williams B. The National Early Warning Score: from concept to NHS implementation. Clin Med (Lond). 2022 Nov;22(6):499-505. doi: 10.7861/clinmed.2022-news-concept. PMID: 36427887; PMCID: PMC9761416.
9. Hugh Logan Ellis, Lubna Alabdallat, Elyse Futrell et al. What Clinicians Want AI to Measure, and How They Want It Done, 23 April 2025, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-6346779/v1]
10. Champely S (2020). _pwr: Basic Functions for Power Analysis_. R package version 1.3-0, <https://CRAN.R-project.org/package=pwr>.
11. Braun, V. & Clarke, V. “Using Thematic Analysis in Psychology.” Qualitative research in psychology 3.2 (2006): 77–101. https://doi.org/10.1191/1478088706qp063oa
12. Lumivero (2023) NVivo (Version 14) www.lumivero.com
13. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care. 2007;19(6):349-357.
14. Bulus, M., & Jentschke, S. (2025). pwrss: Statistical Power and Sample Size Calculation Tools. R package version 1.0.0. https://doi.org/10.32614/CRAN.package.pwrss