



A color-coded urgency bar and intervention thresholds in sequential health decisions

Kolesnikov A

Submitted: June 22, 2026, Revised: version 1, July 23, 2026

Accepted: August 1, 2026

Abstract

Delayed help-seeking is often viewed as a failure to recognize warning signs. However, health decisions are not simple recognition tasks: early symptoms may be ambiguous, warning signs accumulate over time, and individuals must decide when the available evidence becomes sufficient to justify action. This study examined whether different presentations of escalating health-risk information were associated with when people chose to seek help during hypothetical sequential decision tasks. Two hundred and seventy-seven participants completed four scenarios in which they repeatedly chose either to continue monitoring or to seek help as symptoms accumulated. Six information formats were compared: symptoms alone; symptoms with a numerical urgency score, urgency label, visual bar, or color-coded bar; and a multimodal display combining symptoms, score, label, and color. Author-generated urgency scores were used solely to construct internally ordered escalation trajectories and have no external clinical interpretation. Of 1,108 scenario sequences, 25 ambiguous sequences were excluded, leaving 1,083 analyzable scenarios and 3,047 stage-level at-risk decisions. A participant-clustered logistic generalized estimating equation model adjusted for experimental urgency, trajectory, and scenario pair. Compared with symptoms alone, the tested green-yellow-orange-red urgency bar was associated with higher odds of intervention (OR = 2.06, 95% CI 1.52–2.80), and it also exceeded the otherwise identical non-colored visual bar (OR = 1.70, 95% CI 1.26–2.28; Holm-adjusted $p = .0027$). The multimodal display did not differ from symptoms alone (OR = 0.99, 95% CI 0.72–1.37), and the colored-bar association remained in a score-independent sensitivity analysis. These findings suggest that behavioral intervention thresholds depend not only on the amount of information presented but also on how that information is organized and communicated. Within this sequential vignette task, the tested color-coded urgency bar was associated with earlier intervention than simpler or denser alternatives, supporting the importance of interface design in communicating escalating health risk. These findings do not establish clinical benefit or generalize beyond the specific palette, interface, and hypothetical scenarios studied.

Keywords

Risk communication, Decision thresholds, Help-seeking behavior, Clinical decision support, Digital health, Human-computer interaction, Symptom checker, Visual risk communication, Sequential decision-making, Health informatics

Alex Kolesnikov, Eton College, Windsor, Berkshire, SL4 6DW, United Kingdom. alexkolesnikov2010@gmail.com

1. Introduction

Public health communication often begins from a knowledge-deficit assumption: people delay seeking help because they fail to recognize warning signs, underestimate their importance, or lack knowledge about what serious illness looks like. Improving symptom awareness is therefore important. Yet understanding that something may be wrong does not necessarily tell a person whether the situation is serious enough to warrant immediate action. Models of help-seeking describe a multi-stage process in which bodily changes are noticed, interpreted, evaluated as potentially requiring care, and eventually acted upon (1–4). Delay may therefore occur even after a possible warning sign has been identified.

This study focuses on that threshold problem. An intervention threshold is the point at which accumulating evidence becomes action-worthy: continued monitoring is no longer judged acceptable and protective action is taken. Such decisions are often made under uncertainty. A single symptom may be ambiguous, while a developing combination of symptoms may become increasingly concerning. The practical question is therefore not only whether warning signs are understood, but when the emerging pattern crosses a person's threshold for seeking help.

This issue is increasingly relevant not only to digital health but also to emerging AI-assisted safeguarding systems. Symptom checkers, triage websites, app-based health tools, and AI-driven early warning systems designed to detect accumulating risk may need to communicate changing levels of concern through an interface and support decisions about when continued

observation should give way to intervention. Research has documented substantial variation in the diagnostic and triage performance of symptom checkers (5–7), while studies of user experience have identified clarity, usability, trust and the presentation of recommendations as important influences on how such tools are understood and used (8). A system may therefore provide medically relevant information while still leaving the user uncertain about whether waiting has become unsafe. Technical accuracy and effective communication are related but distinct design problems.

Risk-communication research shows that presentation format can shape interpretation (9–12). Numerical formats provide precision, but only when users understand the scale and its reference points. Verbal labels can reduce interpretive effort but may conceal differences between values. Graphical representations can make magnitude or progression easier to see. Color can add a salient categorical signal, although its effect may depend on the palette, learned conventions and surrounding design. A recent review of clinical-deterioration decision-support systems found that color was frequently used to classify risk, often through traffic-light schemes, but also identified substantial variation in implementation (13). Evidence from one palette or graphical design should therefore not be treated as evidence for color coding in general.

More information is not necessarily better. Cognitive-load and multimedia-learning research suggests that users have limited capacity to integrate several simultaneous representations, particularly when those

representations compete for attention or fail to provide a clear decision structure (14,15). A format combining symptoms, a number, a verbal label, a graphical bar and color may appear more complete than a simpler alternative without necessarily providing a clearer signal for action. Signal-detection and evidence-accumulation models similarly describe decisions as the crossing of a response boundary as evidence builds under uncertainty (16–18). This provides a useful framework for asking not only whether information is understood, but whether its presentation is associated with the point at which action occurs.

The present study examined whether six participant-facing information formats were associated with different intervention decisions in hypothetical sequential health-risk scenarios. Participants repeatedly selected either “Wait and monitor” or “Seek help now” as evidence accumulated across up to five stages. The formats were symptoms only, numerical urgency score, urgency label, ordinary visual bar, colored visual bar and a multimodal display. The urgency values were author-generated experimental quantities used to construct internally ordered escalation trajectories; they were not validated clinical scores and had no external clinical interpretation.

The primary question was whether information format was associated with stage-level intervention probability after accounting for current experimental urgency, escalation trajectory and scenario pair. Particular attention was given to the contrast between the colored bar and the otherwise similar non-colored visual bar, because this comparison more narrowly isolates the addition of the tested green-yellow-

orange-red mapping than comparisons with symptoms or numerical information alone. The study also examined whether the denser multimodal display differed from simpler formats. It did not test diagnostic accuracy, identify a clinically optimal intervention threshold, or determine whether earlier intervention would have been medically preferable. Its purpose was to test whether the interfaces studied were associated with shifts in observed behavioral thresholds.

2. Materials and Methods

2.1 Participants and study procedures

A total of 279 survey responses were submitted. Two submissions did not include consent and were excluded before preparation of the analytical dataset, leaving 277 consenting participants: 224 teenagers and 53 adults. Teenagers were recruited primarily through several independent and state schools in the United Kingdom; adults were mainly parents and teachers. The sample was a convenience sample rather than a nationally representative sample, and the adult subgroup was substantially smaller than the teenage subgroup.

Participants completed the electronic survey between May and June 2026. The task examined decisions under uncertainty in hypothetical scenarios; it did not assess medical knowledge, diagnostic accuracy, clinical triage performance, or actual care-seeking behavior.

Participation was voluntary and electronic consent was required before the survey proceeded. The public analytical file contains no names, email addresses, contact details, IP addresses, school identifiers, timestamps,

personal medical histories, or clinical records. Random participant identifiers were assigned after the two non-consenting submissions had been removed. The study involved hypothetical vignettes, no deception or medical intervention, and no collection of health records. Participant-facing materials and procedures were reviewed for school distribution by the designated safeguarding authority at the participating schools.

2.2 Sequential scenario design

The study used a sequential vignette design in which each scenario unfolded across a maximum of five displayed stages. At every stage, participants received updated symptom information and chose either “Wait and monitor” or “Seek help now”. The first “Seek help now” response ended that scenario, so later stages were not displayed and did not enter the

analytical risk set. Participants completed eight scenarios in the broader study: four in Part A and four in Part B.

This article reports Part B, which used the same underlying scenario structures while varying the information shown at the current decision screen. Each participant therefore contributed four Part B scenario sequences and encountered four of the six information formats. The allocation procedure is described in Section 2.6.

Figure 1 summarizes the sequential stopping rule. A participant who waited through Stage 5 therefore contributed five observed non-event decisions, and their intervention threshold was known only to exceed the final displayed stage. Only displayed stages entered the stage-level risk set.

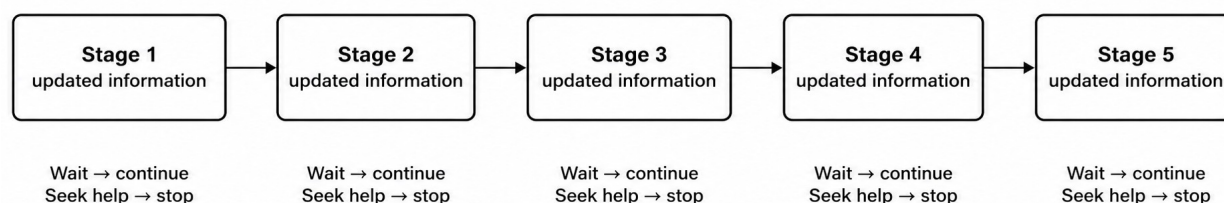


Figure 1. Sequential scenario structure and stopping rule.

2.3 Scenario content and experimental urgency scores

The scenarios were not presented as a single named disease or diagnostic category. Instead, they used varied combinations of common but potentially concerning symptoms, including fever, headache, fatigue, dizziness, vomiting, rapid heart rate, pale skin, blurred vision, sensitivity to light, shortness of breath, drowsiness, chills, and stiff neck. This design was intended to create escalating uncertainty

without making responses depend on familiarity with one named condition.

Before data collection, the author assigned each symptom an experimental weight ranging from 3 to 8. The general design approach was conceptually informed by established methods for structuring and grading health signals, including patient-reported symptom-item systems, adverse-event severity grading, early-warning scores, and clinical triage

classifications. PRO-CTCAE provides a standardized symptom-item library, CTCAE grades symptom-specific adverse-event severity, NEWS2 assigns and aggregates weighted physiological abnormalities to identify acute deterioration, and NICE uses an ordered green-amber-red traffic-light system for signs and symptoms in children with fever (23-26). These instruments were not directly applicable to the present task because they address cancer-treatment toxicities, adverse-event grading, objective physiological observations, or a defined paediatric fever population rather than a general additive urgency coordinate for lay-described symptoms in controlled vignette trajectories. Lower values were therefore generally assigned to relatively common or nonspecific cues, whereas higher values were

assigned to symptoms intended to represent more concerning warning signs within the experimental framework.

The stage score was the sum of the weights for symptoms present at that stage. The fixed weights were used solely to construct coherent, internally ordered and matched gradual and sudden trajectories. They were not estimated from clinical outcomes, copied from or calibrated against a validated instrument, or intended to represent disease probability, diagnostic severity, or an appropriate clinical triage threshold. Supplementary Tables S1 and S2 provide the complete symptom-weight mapping and scenario urgency trajectories; source files are retained in the OSF repository.

Table 1. Example matched gradual and sudden scenario pair.

Step	Gradual version symptoms	Gradual score	Sudden version symptoms	Sudden score
1	Moderate fever	3	Moderate fever	3
2	Fever + headache	8	Fever + headache + vomiting	11
3	Previous symptoms + vomiting	11	Previous symptoms + dizziness	14
4	Previous symptoms + dizziness	14	Previous symptoms + stiff neck	21
5	Previous symptoms + stiff neck	21	No new symptoms; previous symptoms	21

Table 1 illustrates one matched pair. The gradual version accumulated symptoms step by step, whereas the sudden version reached the same final experimental urgency through a more front-loaded path. The values are therefore meaningful only as internal coordinates of this experimental task.

2.4 Calibration and prior exposure

Before Part B, all participants received common instructions explaining the experimental score framework and examples of how increasing

values related to the study's ordering of concern. They also completed Part A, in which numerical scores and graphical bars were encountered before the Part B format comparison. Part B therefore tested the cue displayed on the current decision screen after shared calibration and prior exposure; it was not a comparison between completely naïve participants encountering an unexplained number and participants receiving an interpreted display.

The experimental categories were Low (0–4), Moderate (5–12), High (13–20), and Critical (21 or higher). In colored displays these categories were mapped to green, yellow, orange, and red, respectively. These boundaries were part of the experimental interface and must not be interpreted as externally validated clinical cut-offs.

2.5 Information-format implementation

Part B compared six formats while keeping the underlying symptom sequence fixed: symptoms only; symptoms plus a numerical urgency score;

symptoms plus an urgency label; symptoms plus an ordinary visual bar; symptoms plus a colored bar; and a multimodal display containing symptoms, score, label, and colored bar. Table 2 summarizes the participant-facing information and intended function of each format. In the numerical condition, the screen used wording such as “Urgency score: 14”. A denominator or universal maximum was not repeated on each Part B screen, so this condition offered less immediate interpretive context than the category-based displays despite the shared calibration described above.

Table 2. Information formats compared in Part B.

Format	Information shown	Communication channel	Intended function
Symptoms only	Symptom narrative only	Text baseline	Unaided symptom appraisal
Urgency score	Symptoms plus numerical urgency score	Text plus number	Quantitative severity cue
Urgency label	Symptoms plus verbal urgency category	Text plus label	Categorical interpretation of urgency
Visual bar	Symptoms plus visual urgency bar	Text plus visual scale	Magnitude cue without color coding
Colored bar	Symptoms plus color-coded urgency bar	Text plus color-coded visual scale	Salient visual categorization of urgency
Multimodal	Symptoms, score, label and colored bar	Text, number, label and color-coded visual scale	Dense combined decision-support display

The ordinary visual bar and colored bar used the same scale, category labels, position marker, and arrow. The tested colored-bar implementation added the green-yellow-orange-red category mapping; the multimodal display combined that bar with the numerical score and verbal label. Consequently, the direct ordinary-bar versus colored-bar contrast most narrowly represents the addition of this specific color mapping. Conventional learned traffic-light meanings may have contributed to the effect. The study does not isolate color from all other possible semantic associations and does not test

other palettes or bar designs. Supplementary Figure S1 compares the six participant-facing formats at the same underlying scenario stage. The complete original Part B stimuli are available in the OSF repository.

2.6 Format allocation and analytic dataset

Information-format allocation used a controlled fractional randomization scheme. Participants were randomly assigned to one of twelve predetermined survey versions, each containing four of the six Part B formats and one member of each matched gradual/sudden scenario pair.

Assignment was not based on participant characteristics.

Across the full allocation schedule, every format appeared with each of the four scenario pairs and with both gradual and sudden trajectories. Each participant received four Part B scenarios, and the resulting potential counts ranged from

182 to 187 observations per format. Table 3 summarizes allocation balance and clean scenario counts by information format. Because participants did not each complete all six formats, inference was based on adjusted stage-level models rather than unadjusted within-person comparisons across every display.

Table 3. Part B allocation balance and clean scenario counts.

Information format	Potential N	Clean N	Scenario pairs represented	Gradual/sudden potential N	Teenager/adult potential N
Symptoms only	186	185	1-4	92/94	151/35
Urgency score	186	184	1-4	94/92	148/38
Urgency label	183	177	1-4	88/95	151/32
Visual bar	187	181	1-4	98/89	152/35
Colored bar	182	178	1-4	89/93	146/36
Multimodal	184	178	1-4	93/91	148/36

Note. Potential N is the number assigned before exclusion; Clean N excludes ambiguous response sequences.

The sample derivation was fixed as follows: 277 consenting participants $\times$ four Part B scenarios = 1,108 potential scenario sequences. A strict response sequence was valid when all displayed responses before the first “Seek help now” were “Wait and monitor”, the intervention stage was “Seek help now”, and all later cells were empty because the scenario had stopped. A valid non-intervention sequence contained “Wait and monitor” at all five stages. Twenty-five sequences did not satisfy either rule and were flagged as ambiguous, leaving 1,083 primary clean scenarios. Expansion of those clean scenarios into displayed at-risk stages produced 3,047 stage-level rows. Supplementary Table S7 reports the 25 excluded patterns and reasons; the full audit file is retained in the OSF repository.

2.7 Outcomes and risk-set construction

The primary outcome was intervention at the current at-risk stage. It was coded 1 at the first stage where “Seek help now” was selected and 0 at earlier displayed stages. Valid non-interveners contributed five 0-coded rows. Scenario-level descriptive outcomes were observed intervention, intervention stage, and the experimental urgency score displayed at intervention.

Scenarios ending without intervention were administratively right-censored at Stage 5. Their observed non-event rows remained in the stage-level model, but they were not assigned an artificial scenario-level intervention score beyond the displayed sequence. Accordingly, descriptive means of intervention score included only scenarios with an observed intervention.

Figure 2 confidence intervals were generated by resampling participants, not by treating scenario observations as independent.

2.8 Primary statistical analysis

The primary analysis used a discrete-time logistic hazard representation. Each row was a displayed decision opportunity conditional on no earlier intervention in that scenario; rows after intervention did not exist in the risk set (19,20). A population-average logistic generalized estimating equation (GEE) model was fitted to the 3,047 stage-level rows.

The model predicted intervention at the current stage from continuous current experimental urgency, information format, escalation trajectory, and scenario pair. Symptoms only, gradual trajectory, and Pair 1 were the reference categories. Repeated observations were clustered by participant. The primary working-correlation structure was exchangeable and robust sandwich standard errors were used (21,22). The exchangeable structure was a parsimonious efficiency specification rather than a claim that all true within-participant correlations were identical; the estimated working correlation was 0.323. Analyses were performed in Python 3.13.2 using statsmodels 0.14.6, with the complete frozen environment archived in the repository. Complete coefficients are reported in Supplementary Table S4.

Seven prespecified contrasts formed one Holm-adjusted family: urgency score versus symptoms only; urgency label versus symptoms only; visual bar versus symptoms only; colored bar versus symptoms only; multimodal versus symptoms only; urgency score versus urgency

label; and visual bar versus colored bar. For each contrast, the log-odds estimate, robust standard error, odds ratio, 95% confidence interval, raw p-value, and Holm-adjusted p-value were calculated. The multimodal versus colored-bar comparison was reported separately as exploratory and was not inserted into the planned family. Supplementary Table S3 reproduces the complete planned output.

Model-predicted probabilities were calculated at selected urgency values and over a fine urgency grid, averaging equally across the four scenario pairs and two trajectories. For Figure 2, mean observed intervention scores were accompanied by percentile 95% confidence intervals from 20,000 participant-cluster bootstrap resamples using fixed seed 20260720.

2.9 Secondary and sensitivity analyses

A secondary model added current-urgency-by-format interactions. A joint Wald test assessed the five interaction terms, and individual coefficients—including the colored-bar-by-urgency term—were reported without replacing the prespecified main-effects model.

Five reviewer-driven sensitivity analyses were performed with participant clustering and robust standard errors. First, the primary formula was refitted with a working-independence correlation. Second, a score-independent model replaced numerical urgency with categorical stage and exact scenario version. Third, the 25 ambiguous sequences were conservatively reconstructed: missing responses before a first observed intervention were treated as prior waits, responses after that intervention were excluded from the risk set, and incomplete non-intervention sequences were completed as waits

through Stage 5. Fourth, an exploratory model additionally adjusted for age group, gender, serious-illness experience, and decision confidence. Fifth, a linearity diagnostic replaced the continuous score with the four experimental urgency categories. Supplementary Table S6 compares these colored-bar contrasts; complete files are retained in the OSF repository.

2.10 Analytical assumptions and scope

The analysis assumes that the displayed sequential decisions can be represented as an at-risk process ending at the first intervention or the fixed five-stage endpoint. The primary model treats the internally defined urgency score as linear on the log-odds scale and uses an exchangeable working correlation for repeated decisions; the categorical-urgency and independence analyses probe these specifications. The task measured stated intentions in hypothetical scenarios, not real clinical behavior, and the author-generated score has no external clinical meaning. Prior calibration and Part A exposure also form part

of the tested context. No assumption was made that censoring or ambiguous exclusions were unrelated to information format; the risk-set treatment and conservative ambiguity analysis address those issues directly.

3. Results

3.1 Sample derivation and descriptive intervention thresholds

The 277 consenting participants contributed four PartB scenarios each, yielding 1,108 potential scenario sequences. Twenty-five sequences were classified as ambiguous under the prespecified stopping-rule logic and excluded from the primary reconstruction. The primary scenario-level dataset therefore contained 1,083 clean sequences, which expanded to 3,047 at-risk stage rows across 277 participant clusters. The experimental urgency values used below are internal coordinates of the vignette task and have no external clinical interpretation.

Table 4. Observed intervention outcomes by information format.

Format	Clean N	Observed n	Rate (%)	Mean score	Score SD	Mean step	Step SD
Symptoms only	185	181	97.8	15.10	4.54	2.79	0.95
Urgency score	184	181	98.4	15.59	5.02	2.84	1.03
Urgency label	177	175	98.9	15.52	4.77	2.79	0.98
Visual bar	181	180	99.4	14.99	4.34	2.82	0.94
Colored bar	178	177	99.4	14.37	4.71	2.58	0.84
Multimodal	178	174	97.8	15.28	5.01	2.88	0.89

Note. Intervention rates use Clean N as the denominator. Mean score and mean step include only scenarios with an observed intervention.

Table 4 summarizes observed intervention rates were high in every format, ranging from outcomes by information format. Intervention 97.8% to 99.4%. Among scenarios with an

observed intervention, the tested colored-bar condition had the lowest mean experimental urgency score at intervention (14.37) and the earliest mean intervention step (2.58). The corresponding values were 15.10 and 2.79 for symptoms only, and 15.28 and 2.88 for the multimodal display.

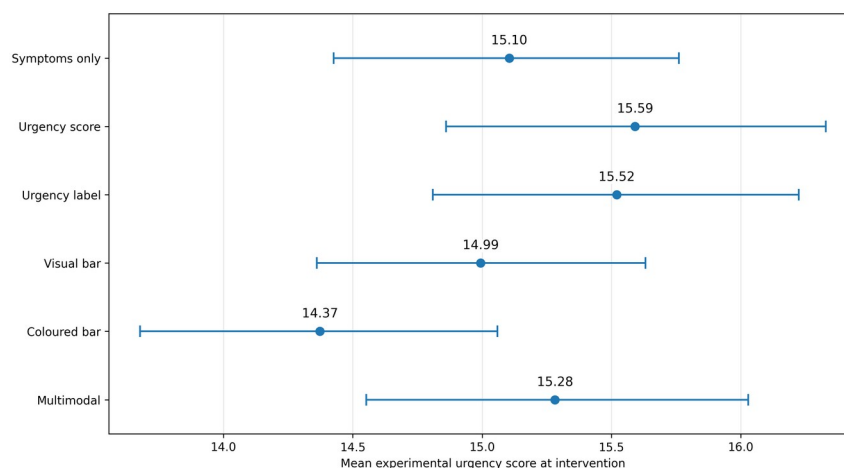


Figure 2. Mean experimental urgency score at observed intervention by information format. Points show means; horizontal bars show 95% participant-cluster bootstrap confidence intervals.

Figure 2 shows the mean experimental urgency score among scenarios with an observed intervention, together with participant-cluster bootstrap 95% confidence intervals. Scenarios without an intervention by Stage 5 had no observed intervention score and therefore did not contribute to these descriptive means, although their observed non-event stages remained in the primary stage-level analysis.

3.2 Primary stage-level model

The primary logistic generalized estimating equation included all 3,047 at-risk stage rows and clustered repeated decisions within 277 participants.

Table 5. Primary logistic GEE for stage-level intervention.

Predictor or format comparison	β	SE	Odds ratio	95% CI	p-value
Current urgency score	0.325	0.019	1.38	1.33–1.44	< .001
Urgency score vs symptoms only	0.058	0.167	1.06	0.76–1.47	.730
Urgency label vs symptoms only	0.239	0.161	1.27	0.93–1.74	.138
Visual bar vs symptoms only	0.193	0.137	1.21	0.93–1.59	.159
Colored bar vs symptoms only	0.723	0.157	2.06	1.52–2.80	< .001
Multimodal vs symptoms only	–0.009	0.166	0.99	0.72–1.37	.959

Note. Symptoms only, gradual trajectory, and Pair 1 were reference categories. The model used participant clustering, an exchangeable working correlation, and robust sandwich standard errors.

The exchangeable working-correlation estimate was 0.323. Current experimental urgency was strongly associated with intervention: each one-point increase corresponded to 1.38 times the odds of intervention, 95% CI [1.33, 1.44], $p < .001$. Information format was jointly associated with intervention probability, Wald $\chi^2(5) = 36.96$, $p < .001$.

After adjustment for current urgency, trajectory, and scenario pair, the tested colored bar was associated with higher intervention odds than

symptoms only, OR = 2.06, 95% CI [1.52, 2.80]. The numerical score, urgency label, ordinary visual bar, and multimodal display did not individually differ from symptoms only in the primary model (Table 5).

3.3 Model-predicted intervention probabilities

Table 6 reports model-derived probabilities at selected experimental urgency scores, averaged equally across the four scenario pairs and two trajectories.

Table 6. Predicted intervention probabilities at selected experimental urgency scores.

Format	Urgency 8 (%)	Urgency 12 (%)	Urgency 16 (%)	Urgency 20 (%)
Symptoms only	26.2	56.4	82.5	94.5
Urgency score	27.3	57.8	83.3	94.8
Urgency label	31.0	62.1	85.7	95.6
Visual bar	30.0	61.0	85.1	95.4
Colored bar	42.1	72.6	90.7	97.3
Multimodal	26.0	56.2	82.4	94.5

Note. Probabilities were averaged equally across four scenario pairs and two trajectories.

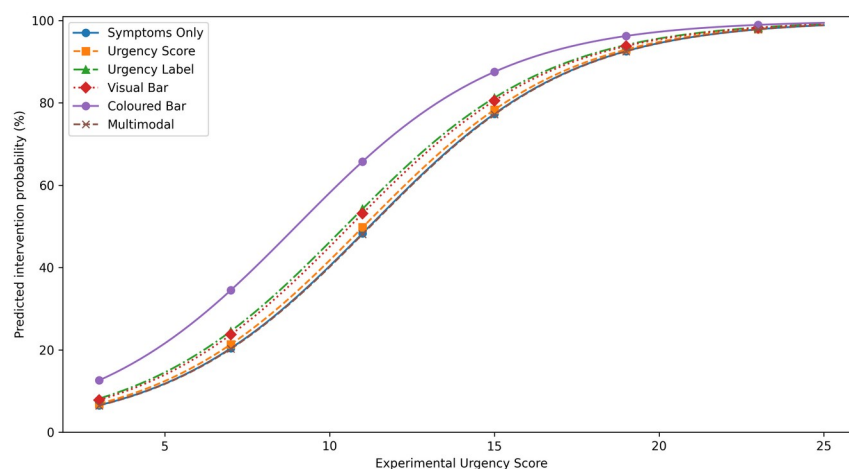


Figure 3. Predicted intervention probability by experimental urgency and information format.

At an experimental urgency score of 12, for symptoms only and 72.6% for the colored predicted intervention probability was 56.4% bar. At urgency 16, the corresponding

probabilities were 82.5% and 90.7%. The multimodal estimates closely tracked the symptoms-only baseline. These values describe the experimental scale and should not be interpreted as probabilities at clinically validated thresholds. Figure 3 displays the predicted probability curves across the experimental urgency range.

3.4 Planned Holm-adjusted contrasts

Table 7 presents the seven contrasts in the prespecified Holm-adjusted family.

Table 7. Prespecified Holm-adjusted contrast family.

Planned contrast	Log-odds	SE	OR	95% CI	Raw p	Holm p
Urgency score vs symptoms only	0.058	0.167	1.060	0.763–1.471	.729692	1.000000
Urgency label vs symptoms only	0.239	0.161	1.270	0.926–1.742	.137528	.687640
Visual bar vs symptoms only	0.193	0.137	1.213	0.927–1.587	.159305	.687640
Colored bar vs symptoms only	0.723	0.157	2.061	1.516–2.802	.000004	.000028
Multimodal vs symptoms only	−0.009	0.166	0.992	0.717–1.372	.958993	1.000000
Urgency score vs urgency label	−0.181	0.160	0.834	0.610–1.141	.256430	.769291
Visual bar vs colored bar	−0.530	0.151	0.589	0.438–0.791	.000449	.002695

Note. Contrasts are shown in their prespecified directions. For visual bar versus colored bar, $OR < 1$ indicates lower intervention odds under the visual bar; the reciprocal OR for colored bar versus visual bar is 1.70.

Two remained significant after adjustment: colored bar versus symptoms only (Holm-adjusted $p = .000028$) and ordinary visual bar versus colored bar (Holm-adjusted $p = .002695$).

Expressed in the more intuitive direction, the latter contrast corresponds to 1.70 times the odds of intervention under the colored bar relative to the otherwise similar ordinary visual bar, 95% CI [1.26, 2.28]. None of the other five planned contrasts was significant after Holm adjustment.

3.5 Interaction and sensitivity analyses

A secondary interaction model provided evidence that urgency slopes were not identical across all formats, joint format-by-urgency Wald $\chi^2(5) = 14.79$, $p = .011$. However, the colored-bar-by-urgency interaction itself was not significant, OR per one-point increase =

1.05, 95% CI [0.93, 1.18], $p = .454$. Thus, the interaction analysis did not show that the colored-bar association systematically increased or decreased across the urgency range, although it indicated some overall slope heterogeneity among the six formats. Supplementary Table S5 reports the complete coefficients and joint Wald test.

The direction of the colored-bar association persisted across all five sensitivity analyses. It was attenuated but remained positive under working independence; it remained present when the numerical urgency score was removed and stage plus exact scenario version were used instead; and it was similar under conservative inclusion of the 25 ambiguous sequences, exploratory demographic adjustment, and categorical urgency bands. Table 8 reports raw

p-values for these robustness analyses; they are not members of the primary Holm family.

Table 8. Sensitivity analyses for the colored-bar comparisons.

Analysis	Contrast	OR	95% CI	Raw p
Working independence	Colored bar vs symptoms only	1.55	1.10–2.18	.0112
Working independence	Colored bar vs visual bar	1.39	1.02–1.88	.0346
Score-independent model	Colored bar vs symptoms only	2.01	1.40–2.88	.000154
Score-independent model	Colored bar vs visual bar	1.76	1.25–2.49	.001310
Conservative ambiguous sequences	Colored bar vs symptoms only	1.99	1.47–2.70	.000008
Conservative ambiguous sequences	Colored bar vs visual bar	1.73	1.28–2.32	.000315
Demographic-adjusted model	Colored bar vs symptoms only	2.33	1.67–3.25	< .000001
Demographic-adjusted model	Colored bar vs visual bar	1.88	1.38–2.57	.000057
Categorical urgency diagnostic	Colored bar vs symptoms only	1.76	1.28–2.42	.000488
Categorical urgency diagnostic	Colored bar vs visual bar	1.60	1.18–2.15	.002271

Note. Raw p-values are unadjusted and are not part of the primary Holm family.

3.6 Exploratory multimodal comparison

The multimodal-versus-colored-bar comparison was not part of the planned Holm family and is reported as exploratory. The multimodal display was associated with lower intervention odds than the colored bar, OR = 0.48, 95% CI [0.35, 0.66], raw $p < .001$. This result does not establish that multimodal information is generally inferior; it concerns the specific dense display tested in this vignette task.

4. Discussion

4.1 Observed association

The primary finding was that the tested green-yellow-orange-red urgency bar was associated with intervention at an earlier point in the sequential vignette task. After adjustment for current experimental urgency, escalation trajectory and scenario pair, the colored bar was associated with higher stage-level intervention odds than symptoms alone, OR = 2.06, 95% CI

[1.52, 2.80]. The descriptive results pointed in the same direction: among scenarios with an observed intervention, the colored-bar condition had the lowest mean experimental urgency score at intervention. Together, these findings indicate that the way escalating information is organized and signaled can be associated with where participants place their threshold for action.

The other enhanced formats did not individually differ from symptoms alone in the primary coefficient table. The result therefore does not appear to reflect a general advantage of adding information or decision support. Instead, it identifies one specific visual implementation that was associated with a shift in intervention behavior within the experimental setting.

4.2 Evidence from the ordinary visual-bar comparison

The most informative color-related contrast was between the colored bar and the otherwise

similar ordinary visual bar. The two displays shared the same scale, category labels, position marker and arrow; the colored implementation added the green-yellow-orange-red mapping. In the prespecified contrast family, the colored bar was associated with 1.70 times the odds of intervention relative to the ordinary visual bar, 95% CI [1.26, 2.28], Holm-adjusted $p = .0027$. This comparison is more informative than the contrast with symptoms alone because it reduces differences unrelated to the tested color mapping. It suggests that adding a salient categorical layer to an existing graphical scale may affect when users decide that further monitoring is no longer sufficient. However, the colors also carried familiar traffic-light meanings, and participants had received common calibration and prior exposure to the score framework and graphical displays. The finding therefore concerns the complete colored-bar implementation, including its learned semantic associations, rather than an isolated perceptual effect of hue.

4.3 Possible mechanisms and the multimodal result

Several mechanisms could account for the observed association. The color mapping may have made category membership more salient, reduced the effort required to interpret position on the scale, or activated familiar associations between green, yellow, orange and red and increasing levels of concern. Because the bar also represented magnitude and progression, color may have reinforced an already visible trajectory and made the movement towards a higher-risk category easier to interpret.

The multimodal display contained symptoms, a numerical score, a verbal label and the colored

bar, yet it did not differ from symptoms alone in the planned comparison. The separate multimodal-versus-colored-bar contrast favored the colored bar, although that comparison was exploratory. This pattern suggests that adding several simultaneous cues does not necessarily strengthen the signal for action. Redundant or competing elements may have diluted the prominence of the colored threshold cue, or the denser layout may have encouraged further interpretation before a decision was made. These mechanisms were not measured directly, but they provide testable explanations for the difference between the simpler colored display and the multimodal format.

4.4 Implications and remaining uncertainty

The findings may also be relevant to the design of symptom checkers, health apps, wearable dashboards, monitoring systems and emerging AI-driven early warning systems. Such tools are often evaluated primarily by the accuracy of the risk estimate they produce. The present results suggest that the way this estimate is communicated may also affect when a user decides to act. A visually prioritized categorical cue may sometimes provide a clearer behavioral signal than a denser display containing several parallel representations.

This does not mean that interfaces should simply maximize intervention. The appropriate aim is to help users act at the right time. A display that increases intervention probability may reduce harmful delay in some settings but may also increase unnecessary alarm or service use in others. The tested experimental thresholds should therefore not be implemented as clinical guidance. Instead, the results support evaluating interface design alongside predictive

accuracy, particularly by comparing action timing, comprehension, false alarms and decision quality.

More broadly, the study demonstrates the value of treating intervention timing as an outcome in its own right. Risk communication is not only a question of whether information is noticed or understood; it is also a question of whether the presentation helps users decide when accumulating evidence has become sufficient to justify action.

4.5 Limitations

Several limitations constrain the interpretation and generalizability of these findings. First, participants formed a convenience sample rather than a nationally representative sample. Teenagers were recruited mainly through United Kingdom schools, adults were mainly parents and teachers, and the adult subgroup was substantially smaller than the teenage subgroup. The observed associations may therefore differ in other age groups, healthcare settings, countries, or populations with different levels of health and digital literacy.

Second, the scenarios were hypothetical and measured stated intentions rather than real care-seeking. Actual decisions may be shaped by emotion, time pressure, prior symptom experience, embarrassment, cost, access to care, family involvement and the consequences of false alarms. The vignette task provided a controlled test of intervention thresholds but not a direct estimate of behavior during a real medical event.

Third, the urgency scale was author-generated for the experiment. Its use of a structured

symptom library, graded values and an aggregate score drew on established measurement approaches, but the exact additive weights were assigned before data collection for stimulus construction rather than estimated from clinical outcomes. The score therefore functioned as an internal experimental coordinate and has no external diagnostic, prognostic or triage interpretation. Treating it as continuous in the primary model also assumed that a one-point increase had a common log-odds association across the experimental range. The categorical-urgency analysis reduced dependence on that assumption but did not convert the scale into a clinical measure.

Fourth, the six conditions did not provide identical amounts of immediate interpretive context. In particular, the numerical-score screen displayed wording such as “Urgency score: 14” without repeating a denominator or universal maximum, whereas the label and bar formats displayed category information directly. Comparisons therefore concern the complete participant-facing implementations tested rather than isolated communication channels under perfectly matched contextual conditions.

Fifth, Part B followed common score calibration and prior exposure in Part A to numerical scores and graphical bars. The results therefore describe how the cue shown on the Part B decision screen operated after shared familiarization. They may not generalize to a first encounter with an unexplained score, label or bar, or to users with different training and prior experience.

Sixth, the color evidence is specific to one green-yellow-orange-red palette and one bar

implementation. Learned traffic-light meanings may have contributed to the association, so the study cannot separate the perceptual effect of color from the categorical meanings attached to those colors. Other palettes, threshold boundaries, layouts and cultural conventions were not tested. Color-vision accessibility, legibility under different display conditions and the adequacy of non-color cues were also not assessed.

Seventh, each participant completed four Part B scenarios rather than all six information formats. The allocation schedule distributed formats across scenario pairs and trajectories, and the models adjusted for those factors, but the design was not a fully within-participant comparison of every display. The study also did not directly measure comprehension, attention, reading time, perceived urgency, cognitive load or trust, so the mechanisms proposed for the colored-bar and multimodal findings remain untested.

Eighth, scenarios were observed for no more than five stages. For participants who continued to wait through Stage 5, the exact intervention threshold remained unobserved and was known only to exceed the final displayed stage. These cases contributed their observed non-event decisions to the stage-level model but could not contribute a scenario-level intervention score to the descriptive mean. The fixed endpoint therefore limits precision about thresholds among non-interveners.

Finally, intervention at lower experimental urgency was not necessarily clinically preferable. Earlier action could reduce harmful delay in some settings, but it could also increase anxiety, false alarms, unnecessary consultations

or pressure on services. The study measured whether the tested interfaces shifted intervention probability and threshold position; it did not determine whether those shifts improved decision quality or matched an appropriate clinical standard.

4.6 Future research

Future studies should test risk-display formats in higher-fidelity digital environments, including interactive symptom checkers, health-app simulations, wearable dashboards and AI-driven early warning systems. These studies could examine whether visually prioritized threshold cues remain effective when users interact with more realistic interfaces and when system outputs change dynamically over time.

Further work should also test whether format effects vary by age, health literacy, prior illness experience, digital familiarity, color vision and trust in automated systems. Measuring comprehension, attention, perceived urgency, cognitive load and decision confidence would help distinguish between the possible mechanisms suggested by the present findings.

Finally, future research should combine threshold measurement with clinical or expert validation. The most useful display is not necessarily the one that produces the earliest action, but the one that improves the match between action timing and actual need.

5. Conclusion

Delayed help-seeking is often framed as a problem of symptom recognition, yet recognizing a warning sign does not necessarily indicate when action should be taken. This study instead examined intervention thresholds—the

point at which accumulating evidence becomes sufficient to justify seeking help—and showed that these thresholds can be associated with how risk information is presented rather than solely with the amount of information available.

Within a sequential health-risk vignette task, the tested green-yellow-orange-red urgency bar was associated with earlier intervention than both symptoms alone and an otherwise identical non-colored visual bar, whereas a denser multimodal display offered no apparent advantage. These findings suggest that interface design is not merely a vehicle for communicating risk but may itself influence the behavioral threshold at which people decide to act.

More broadly, the study highlights the importance of evaluating decision-support systems not only by the accuracy of their predictions but also by their ability to communicate escalating risk in ways that support appropriate action. For symptom checkers, clinical decision-support systems, wearable monitoring platforms, and emerging AI-assisted health applications, optimizing how

risk is presented may be as important as improving the underlying prediction algorithms.

These findings should be interpreted within the study's limitations. The urgency scale was an author-generated experimental construct, the scenarios were hypothetical, and the results do not establish clinical benefit or imply that earlier intervention is inherently preferable. Rather, they demonstrate that communication design can shift observed intervention thresholds. Future research should determine whether such shifts improve the alignment between action timing and genuine clinical need in real-world healthcare settings.

Acknowledgments and AI assistance

The author thanks Ellie Fassman for her guidance as a project mentor and for reviewing the results and overall structure of this paper. OpenAI ChatGPT was used solely for spelling, grammar, and general language editing during manuscript preparation. The author reviewed and approved all final text. All claims, interpretations, and editorial decisions are the author's responsibility.

Data availability statement

Article-specific reproducibility materials are available through a dedicated private OSF view-only link for peer review:

https://osf.io/3k8jt/overview_view_only=7d51a6100eef4644b35a945c2445abc4

The repository provides a README and an audited archive containing the de-identified response-level dataset, data dictionary and codebook, participant-facing instructions and stimuli, reconstruction scripts, statistical analysis scripts, figure-generation scripts, complete derived datasets, tables, figures, model and sensitivity-analysis outputs, and frozen software-environment information. The archive excludes the two non-consenting submissions and direct identifiers and is intended to permit independent reproduction of the revised article-specific analyses.

6. References

1. Safer, M. A., Tharps, Q. J., Jackson, T. C., & Leventhal, H. (1979). Determinants of three stages of delay in seeking care at a medical clinic. *Medical Care*, 17(1), 11–29.
2. Andersen, B. L., Cacioppo, J. T., & Roberts, D. C. (1995). Delay in seeking a cancer diagnosis: Delay stages and psychophysiological comparison processes. *British Journal of Social Psychology*, 34, 33–52.
3. Scott, S. E., Walter, F. M., Webster, A., Sutton, S., & Emery, J. (2013). The Model of Pathways to Treatment: Conceptualization and integration with existing theory. *British Journal of Health Psychology*, 18(1), 45–65.
4. Walter, F., Webster, A., Scott, S., & Emery, J. (2012). The Andersen Model of Total Patient Delay: A systematic review of its application in cancer diagnosis. *Journal of Health Services Research & Policy*, 17(2), 110–118.
5. Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self-diagnosis and triage: Audit study. *BMJ*, 351, h3480.
<https://doi.org/10.1136/bmj.h3480>
6. Wallace, W., Chan, C., Chidambaram, S., Hanna, L., Iqbal, F. M., Acharya, A., Normahani, P., Ashrafi, H., Markar, S. R., & Sounderajah, V. (2022). The diagnostic and triage accuracy of digital and online symptom checker tools: A systematic review. *npj Digital Medicine*, 5, 118.
<https://doi.org/10.1038/s41746-022-00689-w>
7. Riboli-Sasco, E., El-Osta, A., Alaa, A., Webber, I., Karki, M., El Asmar, M. L., Purohit, K., Painter, A., & Hayhoe, B. (2023). Triage and diagnostic accuracy of online symptom checkers: Systematic review. *Journal of Medical Internet Research*, 25, e43803.
<https://doi.org/10.2196/43803>
8. You, Y., Ma, R., & Gui, X. (2023). User experience of symptom checkers: A systematic review. *AMIA Annual Symposium Proceedings*, 2022, 1198–1207.
9. Brewer, N. T., Weinstein, N. D., Cuite, C. L., & Herrington, J. E. (2004). Risk perceptions and their relation to risk behaviour. *Annals of Behavioral Medicine*, 27(2), 125–130.
https://doi.org/10.1207/s15324796abm2702_7

10. Lipkus, I. M., & Hollands, J. G. (1999). The visual communication of risk. *Journal of the National Cancer Institute Monographs*, 1999(25), 149–163.
<https://doi.org/10.1093/oxfordjournals.jncimonographs.a024191>
11. Trevena, L. J., Zikmund-Fisher, B. J., Edwards, A., Gaissmaier, W., Galesic, M., Han, P. K. J., King, J., Lawson, M. L., Linder, S. K., Lipkus, I., Ozanne, E., Peters, E., Timmermans, D., & Woloshin, S. (2013). Presenting quantitative information about decision outcomes: A risk communication primer for patient decision aid developers. *BMC Medical Informatics and Decision Making*, 13(Suppl. 2), S7. <https://doi.org/10.1186/1472-6947-13-S2-S7>
12. Fagerlin, A., Zikmund-Fisher, B. J., & Ubel, P. A. (2011). Helping patients decide: Ten steps to better risk communication. *Journal of the National Cancer Institute*, 103(19), 1436–1443.
<https://doi.org/10.1093/jnci/djr318>
13. Jayawardena, T., Baysari, M., & Bamgboje-Ayodele, A. (2025). Interface design features of clinical decision support systems for real-time detection of deterioration: A scoping review. *International Journal of Medical Informatics*, 201, 105946.
<https://doi.org/10.1016/j.ijmedinf.2025.105946>
14. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
15. Mayer, R. E. (Ed.). (2005). *The Cambridge handbook of multimedia learning*. Cambridge University Press.
16. Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.
17. Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30, 535–574. <https://doi.org/10.1146/annurev.neuro.29.051605.113038>
18. Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873–922.
<https://doi.org/10.1162/neco.2008.12-06-420>
19. Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13, 61–98.
20. Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

21. Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22. <https://doi.org/10.1093/biomet/73.1.13>
22. Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2), 1–11. <https://doi.org/10.18637/jss.v015.i02>
23. Basch, E., Reeve, B. B., Mitchell, S. A., Clauser, S. B., Minasian, L. M., Dueck, A. C., Mendoza, T. R., Hay, J., Atkinson, T. M., Abernethy, A. P., Bruner, D. W., Cleeland, C. S., Sloan, J. A., Chilukuri, R., Baumgartner, P., Denicoff, A., St Germain, D., O'Mara, A. M., Chen, A., ... Trotti, A. (2014). Development of the National Cancer Institute's patient-reported outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). *Journal of the National Cancer Institute*, 106(9), dju244. <https://doi.org/10.1093/jnci/dju244>
24. National Cancer Institute. (2017). *Common Terminology Criteria for Adverse Events (CTCAE) (Version 5.0)*. U.S. Department of Health and Human Services. https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm
25. Royal College of Physicians. (2017). *National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS: Updated report of a working party*. Royal College of Physicians. <https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2>
26. National Institute for Health and Care Excellence. (2021). *Fever in under 5s: Assessment and initial management (NICE Guideline No. NG143)*. <https://www.nice.org.uk/guidance/ng143>