# Interactive Evaluation of an Adaptive-Questioning Symptom Checker Using Standardized Clinical Vignettes

Prathima Madda, Jagadeesh Kondru

Submitted to: JMIRx Med
on: September 02, 2025

# *Table of Contents*

# Interactive Evaluation of an Adaptive-Questioning Symptom Checker Using Standardized Clinical Vignettes

Prathima Madda[1] MBBS; Jagadeesh Kondru[1]

[1]7048 Nueces Dr Healtheja Inc Irving US

**Corresponding Author:**
Prathima Madda MBBS
7048 Nueces Dr
Healtheja Inc
7048 Nueces Dr
Irving
US

## *Abstract*

**Background:** Digital symptom checkers are widely used to guide care-seeking, yet evidence for triage safety in realistic, multi-turn use is limited. Prior evaluations often simulate single-pass inputs and rarely assess whether tools elicit the key clinical features needed for safe recommendations. Rigorous, interactive evaluations that quantify triage performance alongside the quality and burden of history-taking are needed.

**Objective:** To evaluate the triage performance and history-taking quality of an adaptive-questioning symptom checker (CareRoute) using an interactive protocol on standardized clinical vignettes (Semigran et al., BMJ 2015; 45 cases).

**Methods:** Each session began with only the presenting complaint; CareRoute asked follow-up questions adaptively, and the evaluator answered concisely per the vignette. At the end of questioning, CareRoute issued a triage recommendation. We compared CareRoute's issued triage with the reference triage and computed history-taking quality from normalized features derived from each vignette's Condensed Format. History-taking quality comprised (i) elicitation coverage—the percentage of a vignette's normalized features obtained through questioning, and (ii) elicitation fraction—the proportion of surfaced normalized features (elicited or volunteered) that were obtained through questioning. Primary outcomes were triage concordance and history-taking quality; the secondary outcome was user burden (time spent answering questions). We did not evaluate possible diagnoses, though CareRoute issues them.

**Results:** Exact 3-tier triage concordance was 88.9% (40/45; 95% CI 76.5–95.2%). Elicitation coverage had a median of 67% (IQR 60–71%), and elicitation fraction had a median of 70% (IQR 62–75%). CareRoute asked a median of 19 questions overall (IQR 16–20), with urgency-conditioned questioning: Emergency Care median 10 questions (IQR 4–14), Doctor Visit median 19 questions (IQR 18–20), Self Care median 19 questions (IQR 17–20).

**Conclusions:** In an interactive, vignette-constrained evaluation starting from only the presenting complaint, CareRoute achieved high 3-tier triage concordance (88.9%) with no under-triage on Emergency-reference vignettes, while eliciting most normalized features (median elicitation coverage 67%; median elicitation fraction 70%) with acceptable user burden via urgency-conditioned questioning.

(JMIR Preprints 02/09/2025:83429)
DOI: https://doi.org/10.2196/preprints.83429

## 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

# Interactive Evaluation of an Adaptive-Questioning Symptom Checker Using Standardized Clinical Vignettes

**Prathima Madda, Jagadeesh Kondru**

## Abstract

**Objective:** To evaluate the triage performance and history-taking quality of an adaptive-questioning symptom checker (CareRoute) using an interactive protocol on standardized clinical vignettes (Semigran et al., BMJ 2015; 45 cases).

**Methods:** Each session began with only the presenting complaint; CareRoute asked follow-up questions adaptively, and the evaluator answered concisely per the vignette. At the end of questioning, CareRoute issued a triage recommendation. We compared CareRoute's issued triage with the reference triage and computed history-taking quality from normalized features derived from each vignette's Condensed Format. History-taking quality comprised (i) elicitation coverage—the percentage of a vignette's normalized features obtained through questioning, and (ii) elicitation fraction—the proportion of surfaced normalized features (elicited or volunteered) that were obtained through questioning. Primary outcomes were triage concordance and history-taking quality; the secondary outcome was user burden (time spent answering questions). We did not evaluate possible diagnoses, though CareRoute issues them.

**Results:** Exact 3-tier triage concordance was 88.9% (40/45; 95% CI 76.5–95.2%). Elicitation coverage had a median of 67% (IQR 60–71%), and elicitation fraction had a median of 70% (IQR 62–75%). CareRoute asked a median of 19 questions overall (IQR 16–20), with urgency-conditioned questioning: Emergency Care median 10 questions (IQR 4–14), Doctor Visit median 19 questions (IQR 18–20), Self Care median 19 questions (IQR 17–20).

**Conclusions:** In an interactive, vignette-constrained evaluation starting from only the presenting complaint, CareRoute achieved high 3-tier triage concordance (88.9%) with no under-triage on Emergency-reference vignettes, while eliciting most normalized features (median elicitation coverage 67%; median elicitation fraction 70%) with acceptable user burden via urgency-conditioned questioning.

## Introduction

Symptom checkers are consumer-facing digital tools that collect brief medical histories and return preliminary triage advice (and often possible diagnoses). Their practical value hinges on two capabilities: (1) asking the

right follow-up questions to surface key findings, and (2) translating those findings into safe, actionable triage guidance. Modern AI-based, adaptive-questioning symptom checkers—such as CareRoute (evaluated in this study)—are designed to elicit missing information and issue a triage recommendation. Evaluations should therefore begin from a presenting complaint and measure both elicitation quality and triage accuracy.

However, most symptom checker evaluations to date use pre-extracted symptom lists entered once, leaving the quality of interactive questioning unmeasured[1–3]. To address this gap and evaluate both questioning quality and triage accuracy in a standardized way, we conducted an interactive evaluation using a well-established set of standardized clinical vignettes. Specifically, we used the 45 standardized clinical vignettes curated by Semigran et al. (BMJ, 2015)[1,4], which provide validated reference standards and have been widely used in the symptom checker evaluation literature. For brevity, we refer to this dataset as “Semigran-45.”

In Semigran-45, each vignette is labeled with one of three triage categories: *Requires Emergent Care*, *Requires Non-emergent Care*, or *Self Care appropriate*. Each vignette is also represented in a Condensed Format— a list of findings intended for symptom checker testing[1,4]. We used a normalized version of this list for scoring history-taking quality.

Our study had three objectives: (1) to evaluate CareRoute’s triage accuracy and safety using an interactive protocol that begins with only the presenting complaint, (2) to evaluate CareRoute’s ability to elicit key clinical features through adaptive-questioning, and (3) to establish a reproducible methodology for evaluating history-taking quality in symptom checkers. We pre-specified two primary outcomes: triage concordance (including safety assessment) and history-taking quality (via elicitation coverage and elicitation fraction metrics). We also report user burden (time spent answering questions) as a secondary outcome.

# Methods

We evaluated CareRoute (version 0.3.71) using an interactive, vignette-constrained protocol. Details of the protocol are described below.

## Dataset and triage mapping

We used all 45 vignettes from Semigran-45[1,4] (15 per category: *Requires Emergent Care*, *Requires Non-emergent Care*, *Self Care appropriate*). For consistency with CareRoute, we refer to the vignette triage categories as **Emergency Care**, **Doctor Visit**, and **Self Care**, respectively.

CareRoute provides four triage levels (Emergency Care, Urgent Care, Doctor Visit, Self Care), but our analysis uses a conservative 3-tier mapping that collapses Urgent Care to Doctor Visit.

## Condensed Format (official findings list)

For each vignette we started with the Condensed Format list, as published in the Semigran-45 materials, and derived a normalized features list by breaking down composite phrases into discrete features (e.g., “sudden onset severe abdominal pain” into “sudden onset”, “severe”, “abdominal pain”). We treated this normalized

list as the set of findings an ideal history should surface for correct triage. The Condensed Format and the normalized features were not shown to CareRoute; they were used only for computing elicitation coverage and elicitation fraction.

## Normalized features: example mapping

We normalized composite items from the Semigran-45 Condensed Format into discrete features. For example:

*Condensed Format of Vignette:*

> 12 y/o f, sudden onset severe abdominal pain, nausea, vomiting, diarrhea, T=104

*Normalized features:*

- 12 y/o f
- sudden onset
- severe
- abdominal pain
- nausea
- vomiting
- diarrhea
- fever 104°F

The normalization process enables granular assessment of history-taking quality via the elicitation coverage and elicitation fraction metrics.

## Interactive evaluation protocol

A single physician evaluator (P.M.) completed all 45 vignettes interactively. The evaluator was not presented with the vignette's reference triage tier, Condensed Format, or normalized features list during questioning. Each session started with only the presenting complaint. CareRoute asked questions adaptively. The evaluator answered to the point, consistent with the vignette. About half of the questions were selection-type; most were multi-select (except yes/no items, which were single-select). Where the vignette included clinician-generated information (e.g., a documented test), the evaluator could report it (e.g., "I was told the test was positive") when specifically asked by CareRoute. The session ended when CareRoute issued a triage recommendation. Each surfaced normalized feature was labeled *volunteered* if first mentioned before any targeted question about that feature (including in the presenting complaint), and *elicited* if first stated only in response to a targeted question.

## Outcomes and statistics

**Primary outcomes:**

- Triage concordance with reference values

- Elicitation coverage, $C = \frac{|Q \cap F|}{|F|} \times 100\%$
- Elicitation fraction, $E = \frac{|Q \cap S|}{|S|} \times 100\%$

Notation: For each vignette, $F$ is the set of normalized features derived from the Condensed Format; $V$ are features volunteered by the evaluator (including the presenting complaint); $Q$ are features elicited through questioning; $S = V \cup Q$.

**Meaning of the elicitation metrics.** *Elicitation coverage* asks: "Of the vignette's normalized features (the target findings an ideal history should surface), how many did CareRoute obtain by asking questions?" *Elicitation fraction* asks: "Of all surfaced normalized features (volunteered + elicited), what share came from CareRoute's questions rather than being volunteered?" Higher values on both mean CareRoute's questioning did more of the work.

**Secondary outcomes:**

- User burden (questions asked and estimated time)

**Reporting standards:**

Proportions are reported with Wilson 95% confidence intervals[5]. Medians are reported with interquartile ranges (IQRs).

# Results

In the following subsections, we discuss evaluation results in detail. Summarized evaluation results are available in Table A1 (Appendix).

## Triage concordance and safety

CareRoute exactly matched the reference in 40/45 cases (88.9%; 95% CI 76.5–95.2%). Under-triage occurred in 1/45 (2.2%) and over-triage in 4/45 (8.9%). On Emergency-reference vignettes, there was no under-triage (15/15; Wilson 95% CI 79.6–100.0%).

**Characterization of under-triage.** The single under-triage (reference: Doctor Visit; CareRoute: Self Care) was an "Influenza" vignette for a 30-year-old woman without risk factors. Current CDC guidance indicates otherwise-healthy adults with uncomplicated influenza can recover at home unless emergency warning signs arise or they belong to higher-risk groups; we therefore regard this as a defensible disagreement with the historical reference rather than a safety concern[6].

**Characterization of over-triage.** All four over-triage cases were reference *Self Care* vignettes that CareRoute escalated to *Doctor Visit* based on conservative safety cues appropriate for non-urgent outpatient review: infant constipation with intermittent blood-streaked stool and straining; pediatric eczema with sleep disruption and atopy; an upper respiratory illness with ~6 days of fever and purulent nasal discharge; and acute

bronchitis in an older adult with recent fever and a productive cough. These represent safety-favoring shifts to primary care rather than emergency referral.

**Benchmarking against prior evaluations.** Published evaluations report wide variability in triage accuracy for symptom checkers, typically spanning ~50–90% depending on study design, vignette set, and outcome definitions[2,3]. Within this landscape, CareRoute's 88.9% exact 3-tier concordance lies near the upper end of reported ranges. Reviews also note that higher-acuity presentations are often triaged more accurately than non-urgent ones, aligning with our finding of no under-triage on Emergency-reference vignettes[2].

## Confusion matrix (rows: Semigran-45 reference; columns: CareRoute results)

| Reference | Emergency Care | Doctor Visit | Self Care | Row total |
|---|---|---|---|---|
| Emergency Care (n=15) | **15** | 0 | 0 | 15 |
| Doctor Visit (n=15) | 0 | **14** | 1 | 15 |
| Self Care (n=15) | 0 | 4 | **11** | 15 |
| Column total | 15 | 18 | 12 | 45 |

## Elicitation coverage and fraction

Elicitation coverage achieved a median of 67% (IQR 60–71%) across all vignettes, with similar medians across tiers but greater variability for Emergency cases: Emergency 67% (IQR 38–71); Doctor Visit 67% (IQR 65–71); Self Care 67% (IQR 63–73).

Elicitation fraction was 70% (IQR 62–75%). Per-tier medians—Emergency 71.4% (IQR 55.0–75.0), Doctor Visit 66.7% (IQR 64.6–72.1), Self Care 70.0% (IQR 66.7–76.4)—tracked the elicitation coverage medians. Despite markedly fewer questions in Emergency cases, the elicitation fraction remained high, indicating that questioning—not volunteering—drove most surfaced normalized features.

## User burden

CareRoute demonstrated urgency-conditioned questioning, asking substantially fewer questions for high-acuity cases: Emergency (median 10; IQR 4–14) vs. Doctor Visit (median 19; IQR 18–20) and Self Care (median 19; IQR 17–20).

## Time estimates

Using measured 99th percentile app response times (5 seconds) and estimated user input time (10 seconds per question), median session duration was 4.8 minutes overall (IQR 4.0–5.0), with notably shorter sessions for emergency cases (2.5 minutes; IQR 1.0–3.5) compared to Doctor Visit and Self Care cases (~4.8 minutes

each).

## Case example: Kidney stones

**Vignette:** “A 45-year-old white man presents to the emergency department with a 1-hour history of sudden onset of left-sided flank pain radiating down toward his groin. The patient is writhing in pain, which is unrelieved by position. He also complains of nausea and vomiting.”

**Condensed Format:** “45 y/o m, 1 hour severe left-sided flank pain radiating into groin, nausea, vomiting, pain unrelieved by position”

**Normalized features (9 items):** - 45 y/o m - 1 hour - severe - left-sided - flank pain - radiating into groin - nausea - vomiting - pain unrelieved by position

**Interactive session:** - *Presenting complaint:* “flank pain” - *Questions asked:* 16 - *Features elicited:* 8/9 (89%) - *Triage issued:* Emergency Care (matches reference triage)

# Discussion

In this interactive evaluation, CareRoute achieved high triage concordance (88.9%) while minimizing user burden in high-acuity cases. Most normalized features were surfaced through questioning rather than volunteering, demonstrating that adaptive-questioning contributes meaningfully to safe triage.

## Comparison with existing evaluation methods

Prior symptom checker evaluations fall into four categories, none of which adequately measure history-taking quality:

- **Static-input:** Pre-extracted symptom lists entered once, scoring only triage/diagnosis[1,7]
- **Interactive-but-unquantified:** Symptom checkers ask questions but studies don’t measure elicitation effectiveness[8]
- **Patient-based studies:** Real-world use with clinical adjudication, but still not isolating elicitation quality[9]
- **Retrospective reviews:** Re-analysis of cases focusing on diagnostic accuracy, not questioning strategy[2,3,10]

Reviews note that vignette-only evaluations may inflate performance and fail to capture real interaction dynamics[2,11,12].

## Our methodological innovation

Our protocol uniquely measures three key dimensions:

1. Triage concordance/safety
2. History-taking quality: elicitation coverage and elicitation fraction
3. User burden (questions asked and time required)

This approach directly evaluates history-taking quality while demonstrating that urgency-conditioned questioning (fewer questions for emergencies) can maintain safety while improving efficiency.

## Limitations

This study has the following limitations: (1) the modest sample (45 vignettes) limits precision and may limit generalizability; (2) a single evaluator ensures consistency but does not reflect varied user communication styles; (3) the behavior of generative AI-based applications can vary across versions and runs, and we did not assess robustness.

# Conclusions

Our interactive evaluation protocol addresses a critical gap in symptom checker assessment by measuring both triage accuracy and history-taking quality. Using this approach, CareRoute achieved 88.9% triage concordance on the Semigran-45 benchmark with perfect safety on emergency cases, while demonstrating strong history-taking quality (median elicitation coverage 67%; elicitation fraction 70%) via adaptive-questioning.

CareRoute's urgency-conditioned approach—asking fewer questions for high-acuity cases—demonstrates that symptom checkers can balance thoroughness with efficiency. These findings support incorporating history-taking quality metrics into standard evaluation protocols for symptom checkers.

# Data and materials availability

Per-case evaluation data, session transcripts, and code for metrics computation are available at https://github.com/healtheja/carerouteai-data. Reviewers can reproduce the protocol using the free CareRoute app (https://careroute.ai/get), version 0.3.71 or later.


# Funding

None.


# Competing interests

Authors are co-founders of Healtheja Inc., developer of the CareRoute app.

# Ethics statements

Not applicable (vignette-based study; no human participants).

# Disclaimer

CareRoute is not a substitute for professional medical advice, diagnosis, or treatment. Users should seek professional medical advice for concerning symptoms.

# References


1. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: Audit study. BMJ. 2015;351:h3480.

2. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: A systematic review. NPJ Digital Medicine. 2022;5:118.
3. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and diagnostic accuracy of online symptom checkers: Systematic review. Journal of Medical Internet Research. 2023;25:e43803.
4. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: Supplementary materials. 2015.
5. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. The American Statistician. 2001;55(1):101–11.
6. Centers for Disease Control and Prevention. Treating flu. https://www.cdc.gov/flu/treatment/index.html; 2024.
7. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in australia. Medical Journal of Australia. 2020;212(11):514–9.
8. Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. BMJ Open. 2020;10(12):e040269.
9. Fraser HSF et al. Evaluation of diagnostic and triage accuracy and usability of a symptom checker in an emergency department: Observational study. JMIR mHealth and uHealth. 2022;10(9):e38364.
10. Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: Systematic review. BMJ Open. 2019;9(8):e027743.
11. El-Osta A, Webber I, Frame I, Bolus A, Jurgutis A, Orton C, et al. Evaluating the use of digital symptom checkers in primary care: A mixed-methods study. BMJ Open. 2022;12(3):e053645.
12. Kopka M, Schmieding ML, Feufel MA, et al. How suitable are clinical vignettes for the evaluation of symptom checker apps? A test theoretical perspective. Digital Health. 2023;9:20552076231194929.

# Appendix

## Table A1. Summary of Evaluation Results

| Metric | Value | 95% CI / IQR | Notes |
|---|---|---|---|
| **Primary outcomes** | | | |
| Triage concordance | 88.9% (40/45) | 76.5–95.2% | |
| • Under-triage | 2.2% (1/45) | — | Influenza case |
| • Over-triage | 8.9% (4/45) | — | 4 Self Care → Doctor Visit |
| • Safety on Emergency cases | 100% (15/15) | 79.6–100.0% | No under-triage |
| Elicitation coverage (median) | 67% | IQR 60–71% | Normalized features |
| • Emergency Care | 67% | IQR 38–71% | |
| • Doctor Visit | 67% | IQR 65–71% | |
| • Self Care | 67% | IQR 63–73% | |
| Elicitation fraction | 70% | IQR 62–75% | % of surfaced |

| Metric | Value | 95% CI / IQR | Notes |
|---|---|---|---|
| (median) | | | normalized features obtained through questioning |
| • Emergency Care | 71.4% | IQR 55.0–75.0% | |
| • Doctor Visit | 66.7% | IQR 64.6–72.1% | |
| • Self Care | 70.0% | IQR 66.7–76.4% | |
| **Secondary outcomes** | | | |
| Questions asked (median) | 19 | IQR 16–20 | User burden (questions) |
| • Emergency Care | 10 | IQR 4–14 | Urgency-conditioned |
| • Doctor Visit | 19 | IQR 18–20 | |
| • Self Care | 19 | IQR 17–20 | |
| Session time estimate | 4.8 min | IQR 4.0–5.0 | 15s per turn |
| • Emergency Care | 2.5 min | IQR 1.0–3.5 | |
| • Doctor Visit | 4.8 min | IQR 4.5–5.0 | |
| • Self Care | 4.8 min | IQR 4.3–5.0 | |