NEJM AI

DOI: 10.1056/AIoa2500833

ORIGINAL ARTICLE

From Advice to Action — Real-World Behavior of Patients Using an Integrated Diagnostic Decision Support System for Navigating the Health Care System

Fabienne Cotte, M.D.,[1] Filipa Dias Lourenço, M.D.,[2] Miguel Paiva Pereira, M.D.,[2] Nisha Kini, M.B.B.S., M.P.H.,[1] Marcel Schmude, M.D.S.,[1] Andreia Pimenta, Ph.D.,[1] Athena Lemesiou, Ph.D.,[1] Stephen Gilbert, Ph.D.,[3,4] Tauseef Mehrali, M.B.Ch.B. (Hons.),[1] Micaela Seemann Monteiro, M.D.,[2] and Pedro Flores, Ph.D.[2]

Received: July 24, 2025; Revised: December 9, 2025; Accepted: January 20, 2026; Published: March 5, 2026

Abstract

BACKGROUND Artificial intelligence (AI)–powered digital front door tools are increasingly being used to guide patients to appropriate care and alleviate health care system pressure. However, most evaluations offer limited insight into stated care intent, real-world behavior, or care appropriateness.

METHODS The E-Health Self–Symptom Assessment as a Front Door and Facilitator of Care (ESSENCE) study was a prospective real-world quality improvement evaluation embedded in Portugal's largest private health care network (CUF). Adults using the Ada Health diagnostic decision support system through the myCUF app reported their care intent before and after symptom assessment. We tracked actual behavior through electronic health records and surveys. Physician panels retrospectively assessed the appropriateness of intended and observed care.

RESULTS A total of 1470 adults (≥18 years of age; mean age, 38.5 years; 57.7% female) were enrolled. Of the 1338 participants with pre- and postassessment intentions, 33.0% revised their planned care level immediately after assessment. Uncertainty decreased from 12.6% to 5.0% ($P<0.001$). Among 721 participants with observed behavior, 59.1% changed their care pathway: 28.9% de-escalated, 17.2% escalated, and 13.0% resolved prior uncertainty. Primary care consultations increased from 16.3% to 42.1% ($P<0.001$), whereas specialist visits decreased from 49.7% to 29.8% ($P<0.001$). Among nonemergency participants with preassessment intentions and sufficient clinical documentation (n=382), appropriate care increased from 29.8% preassessment to 64.4% postbehavior (95% confidence interval, 27.8 to 41.4; $P<0.001$). Of the 96 participants who planned an emergency department (ED) visit, 38.5% selected lower-acuity care after assessment. In the subset with clinician-rated follow-up, 93% (27 of 29; 95% CI, 78.0 to 98.1%) were judged to have appropriately avoided an unnecessary ED visit.

CONCLUSIONS Integrating an AI-supported symptom assessment and follow-up service options within a digital front door was associated with shifts in patient intentions and

Fabienne Cotte and Filipa Dias Lourenço contributed equally to this article.

The author affiliations are listed at the end of the article.

Pedro Flores can be contacted at pedro.flores@sapo.pt.

behaviors, reducing uncertainty and promoting appropriate health care use. These findings suggest that diagnostic decision support systems shape real-world decision-making in addition to generating accurate recommendations, warranting further evaluation across diverse health care settings. (Funded by the Federal Ministry of Research, Technology and Space [Bundesministerium für Forschung, Technologie und Raumfahrt].)

Introduction

Health care systems worldwide face growing strain from rising chronic disease burdens, aging populations, and health care professional shortages, which stretch resources, increase wait times, and limit access to timely care.[1] Scalable solutions are needed that guide patients to the right care at the right time — without compromising quality or safety.

One such solution is the digital front door (DFD): digitally enabled entry points through which patients access health care services. These tools — including symptom assessments, online triage, appointment booking, and remote consultations — are designed to streamline access, reduce unnecessary demand, and guide users to the most appropriate care setting.[2] Critically, they hold potential to support patients in one of the most challenging aspects of health care: deciding whether, when, where, and how urgently to seek care.

Evidence shows that this decision-making process is filled with uncertainty. Self-triage accuracy remains moderate (47%–62%), with frequent overestimation of urgency and occasional failure to recognize serious conditions.[3,4] Although most individuals correctly identify emergencies (82%), overtriage occurs in 65% of cases, and 8% of critical situations are missed entirely.[5] As highlighted during the coronavirus disease 2019 pandemic, delayed or misdirected care can lead to worse health outcomes.[6] In response, health care systems are increasingly embedding artificial intelligence (AI)–based symptom assessment and navigation into patient-facing platforms. These integrations aim to reduce uncertainty, curb unnecessary demand, and improve timely access to care.

However, most evaluations of digital triage tools focus narrowly on output accuracy or stated intent, offering limited insight into actual patient behavior. For example, a study in Portugal reported that 22.8% of users revised their care plans after AI-supported triage, but did not assess whether patients followed through with their intention or whether the resulting care was appropriate.[7] Similarly, an evaluation in Australia described reductions in overtriage compared with the previous system but did not examine adherence to recommendations or downstream outcomes.[8] These approaches miss the complex, often nonlinear ways patients interpret information and make decisions in real-world settings. Users are often exposed to more than a single triage recommendation. They receive a mix of condition-specific advice, explanatory information, and follow-up options — all of which, alongside additional contextual factors, influence their decisions. Understanding this interplay requires evaluation that tracks actual behavior and clinical appropriateness.

To address this gap, we conducted the E-Health Self-Symptom Assessment as a Front Door and Facilitator of Care (ESSENCE) study. Embedded within Portugal's CUF private health care network, the study assessed how integrating Ada Health's diagnostic decision support system (DDSS) into a DFD influenced patient behavior, service use, and care appropriateness in a real-world setting. This article focuses on the behavioral impact of the Ada Health–CUF integration; findings related to clinical workflow and physician perspectives will be reported separately.

Methods

STUDY DESIGN

This prospective, observational quality improvement study evaluated the real-world impact of the Ada Health DDSS on health care–seeking behavior within the CUF private health care network. Adults (≥18 years of age) using the myCUF app in Portuguese to complete a full symptom assessment for themselves were eligible. Inclusion required completion of pre- and postassessment questions, informed consent, and report sharing with a health care provider. Participants were instructed not to use the DDSS in emergency situations. Enrollment ran from November 2023 to October 2024.

DATA COLLECTION

Participants reported their care-seeking intention before the assessment and again after reviewing the symptom report, which listed likely conditions with probability ratings and care recommendations for each condition and the overall case. Each condition included a "What's next?"

button linking to follow-up service options (e.g., booking a teleconsultation), which was only shown after the postassessment intention question was completed.

Health care–seeking behavior was collected in two steps. First, all CUF encounters occurring within 14 days of completion of the symptom assessment were retrieved from the electronic health record (EHR) and classified as observed behavior. Second, for participants without a documented CUF encounter in this 14-day window, an email survey was sent between day 14 and day 28 to capture self-reported care use outside CUF or home management. For the primary analyses, behavior was considered "known" if confirmed either through the 14-day EHR window or through the survey response, using the same care categories as in the intention questions.

For CUF consultations, additional data were extracted (e.g., clinical notes, *International Statistical Classification of Diseases and Related Health Problems*, Tenth Revision codes). If the participant was seen by a physician participating in the study at one of three participating CUF centers, a survey evaluating the symptom report was completed directly after the consultation or later if necessary.

At study end, a panel of three senior CUF physicians who were not involved in treating patients assessed in the study reviewed each visit to determine the most appropriate level of care (i.e., self-care, primary care, specialist care, or emergency care). Raters were blinded to the DDSS recommendation and to participants' pre- and postassessment intentions and did not have access to downstream outcomes (e.g., diagnostics, treatments, revisits) when making their appropriateness ratings. Emergency department (ED) visit appropriateness was assessed separately from non-ED visits using a three-part framework evaluating clinical appropriateness, potential for primary care evaluation, and type of resource use in the ED.[9] Visits were classified as appropriate if they were either clinically justified or required ED-specific interventions.

To assess potential safety signals, we additionally reviewed downstream acute care utilization for users who de-escalated from an initial emergency care intention and for those who directly attended the ED, including repeat ED visits within 72 hours and any hospitalizations recorded in CUF's EHR within 14 days.

DATA ANALYSIS

All statistical analyses were performed using R software (version 4.2.1). Descriptive statistics were used to summarize participant demographics, care intentions, observed behaviors, and DDSS recommendations. Continuous variables are presented as means (± standard deviation) or medians (interquartile range), and categorical variables as frequencies and percentages. Care transitions across the time points (i.e., preassessment, postassessment, observed behavior) were visualized using Sankey diagrams.

To evaluate whether the overall distribution of care levels changed across time points, paired comparisons were analyzed using the Stuart–Maxwell test of marginal homogeneity. Agreement between care selections at different time points was assessed using Cohen's kappa with 95% confidence intervals and interpreted according to conventional strength categories.

To further describe direction-specific movement between categories, McNemar's test was applied to paired proportions. These tests were used descriptively to illustrate specific category shifts. The associations between DDSS advice and observed behavior were analyzed using Pearson's chi-square test. Rankers' agreement was studied by Fleiss' kappa test.

A multivariable logistic regression was conducted to examine predictors of appropriate real-world behavior. The dependent variable was actual behavior appropriateness (appropriate vs. inappropriate). Independent variables were selected a priori based on theoretical relevance: preassessment appropriateness (Y [yes]/N [no]/unsure), DDSS case-advice appropriateness (Y/N), postassessment appropriateness (Y/N/unsure), age (continuous), and sex (female/male). Categorical variables were dummy coded. Variables such as day of the week, number of symptoms, and site were not included due to a high proportion of missing data or limited clinical relevance. Maximum likelihood estimation was used with an alpha equal to 0.05. The Hosmer–Lemeshow test was applied to assess model fit ($P>0.05$ indicating adequate fit).

A significance level of 5% was adopted for all statistical analyses.

GENERALIZABILITY

To assess selection effects, we compared care intentions and user characteristics between the study population and all other DDSS users who completed an assessment within the same integration (excluding skipped responses). Within the study population, we examined differences between participants with and without follow-up data, including demographics, DDSS-recommended care level, and changes in intended care.

DATA MANAGEMENT AND ETHICS

Data were securely stored in a validated electronic system (Teamscope) using unique identifiers. Ethical approval was granted by the National Ethics Committee for Clinical Research (Comissão de Ética para a Investigação Clínica) (No. 2204JJ351 and No. 2309JJ660). The study was registered at ClinicalTrials.gov (as NCT06846957) and complied with the Declaration of Helsinki and International Organization for Standardization (ISO) 14155:2020 guidelines. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines were followed throughout.

Results

PARTICIPANT DEMOGRAPHICS

The study enrolled 1470 participants, with a mean age of 38.5 years (±12.5); 57.7% (848 of 1470) were female. Care-seeking behavior was determined for 721 participants, of whom 529 (73%) were validated through EHRs and 192 (27%) through follow-up survey responses. Full demographic, symptom, and DDSS recommendation distributions are shown in Table 1. The participant selection and inclusion process is outlined in a STROBE flowchart (Fig. 1).

Table 1. Participant Demographics (N=1470).*

Continuous Variable	Mean±SD or (%)
Age (years)	38.5±12.5 (range, 18–83)
Number of symptoms entered initially	2.3±1.9 (IQR, 1–3)
Sex distribution	
Female	848 (57.7)
Male	615 (41.8)
Missing data	7 (0.5)
DDSS recommendation	
Self-care	24 (1.6)
Self-care/pharmacy	154 (10.5)
Primary care — 2–3 weeks	146 (9.9)
Primary care — 2–3 days	326 (22.2)
Primary care — same day	547 (37.2)
Primary care — 4 hours	80 (5.4)
Emergency care	144 (9.8)
Call ambulance	42 (2.9)
Missing data	7 (0.5)
Site visited	
Virtual health care visits	211 (14.3)
In-person health care visits	460 (31.3)
Managed symptoms at home	50 (3.4)
No information available	749 (51.0)

*DDSS denotes diagnostic decision support system; IQR, interquartile range; and SD, standard deviation.

PARTICIPANT INTENTIONS (PRE- VS. POSTASSESSMENT)

Participant intentions before and after the assessment are summarized in Tables 2–4 and Table S1 in the Supplementary Appendix, and are visualized in Figure 2. In the paired cohort (n=1338), uncertainty declined substantially from 12.6% to 5.0% (McNemar χ^2=57.14; P<0.0001), corresponding to a −7.54 percentage point change (95% confidence interval [CI], −9.52 to −5.56). Intentions to consult a specialist increased from 53.0% to 57.3% (+4.33 percentage points; 95% CI, 2.66 to 6.00; P=0.0003), and intentions to consult a primary care provider (PCP) increased from 17.8% to 19.7% (+1.87 percentage points; 95% CI, 0.42 to 3.32; P=0.082). The proportion of patients intending to seek emergency care (9.4% to 10.8%; +1.35 percentage points; 95% CI, 0.13 to 2.57; P=0.07) or to manage symptoms at home (7.2% to 7.2%; 0.00 percentage points; 95% CI, −1.39 to 1.39; P=1.00) showed no meaningful change. Overall, only 67.0% (897 of 1338) of participants maintained their original intention.

PARTICIPANTS' BEHAVIOR COMPARED WITH PREASSESSMENT INTENTIONS

The overall distribution of care level shifted significantly from preassessment intentions to observed behavior (P<0.001), driven by a marked reduction in specialist care and increased PCP consultations. Among 717 participants for whom both preassessment intentions and observed behavior were available, only 41% (294 of 717) followed through with the care level they had initially planned. Among participants, 17.2% (123 of 717) escalated to a higher-urgency option, whereas 28.9% (207 of 716) de-escalated to a lower-urgency one, indicating a net shift toward lower-acuity care (Tables 3 and 4). Primary care use increased from 16.3% (117 of 717) to 42.1% (302 of 717) (95% CI, 22.5 to 29.1; P<0.0001), while specialist visits decreased from 49.7% (356 of 717) to 29.8% (214 of 717) (95% CI, −23.0 to −16.6; P<0.0001). Self-management remained stable (7.7% to 6.8%; 95% CI, −3.1 to 1.5; P=0.54) and emergency care use increased from 13.4% (96 of 717) to 21.2% (152 of 717) (95% CI, 5.2 to 10.4; P<0.0001; Fig. 3).

Agreement between care choices across time points was moderate for pre- and postassessment intentions (κ=0.52; 95% CI, 0.48 to 0.56) and fair for both preassessment and behavior (κ=0.21; 95% CI, 0.16 to 0.26)

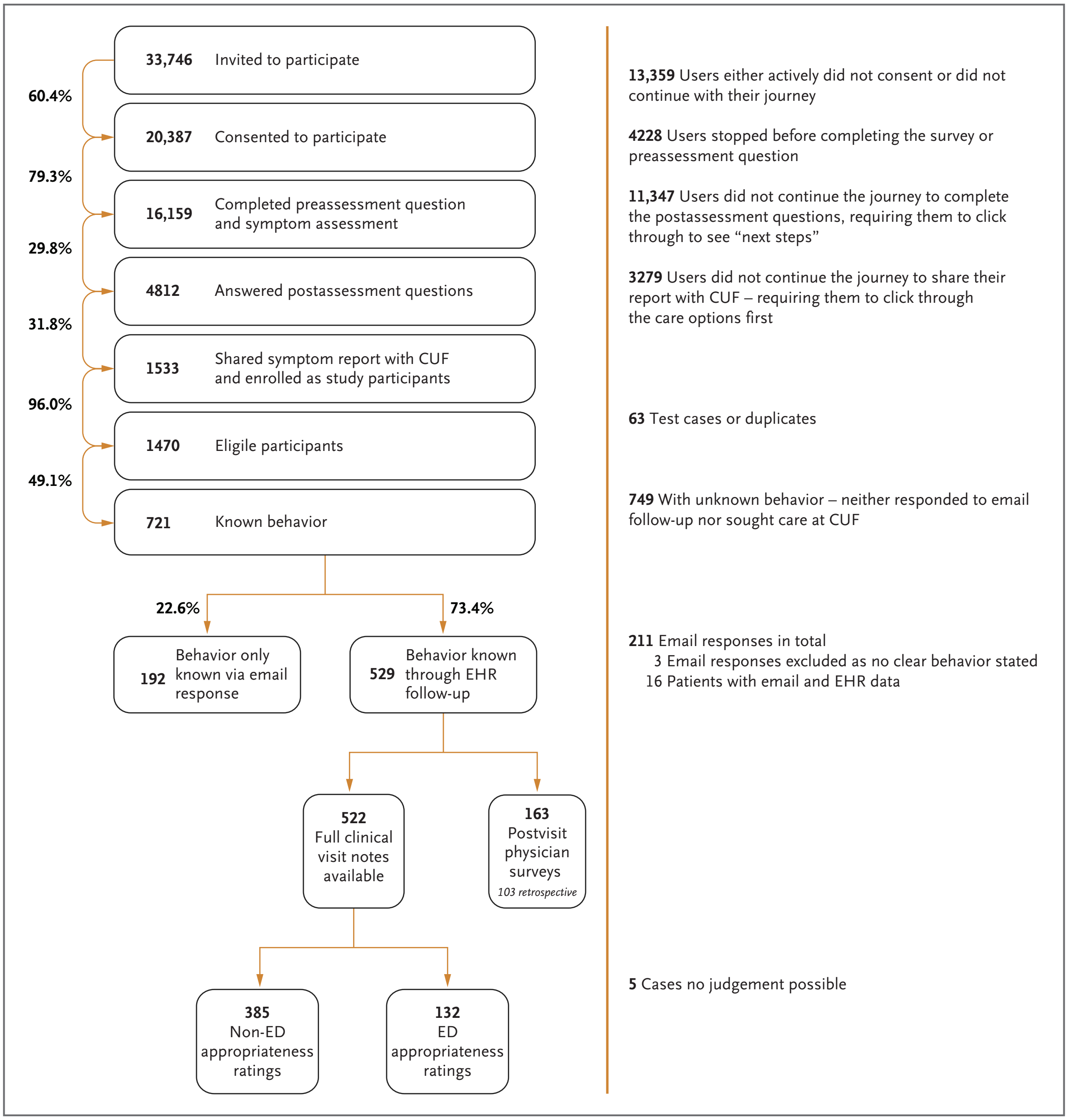

Figure 1. Strengthening the Reporting of Observational Studies in Epidemiology Inclusion Flowchart.

ED denotes emergency department; and EHR, electronic health record.

Table 2. Changes in Care Intentions before and after the Symptom Assessment (Paired Cohort, N=1338).*

Care Category	Preassessment, N (%)	Postassessment, N (%)	Net Change (pp)	95% CI	McNemar χ^2 (CC)	P Value
Consulting a specialist	709 (53.0)	767 (57.3)	4.33	2.66 to 6.00	12.996	0.0003
Consulting a primary care physician	238 (17.8)	263 (19.7)	1.87	0.42 to 3.32	3.015	0.082
I’m not sure	168 (12.6)	67 (5.0)	−7.54	−9.52 to −5.56	57.14	<0.0001
Seeking emergency care	126 (9.4)	144 (10.8)	1.35	0.13 to 2.57	3.28	0.07
Managing at home/consulting pharmacist	97 (7.2)	97 (7.2)	0.00	−1.39 to 1.39	0	1

*A total of 132 users were excluded due to missing intention data, either skipped before or after the symptom assessment, or both. CC denotes correlation coefficient; CI, confidence interval; and pp, percentage points.

Table 3. Changes in Care Intentions before the Symptom Assessment and Subsequent Actual Behavior (Paired Cohort, N=717).*

Care Category	Preassessment, N (%)	Behavior, N (%)	Net Change (pp)	95% CI (pp)	McNemar χ^2 (CC)	P Value
Consulting a PCP	117 (16.3)	302 (42.1)	25.8	22.5 to 29.1	136.1	<0.0001
Consulting a specialist	356 (49.7)	214 (29.9)	−19.8	−23.0 to −16.6	84	<0.0001
Managing at home/consulting pharmacist	55 (7.7)	49 (6.8)	−0.8	−3.1 to 1.5	0.378	0.539
Seeking emergency care	96 (13.4)	152 (21.2)	7.8	5.2 to 10.4	24.27	<0.0001
I'm not sure	93 (13.0)	0 (0.0)	−8.8	−11.3 to −6.3	61	<0.0001

*Behavioral data are only known for 721 participants, of whom 4 initially had skipped the preassessment intent question. CC denotes correlation coefficient; CI, confidence interval; PCP, primary care provider; and pp, percentage points.

Table 4. Transition Outcomes for Pre- versus Postassessment Intentions and Observed Behavior.*

Metric	Pre→Post (N=1338)	Pre→Behavior (N=717)	Post→Behavior (N=659)
No change	897 (67.0%)	294 (41.0%)	299 (45.4%)
Escalated†	126 (9.4%)	123 (17.2%)	122 (18.5%)
De-escalated†	110 (8.2%)	207 (28.9%)	209 (31.7%)
Made certain‡	138 (10.3%)	93 (13.0%)	29 (4.4%)
Made uncertain‡	37 (2.8%)	0 (0.0%)	0 (0.0%)
Stayed uncertain§	30 (2.2%)	0 (0.0%)	0 (0.0%)
Stuart–Maxwell test of marginal homogeneity	Q=48.20, df=4, P<0.001	Q=124.4, df=6, P<0.001	Q=90.39, df=3 P<0.001
Cohen's kappa	Cohen's κ=0.52 (95% CI, 0.48 to 0.56), moderate agreement	Cohen's κ=0.21 (95% CI, 0.16 to 0.26), fair agreement	Cohen's κ=0.24 (95% CI, 0.18 to 0.29), fair agreement

* CI denotes confidence interval; and df, degrees of freedom.
† Escalated indicates a move to a higher-urgency category; de-escalated indicates a move to a lower-urgency category.
‡ Made certain indicates a shift from "I'm not sure" to a specific category; made uncertain indicates a shift from a specific category to "I'm not sure."
§ Stayed uncertain indicates "I'm not sure" at both time points.

and postassessment and behavior (κ=0.23; 95% CI, 0.18 to 0.29). Preassessment intention was the only statistically significant factor driving whether users stuck with or revised their care plans (P<0.001) (Table S5). The level of the DDSS's advice was associated with behavior (χ^2=25.44; degrees of freedom=4; P<0.001).

APPROPRIATENESS OF PLANS FOR SEEKING CARE AND OBSERVED BEHAVIOR

Appropriateness ratings before and after the symptom assessment were available for 363 participants with complete paired data (Table 5). The proportion of appropriate planned care increased from 30.0% to 35.3%, a statistically significant improvement of +5.23 percentage points (95% CI, 0.88 to 9.59; P=0.026). The proportion of inappropriate plans remained stable (59.2% vs. 59.5%, +0.28 percentage points; 95% CI, −4.46 to 5.01; P=1.00), while the proportion of participants uncertain about how to manage their symptoms decreased markedly, from 10.7% to 5.2% (−5.51 percentage points; 95% CI, −9.01 to −2.01; P=0.0034).

The appropriateness of intended care versus actual care taken was compared in 382 participants with complete paired observations (Table 6). The proportion of users whose behavior was rated clinically appropriate was substantially higher than the proportion whose initial plans were appropriate: 29.8% at preassessment versus 64.4% at the point of care (+34.55 percentage points; 95% CI, 27.75 to 41.36; P<0.0001). Conversely, inappropriate intentions decreased from 59.2% to 35.6% (−23.56 percentage points; 95% CI, −30.05 to −17.07; P<0.0001). Among initially unsure participants, 69% (29 of 42; 95% CI, 54.6 to 80.8) sought appropriate care. Overall, 77.0% (296 of 385; 95% CI, 72.5 to 81.0) of cases received the most appropriate urgency advice. Appropriateness could not be evaluated for 50 users (6.9% of 721 users with known behavior) who managed their symptoms at home, as no clinical documentation was available for review.

Interrater agreement was high: The three raters agreed in 344 of 385 cases (89.4%), with disagreement occurring in 41 cases. In 13 cases, rater A differed from raters B and C; in 12 cases, rater B differed; in 11 cases, rater C differed; and in 5 cases, all three raters assigned different classifications. Fleiss' kappa was approximately 0.85, indicating substantial reliability. All disagreements were resolved through consensus after joint review of the clinical files.

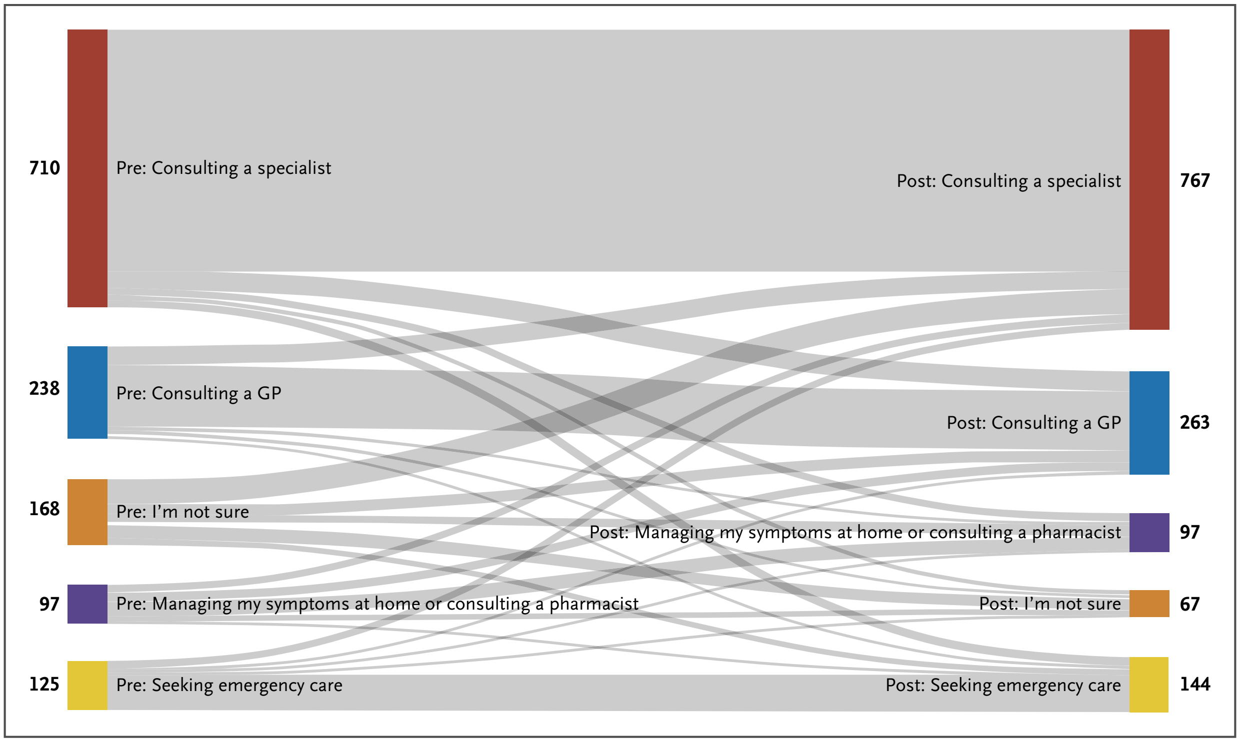

Figure 2. Transitions in Care Intentions from Preassessment to Postassessment Time Points.

This figure presents a Sankey diagram illustrating transitions between merged care categories before and after the symptom assessment. Node sizes represent the number of users in each category; link widths indicate the number transitioning between categories. Categories include consulting a specialist, consulting a primary care provider, managing symptoms at home or consulting a pharmacist, "I'm not sure," and seeking emergency care. Participants with missing responses at either time point were excluded, yielding a paired dataset of 1338. For each category, the number of users who remained in the same category, moved out, or moved in was as follows: consulting a primary care provider: 155 stayed, 83 moved out, 108 moved in; consulting a specialist: 614 stayed, 96 moved out, 153 moved in; I'm not sure: 30 stayed, 138 moved out, 37 moved in; managing at home or consulting a pharmacist: 38 stayed, 59 moved out, 59 moved in; seeking emergency care: 91 stayed, 34 moved out, 53 moved in; GP denotes general practitioner.

EMERGENCY DEPARTMENT UTILIZATION AND APPROPRIATENESS

Among users who initially planned to seek emergency care, 27.8% (35 of 126) revised their intention immediately after the assessment, and 38.5% (37 of 96) ultimately chose a different care level. Of these 96 users, 29 had a follow-up consultation within CUF with an appropriateness rating. In 27 cases (27 of 29; 95% CI, 78.0 to 98.1), emergency care was not considered necessary, suggesting patients' decision to avoid the ED was clinically appropriate.

In the two remaining cases where patients had a follow-up consultation, the DDSS recommended emergency care — one for pneumonia, the other for a suspected wrist fracture. Despite appropriate intentions and DDSS advice, one participant saw a PCP and the other a specialist.

Table 5. Change in Appropriateness of Care Plans before and after the Symptom Assessment (Paired Cohort, N=363).*

Care Category	Preassessment, N (%)	Postassessment, N (%)	Net Change (pp)	95% CI (pp)	McNemar χ^2 (CC)	P Value
Appropriate	109 (30.0)	128 (35.3)	5.23 pp	0.88 to 9.59	4.985	0.0256
Inappropriate	215 (59.2)	216 (59.5)	0.28 pp	−4.46 to 5.01	0	1
Unsure	39 (10.7)	19 (5.2)	−5.51 pp	−9.01 to −2.01	8.595	0.0034

*Overall, 385 non–emergency department participants were judged for appropriateness. Three participants had skipped the preassessment intention question and 20 participants skipped the postassessment intention question. Paired analysis includes only users with both pre- and postappropriateness classifications (n=363). "Unsure" indicates the participant had selected "I don't know" in response to the intent question before or after the assessment. CC denotes correlation coefficient; CI, confidence interval; and pp, percentage points.

Table 6. Change in Appropriateness between Preassessment Plans and Actual Behavior (Paired Cohort, N=382).*

Care Category	Preassessment, N (%)	Behavior, N (%)	Net Change (pp)	95% CI (pp)	McNemar χ^2 (CC)	P Value
Appropriate	114 (29.8)	246 (64.4)	34.55 pp	27.75 to 41.36	97.506	<0.0001
Inappropriate	226 (59.2)	136 (35.6)	−23.56 pp	−30.05 to −17.07	49.506	<0.0001
Unsure	42 (11.0)	0 (0.0)	N/A	N/A	N/A	N/A

*Overall, 385 non–emergency department participants were judged for appropriateness. Three participants had skipped the preassessment intention question. This yields 382 paired observations. CC denotes correlation coefficient; CI, confidence interval; N/A, not applicable; and pp, percentage points.

Overall, 72.0% of ED visits (95 of 132; 95% CI, 63.6 to 79.0) were classified as appropriate. This included 55 visits judged clinically appropriate, and an additional 40 that involved ED-specific interventions such as in-hospital diagnostics, treatment, or admission.

Among appropriate ED visits, 56.8% (54 of 95) were from patients who had not initially intended to seek emergency care. These visits involved slightly older patients (41.1±13.6 years of age) than inappropriate ones (36.0±12.9 years of age; P=0.053). When the DDSS recommended emergency care, 82.1% (32 of 39; 95% CI, 67.3 to 91.0) of those visits were later rated as appropriate (Table S7). Based on the Manchester Triage Scale, 33.7% of appropriate visits were classified as urgent, compared with 8.1% of inappropriate visits (P=0.003). Weekday visits were more often appropriate (74.7% vs. 57.5%; P=0.019).

To evaluate potential safety risks following changes in care intention, we examined downstream acute care utilization among users who de-escalated from an initial emergency care intention and among those who directly attended the CUF ED. Among the 207 users who de-escalated to primary care or specialist care after the assessment, 5 (2.4%)

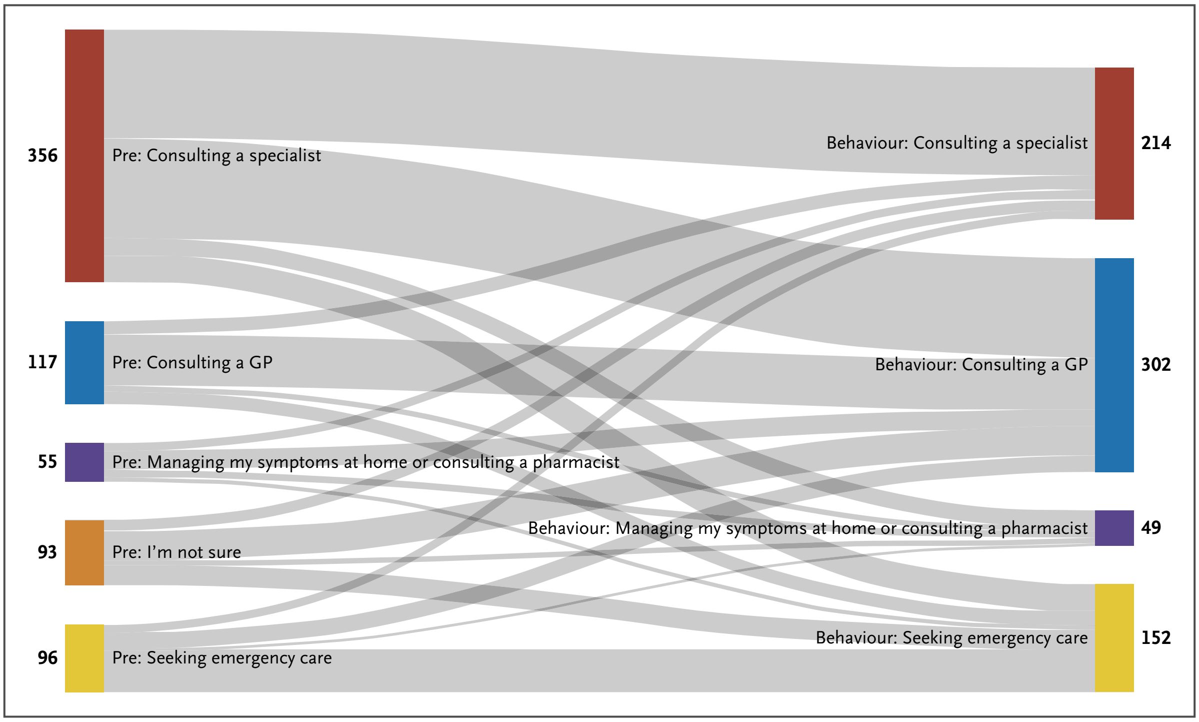

Figure 3. Transitions between Preassessment Intentions and Actual Behavior.

This figure presents a Sankey diagram illustrating transitions between care intentions expressed before the symptom assessment and the care actions subsequently taken. Node sizes represent the number of users in each category; link widths reflect the number transitioning between categories. The paired dataset included 717 users with both preassessment intention and observed behavior. The transition summary for each preassessment category is as follows: consulting a specialist: 152 stayed, 204 moved out, and 62 moved in; consulting a primary care provider: 74 stayed, 43 moved out, and 228 moved in; managing symptoms at home or consulting a pharmacist: 9 stayed, 46 moved out, and 40 moved in; seeking emergency care: 59 stayed, 37 moved out, and 93 moved in; I'm not sure: 0 stayed, 93 moved out, and 0 moved in; GP denotes general practitioner.

subsequently required ED observation. These events occurred between day 3 and day 11 and involved complications of tonsillitis, a nonresolving urinary tract infection, severe urticaria (following a teleconsultation referral), worsening jaundice in hepatitis, and prostatitis (one case with two ED contacts).

In the group of 132 users who directly attended the ED, 13 (9.8%) required a return ED visit and 2 of these were admitted during that return encounter. The proportion of users requiring emergency care after their initial decision was significantly lower in the de-escalated group than in the direct-to-ED group (5 of 207 [2.4%] vs. 13 of 132 [9.8%]; P=0.003).

MULTIVARIABLE ANALYSIS OF PREDICTORS OF APPROPRIATE BEHAVIOR

Because the logistic regression aimed to examine predictors of appropriate behavior across the whole cohort, both ED and non-ED users were included in the model. Of the 522 users with known behavior, 5 lacked an appropriateness judgement and 38 had missing predictor data, resulting in 479 complete cases for analysis.

The model was statistically significant (likelihood ratio $\chi^2(6)=13.82$, P=0.00011; pseudo $R^2=0.045$). Appropriate postassessment intention was the strongest predictor of appropriate real-world behavior (odds ratio, 2.84; 95% CI, 1.55 to 5.34; P=0.001). Female sex was also associated with higher odds of appropriate behavior (odds ratio, 1.69; 95% CI, 1.10 to 2.63; P=0.018). Age, preassessment appropriateness, and DDSS case advice appropriateness were not statistically significant predictors after adjustment. Full model coefficients are provided in Table S8.

GENERALIZABILITY

Compared with all users who completed the assessment (n=16,158), the study cohort (n=1470) showed lower rates of "I'm not sure" intentions (13.0% vs. 20.7%; P=0.012) and home management intentions (7.4% vs. 13.7%; P=0.0066). The cohort also had a higher proportion of primary care DDSS advice (74.8% vs. 68.5%; P<0.0001) and lower proportions of both self-care advice (12.1% vs. 14.9%; P=0.002) and emergency or ambulance advice (12.7% vs. 16.7%; P<0.0001). Sex distribution differed, with fewer women in the study population (57.7% vs. 65.0%; P<0.0001).

Users with known behavior (n=721) were slightly older than those with unknown behavior (n=749) (mean age, 39.5 vs. 37.7 years; P=0.005), while sex distribution (P=0.153) and number of symptoms (P=0.083) did not differ significantly. Differences were observed in care intentions and DDSS advice: Users with known behavior had higher emergency care intention both pre- and postassessment, whereas users with unknown behavior more often intended specialist care and selected home management options. These patterns aligned with DDSS advice distribution, with users with known behavior receiving more emergency-level advice and users with unknown behavior more often receiving specialist-aligned advice (P=0.033). Full results are provided in Tables S9 and S10.

Discussion

The ESSENCE study evaluated an AI-supported digital symptom assessment pathway integrated into Portugal's largest private health care network to support more appropriate health care utilization. Use of the tool was associated with measurable shifts in patient intentions and behaviors, reducing uncertainty and promoting care choices more aligned with clinical appropriateness. These changes corresponded to meaningful adjustments in service use, including increased use of primary care and reduced specialist and emergency visits, highlighting the potential of digital decision support tools to shape real-world health care behavior.

AI IN PATIENT EMPOWERMENT AND HEALTH CARE UTILIZATION

AI is increasingly recognized as a tool to support patients in navigating care decisions. By synthesizing symptom data and offering personalized guidance, AI-supported symptom assessment, when integrated into an actionable digital pathway, can contribute to improved efficiency and appropriateness. System-level benefits of AI-enabled triage have been demonstrated in prior work.[10] Early evaluations focused on diagnostic accuracy, but interest is now shifting toward real-world impact on behavior and outcomes. AI integrations have already been shown to influence health behaviors in areas such as hypertension management, smoking cessation, and digital cognitive behavioral therapy for pain and anxiety.[11-13] However, evidence on how DFD tools affect real care decisions remains limited.

A study evaluating the integration of AI-supported virtual triage with live nurse triage in a Portuguese health insurance plan found 22.8% of patient-members revised their care intent after triage — compared with over 30% in the present study — but behavior was only measured through outpatient bookings, excluding those choosing ED or self-care, with the focus of the study being mainly on intent comparison.[7] Australia's national virtual front door redirected 55% of ED-intent users to lower-acuity options, but adherence and appropriateness were not assessed.[8] A related large language model (LLM)–based tool showed good technical performance but also lacked behavioral validation.[14]

Together, these studies reflect a growing interest in AI-supported navigation tools, but also highlight a common limitation: reliance on AI performance or stated intent rather than observed behavior. In the present study, the poor agreement between postassessment intent and actual behavior (Cohen's κ=0.25) reinforces the need for robust real-world evaluation to understand how DFD tools influence patient actions and system outcomes.

DRIVERS OF BEHAVIOR CHANGE

Understanding health care–seeking behavior is inherently complex. In this study, one third of users revised their care intentions immediately after reviewing the report, suggesting that the assessment process and structured information significantly influenced their initial decisions.

After this, users were shown mapped service options tailored to their condition and urgency, codeveloped by CUF and the DDSS provider. These pathways formed a composite digital intervention, combining the informational content of the DDSS assessment with immediate access to bookable or actionable next-step options (e.g., primary care video consultations, same-day primary care appointments, and emergency care guidance). The high uptake of teleconsultations and same-day appointments suggests these actionable options strongly influenced behavior, and prior research shows that immediacy and convenience often outweigh urgency in care decisions, meaning that clear next steps likely acted as effective behavioral nudges.[15-17] Because our dataset does not include information on which links users viewed, clicked, or used to complete a booking, we could not analytically separate the influence of the DDSS advice from the effect of these service links or their associated convenience. As such, the observed behavior change should be interpreted as the impact of the combined AI-supported pathway rather than the DDSS component alone.

In addition, while triage recommendations guide decisions, patients also weigh whether advice fits their circumstances. Qualitative research shows people act on guidance that feels clear, reassuring, and feasible.[4] These patterns reflect behavioral economics models, where individuals often "satisfice" — choosing acceptable, if not optimal, options under uncertainty.[18] Patient engagement further improves when digital pathways offer transparency, personalization, and perceived empathy, creating a credible foundation for decision-making.[16] Future evaluations incorporating EHR linkage, click-level analytics, A/B testing of pathway elements, or user interviews will be important for isolating which specific report components, recommendations, service links, or contextual factors most strongly drive behavior change.

APPROPRIATENESS OF DECISION-MAKING

AI-driven DFDs aim to guide patients to appropriate care or safe self-management, improving outcomes and efficiency while reducing unnecessary or misdirected use. Yet, visit appropriateness is rarely evaluated — especially outside emergency care, where frameworks are more established.

These evaluations are also not straightforward. Retrospective judgments often lack insight into patient motivations or symptom progression. Tools focus on clinical effectiveness but often overlook barriers, reassurance needs, and cost considerations.[19] Even clinical ratings are subjective: Giannouchos et al. found 40%–60% discordance in ED triage when comparing discharge diagnoses with presenting complaints.[20]

Frameworks also tend to assume a single "correct" option, ignoring that multiple care paths may be reasonable. In practice, ambiguity, overlapping symptoms, and context shape behavior. Weekend visits were more often rated inappropriate than weekday visits; however, this pattern may also reflect reduced availability of primary care services rather than patient preference alone. In our setting, in-person primary care services operate only on Saturdays until 1.00 p.m. and are unavailable from Saturday afternoon through Sunday, during which time only primary care video consultations and the ED remain accessible. As a result, a visit that would be rated inappropriate during the week — when more service options are available — may be considered appropriate on weekends when no alternative in-person care is accessible. This nuance should be explicitly considered in future evaluations of appropriateness. At the same time, lower service availability may interact with factors such as convenience and perceived waiting times, which have previously been shown to influence care-seeking behavior and sometimes outweigh clinical reasoning, echoing findings from Kraaijvanger et al.[21] Both mechanisms represent plausible explanations for the observed weekend pattern.

Multivariable analysis further contextualized these appropriateness patterns. In a logistic regression, the strongest predictor of appropriate real-world behavior was having an appropriate intention after the assessment (odds ratio, 2.84; 95% CI, 1.55 to 5.34). Crucially, this postassessment intention was recorded before any service options or booking pathways were presented, meaning it could only have been shaped by the structured information contained in the symptom assessment itself. This suggests that the assessment process — including urgency framing, differential diagnosis, and explanatory content — contributed meaningfully to aligning users with appropriate care. Female sex was also associated with higher odds of appropriate behavior, whereas age, initial intention, and case-advice appropriateness were not independently predictive, highlighting that behavioral alignment may be driven more by how users interpret assessment outputs than by baseline characteristics.

Further insights come from experimental research outside clinical settings. In a randomized trial, Kopka et al. showed that laypeople using a DDSS improved triage accuracy (53.2% to 64.5%), tripled correct self-care recognition (13.1% to 36.9%), and increased emergency detection by 29%, without raising undertriage.[4] Users were also more likely to revise urgency (27% vs. 17% with a general-purpose LLM), supporting ESSENCE's behavioral findings.

Although few studies assess AI-supported decision-making before care is sought, ED-based evaluations exist. Taylor et al. studied an AI triage tool across EDs in the United States, reporting better critical illness detection (78.8% to 83.1%) and faster care.[10] Rather than binary labels, the study used downstream outcomes, such as critical care need and hospitalization, as proxies for quality.

STRENGTHS

The ESSENCE study is among the first prospective real-world evaluations of DDSS integrated into a health care system to influence patient utilization. Its methodological strength lies in combining self-reported intentions with objective follow-up data from EHRs and email surveys with a high response rate, allowing robust triangulation of planned versus actual behavior, which addresses a common limitation of digital health research. Conducted within Portugal's largest private health care network, the study benefited from a large, heterogeneous user base spanning the full spectrum of care from self-management to specialist and emergency visits, and uniquely applied structured appropriateness criteria beyond emergency settings. The rigorous prospective design, adherence to ISO 14155:2020 good clinical practice guidelines, 1 year of data collection, and inclusion of diverse adult specialties enhance generalizability and provide a strong foundation for understanding the real-world impact DDSSs have on health care utilization.

LIMITATIONS

Inclusion in the study cohort required a second explicit consent step in which users chose to share their assessment report with a CUF physician. Because this action is primarily relevant for individuals intending to seek CUF-based care, it likely introduced an inclusion bias toward more care-seeking users. Consistent with this, the study cohort showed lower uncertainty, fewer self-care intentions, and a higher proportion of primary care advice compared with the broader assessment population. In the wider DDSS user base (n=16,158), self-care intentions increased from 13.7% to 16.9% after the assessment, whereas in the study cohort, self-care remained substantially lower (7.4% to 7.3%). Follow-up data were also uneven: Behavior could only be fully captured for users who engaged with CUF services through EHR linkage, whereas out-of-network care relied on voluntary email surveys, likely underreporting self-care and avoiding visits. Reflecting these method differences, users with known behavior tended to receive higher-acuity advice and they expressed greater emergency care intention, whereas users with unknown behavior more often showed specialist or home management intentions.

Transportability of these findings to public or lower-resource health care systems requires caution. The study was conducted within a well-resourced private network with rapid access to primary and specialist care, low structural barriers, and digitally integrated booking pathways — conditions that differ substantially from many public systems. In settings where primary care access is constrained and specialist care involves long delays, care-seeking behavior may shift differently.

Public systems often aim to increase safe self-management to reduce avoidable demand; whether similar shifts would emerge in contexts with different access barriers, waiting times, and cost structures remains uncertain. Further evaluation in public or lower-resource environments — where behavioral responses, service substitution, and appropriateness thresholds may differ — is needed to assess the generalizability of these utilization patterns.

In addition, in this setting, ED visits are well captured, but private hospital admissions require a high up-front deposit, leading many patients to seek inpatient care in the public sector, which is not linked to CUF's EHR. As a result, downstream hospitalizations are systematically underascertained.

Demographic, intent, and DDSS advice data become available only once a user completes the symptom assessment. No information is collected at earlier attrition steps (e.g., invited→consented→started assessment), which limits the ability to compare users lost before assessment completion. Importantly, we had no data on which elements of the report users interacted with, including which conditions they opened in the app, or which service links they viewed or clicked. As a result, only the eventual care action (e.g., recorded CUF visit or self-reported behavior) could be evaluated, and we could not determine how exposure to specific report components shaped decisions.

In addition, appropriateness ratings were assessed using different methods by care level, and self-care decisions could not be evaluated because no clinical documentation was available for the 50 users who managed symptoms at home (6.9% of those with known behavior). Excluding these low-acuity cases — which would often be considered appropriate — likely introduces a downward bias, meaning that the reported appropriateness estimates may underestimate true appropriateness across all settings.

A further limitation of this study is the absence of a control group. Without a comparator group, we cannot determine how much of the observed change in care intentions and behavior would have occurred naturally over the course of the illness, independent of the digital pathway. Illness severity, symptom progression, or external factors — such as service availability or social input — may influence decisions even without any intervention.

Conclusion

The ESSENCE study provides real-world evidence that an AI-supported digital assessment pathway can contribute to more confident and appropriate patient decision-making. By combining personalized symptom assessment with actionable service options, the integration reduced uncertainty, shifted care intentions, and promoted more appropriate use — highlighting AI's potential as a behavioral nudge within care pathways.

Importantly, AI's impact extends beyond output accuracy to how information is structured, perceived, and acted upon. This underscores the need for user-centered design and thoughtful integration into health care systems.

Future work should prioritize expanding EHR linkage approaches to more accurately capture real-world behavior, while also examining in greater depth which specific report elements, recommendations, service links, and contextual factors — such as service availability, symptom evolution, or social input — drive behavior change. Complementary methods, including A/B testing to isolate the impact of individual components and user interviews to explore decision-making processes, will be valuable for identifying the mechanisms most responsible for influencing care-seeking actions. Broader evaluations in public and underserved settings are needed to assess generalizability, alongside health economics studies quantifying cost and system impact.

Data and statistical analysis governance: Statistical analysis was conducted by author F.C. A detailed independent statistical review and audit was conducted by CUF (independently of the DDSS developer and all coauthors affiliated with the developer) by the CUF Academic Center Team under the leadership of statistician S.D. and supervised by author P.F. Authors S.D. and P.F. had access to all raw data and all analyses and carried out an independent adjudication of all methods and analyses. These authors carried out additional independent analyses where they judged necessary and independently approved the final analyses and statistical reporting.

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

Data collected for the study, including individual participant data and a data dictionary defining each field in the set, will be made available to others after publication upon reasonable request, subject to approval. Requests for access should be made to the study team at Ada Health (science@ada.com). After approval, a signed data-sharing agreement will be required before data release.

This work was supported by the Federal Ministry of Research, Technology and Space (Bundesministerium für Forschung, Technologie und Raumfahrt) through the European Union–financed NextGenerationEU program under grant number 16KISA100K, project PATH — Personal Mastery of Health and Wellness Data.

We thank Dr. Sara Dias, statistics professor at the CUF Academic Center and Polytechnic Institute of Leiria, for her careful review of the statistical procedures required for this article.

Author Affiliations

[1] Ada Health, Berlin, Germany
[2] CUF, Carnaxide, Portugal

[3] Else Kröner Fresenius Center for Digital Health, Dresden University of Technology, Dresden, Germany

[4] Faculty of Business and Economics, TUD Dresden University of Technology, Dresden, Germany

References

1. Organisation for Economic Co-operation and Development. Health at a glance: Europe 2020: state of health in the EU cycle. Paris: OECD, 2020. DOI: 10.1787/82129230-en.
2. Rastogi N. Healthcare's new frontier: the digital front door. BMJ Innov 2022;8:129-132. DOI: 10.1136/bmjinnov-2021-000874.
3. Kopka M, Schmieding ML, Rieger T, Roesler E, Balzer F, Feufel MA. Determinants of laypersons' trust in medical decision aids: randomized controlled trial. JMIR Hum Factors 2022;9:e35219. DOI: 10.2196/35219.
4. Kopka M, von Kalckreuth N, Feufel MA. Accuracy of online symptom-assessment applications, large language models, and laypeople for self-triage decisions: a systematic review. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory. September 14, 2024 (https://www.medrxiv.org/content/10.1101/2024.09.13.24313657v1). Preprint.
5. Trivedi S, Littmann J, Stempien J, Kapur P, Bryce R, Betz M. A comparison between computer-assisted self-triage by patients and triage performed by nurses in the emergency department. Cureus. 2021;13:e14002. DOI: 10.7759/cureus.14002.
6. Oosterhoff M, Kouwenberg LHJA, Rotteveel AH, et al. Estimating the health impact of delayed elective care during the COVID-19 pandemic in the Netherlands. Soc Sci Med 2023;320:115658. DOI: 10.1016/j.socscimed.2023.115658.
7. Gellert GA, Galvão P, Gomes SM, et al. Impact of integrated virtual and live nurse triage on patient care seeking and health care delivery effectiveness and efficiency. Telemed Rep 2024;5:330-338. DOI: 10.1089/tmr.2024.0054.
8. McMahon B, McInerney D. Right care, right place, first time: how AI is improving national virtual front doors. NEJM AI 2025;2(6). DOI: 10.1056/aipc2401260.
9. Naouri D, Ranchon G, Vuagnat A, Schmidt J, El Khoury C, Yordanov Y. Factors associated with inappropriate use of emergency departments: findings from a cross-sectional national study in France. BMJ Qual Saf 2020;29:449-464. DOI: 10.1136/bmjqs-2019-009396.
10. Taylor RA, Chmura C, Hinson J, et al. Impact of artificial intelligence–based triage decision support on emergency department care. NEJM AI 2025;2(3). DOI: 10.1056/aioa2400296.
11. Chiang P-H, Wong M, Dey S. Using wearables and machine learning to enable personalized lifestyle recommendations to improve blood pressure. IEEE J Transl Eng Health Med 2021;9:1-13. DOI: 10.1109/jtehm.2021.3098173.
12. Carrasco-Hernandez L, Jódar-Sánchez F, Núñez-Benjumea F, et al. A mobile health solution complementing psychopharmacology-supported smoking cessation: randomized controlled trial. JMIR MHealth UHealth 2020;8:e17530. DOI: 10.2196/17530.
13. Sadeh-Sharvit S, Camp TD, Horton SE, et al. Effects of an artificial intelligence platform for behavioral interventions on depression and anxiety symptoms: randomized clinical trial. J Med Internet Res 2023;25:e46781. DOI: 10.2196/46781.
14. Naved BA, Ravishankar S, Colbert GE, Johnston A, Slott QM, Luo Y. LLM enabled classification of patient self-reported symptoms and needs in health systems across the USA. NPJ Digit Med 2025;8:390. DOI: 10.1038/s41746-025-01779-9.
15. McDonnell T, Nicholson E, Conlon C, et al. Assessing the impact of COVID-19 public health stages on paediatric emergency attendance. Int J Environ Res Public Health 2020;17:6719. DOI: 10.3390/ijerph17186719.
16. Li Y-H, Li Y-L, Wei M-Y, Li G-Y. Innovation and challenges of artificial intelligence technology in personalized healthcare. Sci Rep 2024;14:18994. DOI: 10.1038/s41598-024-70073-7.
17. Song Y, Wu M, Feng H. Influencing factors of patients' behavior of healthcare seeking: a meta-analysis in China. Front Public Health 2025;13:1583075. DOI: 10.3389/fpubh.2025.1583075.
18. Becerra Pérez MM, Menear M, Brehaut JC, Légaré F. Extent and predictors of decision regret about health care decisions: a systematic review. Med Decis Making 2016;36:777-790. DOI: 10.1177/0272989X16636113.
19. Coulter I, Herman P, Ryan G, Hilton L, Hays RD, Members of CERC Team. The challenge of determining appropriate care in the era of patient-centered care and rising health care costs. J Health Serv Res Policy 2019;24:201-206. DOI: 10.1177/1355819618815521.
20. Giannouchos TV, Ukert B, Wright B. Concordance in medical urgency classification of discharge diagnoses and reasons for visit. JAMA Netw Open 2024;7:e2350522. DOI: 10.1001/jamanetworkopen.2023.50522.
21. Kraaijvanger N, van Leeuwen H, Rijpsma D, Edwards M. Motives for self-referral to the emergency department: a systematic review of the literature. BMC Health Serv Res 2016;16:685. DOI: 10.1186/s12913-016-1935-z.