Evaluation of symptom checker formats to support health literacy and trust in AI: Results from an online randomised-controlled trial

Julie Ayre[1], Karen Gallagher[2], Jenna Smith[1], Kirsten Ingwersen[1], Claire Hudson[1], Angelica Scott[2], Alison Woods[2], Christopher Ng[2], Yuanee Wickramasinghe[2], Ivan CK Ma[3], Waren Nadesan[3], Gauri Kapoor[3], Geoffrey Edlund[3], Lorna Butters[4], Trang Vu[3], Kirsten J McCaffery[1]

[1] Sydney Health Literacy Lab, Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, NSW, Australia

[2] Healthdirect Australia, Sydney, Australia

[3] Co-SHeLL panel, Sydney Health Literacy Lab, Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, NSW, Australia

Abstract

Objectives: Evaluate the impact of online symptom checker formats on symptom management knowledge, symptom checker trust and acceptability, and behavioural intentions.

Design: Two 5-arm parallel-group online randomised controlled trials.

Setting: Online survey.

Participants: Sample of 2062 Australian adults recruited through online research panel in June 2025. Almost half identified as man/male (49%) and woman/female (51%), median age 49 years (IQR=28). Participants viewed a hypothetical health scenario (fever and vomiting) followed by a screenshot of an online symptom checker from national health service provider, healthdirect.

Interventions: Participants were randomised to symptom acuity level (low: self-care at home; or moderate: see a General Practitioner (GP) in 24 hours) and one of five symptom checker formats. The standard format showed the existing healthdirect symptom checker advice. The remaining formats were 'AI-enhanced' versions. These included an AI-enhanced version with e.g. more tailored advice, rationale for acuity level, and AI disclosure statement. The other AI-enhanced formats had additional features: numbered steps, multimedia, and more detailed information about the use of AI.

Main outcome measures: Primary outcomes were intentions to follow the symptom checker's self-care advice and intentions to see a GP in 24 hours. Secondary outcomes were trust in advice, knowledge of symptom management, and acceptability of the tool. All outcomes were assessed immediately post-intervention; knowledge was also assessed after 2 weeks.

Results: When advised to self-care at home, the AI-enhanced groups reported lower intentions to see a GP in 24 hours (median 3.00 out of 5), compared to the standard (original) tool (median 4.00; adjusted $p = 0.003$). There were no other significant effects on intentions. Immediately following the intervention, participants who viewed an AI-enhanced format reported greater knowledge about how to manage current and changing symptoms, across both acuity levels (adjusted p's <0.001). Knowledge gains were not sustained at 2 week follow-up. There were no significant effects on trust or acceptability.

Conclusions: Participants who viewed the more tailored information in the AI-enhanced formats demonstrated stronger knowledge for managing symptoms than those who viewed the standard format. There was also some evidence that an AI-enhanced format may be more effective at reducing use of primary care for symptoms that can be

managed at home. Trust and acceptability were high across formats, and the explicit use of AI did not impact significantly on these outcomes. Future research should investigate these formats using interactive prototypes across a wider variety of health contexts.

Registration: ACTRN ACTRN12625000474459p

Key messages

- **What is already known on this topic:** Although online, evidence-based symptom checkers have been widely available from reputable health organisations for over a decade, they often face poor uptake and may not adequately meet health literacy needs of diverse users.
- **What this study adds**: Symptom checker features that could be implemented with AI, such as tailored information and a clear rationale for triage advice, may help support appropriate symptom management. Statements about the tool's use of AI did not appear to impact trust or acceptability of the symptom checker tool.
- **How this study might affect research, practice or policy**: Findings from this study suggest that using AI to enhance symptom checker advice may not impact negatively on trust and acceptability of the tool, and may improve appropriate symptom management. Further research is needed to investigate AI-enhanced symptom checker formats using interactive prototypes across a wider variety of health contexts.

INTRODUCTION

Online symptom checkers are evidence-informed tools that can help people manage their symptoms and seek appropriate care. These tools have been publicly available for over a decade, via mainstream services such as the UK's National Health Service and Australia's healthdirect.[1] They are increasingly relevant in the current policy climate which emphasises reducing unnecessary and low-value care, and providing accessible, trustworthy health advice.[2 3] Though several studies have highlighted potential issues with accuracy, or argue their advice is too risk-averse to substantially reduce health system burden,[4 5] online symptom checkers remain an important scalable tool for timely tailored health advice, particularly in the face of general-purpose artificial intelligence (AI) tools such as ChatGPT.[6-9]

Research shows that people are generally supportive of online symptom checkers, [5 10-12] but there is also mounting evidence that they could better meet community needs. For example, even if people use an online symptom checker, they may not follow the triage advice. Whilst some studies suggest that intentions to follow symptom checker advice are high (e.g. 87% of 634 UK adults reported following the advice most/all of the time)[11], translation into action is likely much lower.[13 14] Following advice may also be lower when advice is more urgent, although evidence for this is mixed.[13-15]

Current designs for online symptom checkers may also inadvertently contribute to health inequity and the digital divide. For example, there is growing evidence that users are more likely to be female, educated, white, and to have stronger digital literacy skills.[4 5 16] People who find these tools easier to use are up to 8 times more likely to use them, [11] with qualitative research suggesting that usability may be hindered by poor implementation of health literacy principles, particularly around the use of medical jargon.[5 10 17 18]

As is the case with other health information sources, trust likely plays a key role in use of online symptom checkers.[19] Some barriers to trust are common to most digital health tools e.g. lack of endorsement by trusted authorities and concerns about data privacy.[5] Other barriers to trust are more specific, such as users feeling that their symptoms were

poorly represented by the tool or that the health advice was not plausible, vague, or lacked transparency about the tool's decision-making process.[5 10 18 20 21]

Recent technological advances may provide an opportunity to address some of these issues and develop online symptom checkers that better meet the needs of diverse communities. For example, tools that integrate generative AI may be able to more flexibly tailor content for users,[22] and there is growing evidence that generative AI may be an effective tool for creating health materials that more closely adhere to plain language guidelines.[23 24] These potential benefits must be weighed against potential risks to safety and erosion of trust, particularly if the symptom checker is delivered through a reputable organisation.[25 26]

Although there is a well-established literature evaluating the accuracy of AI triage tools,[4 5] only a few studies have investigated the use of AI to make symptom checker communication more accessible and improve end-user experience. Some limited evidence suggests people are just as trusting of symptom checker tools that are explicit about their use of AI as those that are not.[27] There is also some evidence that the manner in which a symptom checker explains triage advice can impact on trust, with one study reporting greater trust in advice for managing unfamiliar symptoms when there was a clear link between symptom input and advice, compared to when no explanation was provided.[28] However, both studies were conducted prior to the release of publicly available generative AI tools such as ChatGPT, and it is likely that the public's relationship with and expectations of AI have changed substantially.

To address these research gaps, this study aimed to evaluate the impacts of different symptom checker formats relating to health literacy and potential AI-enhancements on symptom management knowledge, symptom checker trust and acceptability, and behavioural intentions. This research question was assessed across low and moderate symptom acuity levels i.e. situations where symptoms can be managed through self-care (low acuity) or by accessing primary care in 24 hours (moderate acuity).

METHODS

Study Design

This study involved two 5-arm parallel-group online randomised controlled trials (Figure 1). Participants were randomly allocated to an intervention group using survey platform Qualtrics' randomizer function, a pseudorandom number generator based on the Mersenne Twister. The trial was prospectively registered with the Australian New Zealand Clinical Trials Registry (ACTRN12625000474459p) in May 2024, and approved by the University of Sydney Research Ethics Committee (2025/HE000143).

Digital health context

Healthdirect Australia is a national government-funded health service that provides free health advice and information via a website, app and telephone helpline. The healthdirect online symptom checker was launched in 2014 and is based on an established evidence-based clinical decision support technology. At the time of writing, the existing symptom checker did not use generative AI. Advice is tailored according to triage category (e.g. self-care, see a general practitioner (GP) in 24 hours, go to emergency) rather than specific symptoms. The tool can also connect users with local service information and availability.

Healthdirect also hosts more than 3000 web pages on health topics and medicines information. This provides a curated knowledge library that an AI model could draw from, allowing the symptom checker to provide more tailored advice. For example, retrieval-augmented generation (RAG) frameworks can connect an AI model with external knowledge.[29] Although using a RAG framework cannot completely ensure clinical accuracy, studies demonstrate it can considerably improve the accuracy of generative AI tools compared to those that do not provide a knowledge base.[30 31]

Participants

Eligible participants were adults (18+ years) who resided in Australia and who self-reported good English proficiency. Given that online symptom checkers may be underutilised by people who have not had opportunity for tertiary education,[5] we used quotas to ensure that no more than 50% of the sample had a university education. Participants were recruited between June 19th and 22nd, 2025, via an online research panel comprising more than 550,000 Australian adults. Participants were aware that they would be allocated to different symptom checker formats, but not that some formats incorporated generative AI. Participants completed the follow-up survey at least two weeks later.

Procedure

The survey was hosted using Qualtrics software. After consenting, participants completed questions about demographics, health literacy,[32] and digital health literacy.[33] The survey items also asked about awareness and previous use of healthdirect and AI tools (Appendix B).

Participants across all intervention groups were then asked to imagine the following health scenario:

> *You currently have a fever and have vomited twice. You're worried that you are dehydrated so you try sipping on some apple juice, which you immediately throw up again. Yesterday you had a little less energy than usual, but you were still able to complete the things you needed to do.*
>
> *You are not quite sure what to do, so you use the healthdirect symptom checker. The symptom checker is an online tool to help people decide whether to see a doctor or look after themselves at home. To help you find a local service it will also ask for your postcode, you put in Barangaroo 2000.*
>
> ***The symptom checker does not give a diagnosis and is not a substitute for professional healthcare.***

Participants were then randomised to one of 10 intervention groups (Figure 1), and were asked to view the corresponding health advice for at least 2 minutes. Immediately after viewing the advice, participants completed an attention check that asked them to identify the key symptoms described in the health scenario. Participants then completed primary and secondary outcomes.

After two weeks participants completed a second survey. This reminded them of the original health scenario, and participants were asked to recall their triage advice, and completed the knowledge outcome. At the end of the study all participants were debriefed, including a statement that the scenario was hypothetical and that they did not receive information or advice regarding their personal health situation.

Intervention groups

All participants viewed a screenshot of the advice page (key components shown in Figure 2). Links and video content shown in the screenshot were also available to participants.

Symptom acuity level

Low acuity symptoms

Participants were advised to manage their symptoms at home (Figure 2).

Moderate acuity symptoms

Participants were advised to see a GP within 24 hours and provided with links to local GPs and Urgent Care Services for postcode 2000, the postcode used in the scenario (eFigure 1).

Symptom checker formats

Standard

The standard format reflected the original symptom checker tool's content and format. Advice about symptom management was not specific to the symptoms (eFigures 2 and 3).

AI-enhanced

This format reflected several AI-enhancements, including advice that was tailored to specific symptoms, links to relevant information sources, a statement that the content had been generated by AI, and a reassuring tone for low acuity presentations (eFigures 4 and 5). The advice also explicitly described how the symptoms related to the acuity level. These features appeared in all groups other than the Standard group.

AI-enhanced + numbered steps

This format included numbered steps for following self-management advice (eFigures 6 and 7).

AI-enhanced + multimedia

This format incorporated a video about managing fever and an image showing the symptoms of fever (eFigures 8 and 9).

AI-enhanced + detailed information about AI

This format provided more detailed information about how AI was integrated into the tool (eFigures 10 and 11).

Outcomes

All outcomes were assessed immediately following the intervention. Knowledge was also assessed at 2-week follow-up.

Intention to follow symptom checker advice

The primary outcome was intention to follow symptom checker advice. This was assessed using two items. The first item assessed intentions to follow the self-care advice, using a 4-point scale from *"I would follow all the advice"* to *"I would follow none of the advice".* The second assessed intentions to see a GP in 24 hours; participants responded using a 5-point scale from "*definitely would*" to "*definitely would not*".

Trust

Trust was assessed using three instruments. The first was a single-item 10-point Likert scale that asked participants to rate how trustworthy they found the advice, adapted

from Mac and colleagues.[34] Cognitive trust (4 items) and emotional trust (3 items) were each assessed using a validated instrument[35] that has previously been used to evaluate AI-enhanced symptom checker advice.[36] Reponses were recorded on a 7-point Likert scale.

Knowledge

Knowledge was assessed using purpose-built multiple-choice questions that were based on the content available in the intervention (Appendix B eTable 1). This produced two scores: one for managing current symptoms and the second for managing changing symptoms. Scores are presented as the proportion (%) of correct responses. At 2-week follow-up participants were also asked to recall the acuity level (i.e. to self-care at home for low acuity and to see a GP in 24 hours for moderate acuity).

Acceptability

Acceptability of the symptom checker was assessed using a 5-point Likert scale from (1) “strongly disagree” to (5) “strongly agree.” Three purpose-built statements asked about the extent that they would use the tool in the future, recommend the tool to a friend; and the extent that the tool met their expectations.

Items relating to emotional responses[37] were also collected. They are not analysed here.

Analysis

An intention to treat analysis was performed, comparing the 5 format groups for each symptom acuity level. Descriptive statistics are presented as medians, interquartile ranges and means for continuous variables and relative percentages for categorical variables. Statistical significance of intervention group differences was tested using Kruskal-Wallis tests comparing mean ranks, and one-way ANOVA tests where appropriate, where significance is considered with p values < 0.05. In the case of detecting overall differences, post-hoc comparisons using Dunn’s test with Bonferroni adjustments were planned between each group. The effect of group assignment on the change in the knowledge outcomes repeated at 2 weeks was assessed using linear

regression, adjusting for baseline measurement. Intervention group was not masked prior to this planned analysis.

Effect size and precision (95% confidence interval) for non-parametric tests were estimated using η^2 (rule of thumb for small effect 0.01; medium = 0.06; large = 0.14).[38] Sensitivity analyses investigated the robustness of primary analyses for participants who passed the attention check. Stratified analyses explored differences according to key characteristics: gender (man; woman), age (≤60 years vs >60 years), education (university education vs no university education), chronic disease (no chronic disease; at least one self-reported chronic disease), and previous ChatGPT use (none in the past 6 months; at least a few times in the past 6 months). Stratified analyses for health literacy and digital health literacy were not explored given the relatively small number in some cells. A further sensitivity analysis was performed by imputing the primary outcomes using multiple imputation using chained equations for those who did not complete the survey.

As a low-risk study, we did not anticipate any serious harms.

Sample Size

The total sample size required was 2000 participants. With a 2 x 5 parallel group design, with 90%, alpha = 0.05, and a small effect size, we estimated 160 participants per intervention group. A further 20% was added to account for drop-out at follow-up.

Patient and public involvement

A working group comprising nine people helped shape the intervention designs. After describing the goals of the project and providing foundational information about generative AI, working group members discussed their needs, priorities, and expectations for the tool. Insights from these discussions were incorporated into four prototypes to elicit specific feedback in a second workshop. Design strategies that the working group and healthdirect staff considered essential were incorporated into the AI-enhanced format. For more detail see Appendix C.

RESULTS

Sample characteristics

2546 participants were randomised to an intervention group, of which 2062 comprised the analysis sample (Figure 1; Tables 1 and 2). Across the sample 49% of participants identified as man or male, 51% identified as woman or female, and the median age was 49 years (IQR=28). Fifty-six per cent of participants had obtained a university degree and 16% were identified as having limited health literacy based on the screening item. Two-fifths (43%) of the sample had heard of healthdirect, and 14% had previously used the healthdirect symptom checker tool. The sample was familiar with generative AI tools like ChatGPT, with 56% of participants reporting that they had used ChatGPT at least a few times in the past 6 months (n=1164), and 36% reported that they had used it for health advice (n=740). The most common reason for using ChatGPT for health purposes was to ask about symptoms (19%, n=385). The median digital health literacy score was 31.0 (IQR: 5.0), indicating moderate to high confidence in this domain. No harms were reported in this study. Characteristics of participants who completed follow-up (N=1531, 74.2%), and were similar to those of the baseline sample (eTable 1).

Table 1. Participant characteristics, low acuity symptom level (self-care), by format intervention group (n, %, unless otherwise specified)

Variable	Standard (original tool) N = 215[1]	AI-enhanced N = 206[1]	AI-enhanced + Multimedia N = 202[1]	AI-enhanced + Numbered steps N = 206[1]	AI-enhanced + Detailed info about AI N = 212[1]
Age (years) [1]	49.0 (35.0, 64.0)	47.5 (34.0, 63.0)	49.0 (35.0, 63.0)	51.0 (39.0, 65.0)	48.5 (36.0, 63.0)
Gender					
Man	98 (46%)	97 (47%)	96 (48%)	100 (49%)	110 (52%)
Woman	117 (54%)	109 (53%)	105 (52%)	105 (51%)	101 (48%)
Trans and/or gender diverse	0 (0%)	0 (0%)	0 (0%)	1 (0.5%)	1 (0.5%)
I use a different term / Prefer not to say	0 (0%)	0 (0%)	1 (0.5%)	0 (0%)	0 (0%)
Born in Australia	157 (73%)	162 (79%)	161 (80%)	164 (80%)	152 (72%)
What is the main language you speak at home?					
Other	19 (8.8%)	16 (7.8%)	15 (7.4%)	13 (6.3%)	17 (8.0%)
English	196 (91%)	190 (92%)	187 (93%)	193 (94%)	195 (92%)
Are you of Aboriginal or Torres Strait Islander origin?					
Aboriginal and/or Torres Strait Islander	6 (2.8%)	9 (4.4%)	6 (3.0%)	6 (2.9%)	7 (3.3%)
No	208 (97%)	196 (96%)	196 (97%)	199 (97%)	202 (97%)
Unknown	1 (0.5%)	1 (0.5%)	0 (0%)	1 (0.5%)	3 (1.4%)
Highest level of education					
No school or other qualifications	0 (0%)	3 (1.5%)	2 (1.0%)	5 (2.4%)	2 (0.9%)
School Certificate or intermediate certificate (or equivalent)	11 (5.1%)	10 (4.9%)	6 (3.0%)	8 (3.9%)	17 (8.0%)
Higher school certificate or leaving certificate (or equivalent)	30 (14%)	38 (18%)	28 (14%)	39 (19%)	30 (14%)
Trade Apprenticeship	10 (4.7%)	7 (3.4%)	10 (5.0%)	7 (3.4%)	4 (1.9%)
Diploma or Certificate	41 (19%)	27 (13%)	40 (20%)	34 (17%)	33 (16%)
University degree	123 (57%)	121 (59%)	116 (57%)	113 (55%)	126 (59%)

Variable	Standard (original tool) N = 215[1]	AI-enhanced N = 206[1]	AI-enhanced + Multimedia N = 202[1]	AI-enhanced + Numbered steps N = 206[1]	AI-enhanced + Detailed info about AI N = 212[1]
Health literacy screener					
Adequate health literacy	181 (84%)	166 (81%)	178 (88%)	178 (86%)	176 (83%)
Limited health literacy	34 (16%)	40 (19%)	24 (12%)	28 (14%)	36 (17%)
eHEALS Score (8-40) [1]	31.0 (28.0, 33.0)	31.0 (28.0, 33.0)	31.0 (28.0, 33.0)	32.0 (28.0, 33.0)	31.0 (27.0, 34.0)
Number of chronic conditions					
0	119 (55%)	117 (57%)	118 (58%)	103 (50%)	114 (54%)
1	63 (29%)	56 (27%)	68 (34%)	62 (30%)	68 (32%)
2	27 (13%)	24 (12%)	11 (5.4%)	30 (15%)	19 (9.0%)
3+	6 (2.8%)	9 (4.4%)	5 (2.5%)	11 (5.3%)	11 (5.2%)
healthdirect use					
Helpline	25 (12%)	24 (12%)	22 (11%)	20 (9.7%)	24 (11%)
Symptom Checker	32 (15%)	31 (15%)	26 (13%)	35 (17%)	27 (13%)
Web-based information	43 (20%)	46 (22%)	51 (25%)	42 (20%)	40 (19%)
Videos	7 (3.3%)	7 (3.4%)	13 (6.4%)	8 (3.9%)	10 (4.7%)
Service Finder	16 (7.4)	23 (11%)	14 (6.9%)	18 (8.7%)	19 (9.0%)
Mobile App	11 (5.1%)	17 (8.3%)	10 (5.0%)	12 (5.8%)	16 (7.5%)
ChatGPT use for health information (in the last 6 months)					
A few times	34 (16%)	56 (27%)	35 (17%)	37 (18%)	49 (23%)
Once a month	14 (6.5%)	11 (5.3%)	16 (7.9%)	12 (5.8%)	11 (5.2%)
Once a week	9 (4.2%)	9 (4.4%)	8 (4.0%)	9 (4.4%)	14 (6.6%)
More than once a week	8 (3.7%)	12 (5.8%)	7 (3.5%)	8 (3.9%)	10 (4.7%)
Has not used ChatGPT for health	150 (70%)	118 (57%)	136 (67%)	140 (68%)	128 (60%)

Table 2. Participant characteristics, moderate acuity symptom level (see a GP in 24 hours), by format intervention group (n, %, unless otherwise specified)

Variable	Standard (original tool) N = 183	AI-enhanced N = 209	AI-enhanced + Multimedia N = 219	AI-enhanced + Numbered steps N = 203	AI-enhanced + Detailed info about AI N = 207
Age (years) [1]	50.00 (34.00, 63.00)	44.00 (34.00, 59.00)	49.00 (36.00, 65.00)	48.00 (33.00, 63.00)	49.00 (33.00, 65.00)
Gender					
Man	86 (47%)	107 (51%)	104 (47%)	100 (49%)	103 (50%)
Woman	97 (53%)	100 (48%)	115 (53%)	103 (51%)	104 (50%)
Trans and/or gender diverse	0 (0%)	1 (0.5%)	0 (0%)	0 (0%)	0 (0%)
I use a different term / Prefer not to say	0 (0%)	1 (0.5%)	0 (0%)	0 (0%)	0 (0%)
Born in Australia	136 (74%)	157 (75%)	163 (74%)	157 (77%)	165 (80%)
What is the main language you speak at home?					
Other	9 (4.9%)	13 (6.2%)	19 (8.7%)	19 (9.4%)	15 (7.2%)
English	174 (95%)	196 (94%)	200 (91%)	184 (91%)	192 (93%)
Are you of Aboriginal or Torres Strait Islander origin?					
Aboriginal and/or Torres Strait Islander	4 (2.2%)	5 (2.4%)	9 (4.1%)	9 (4.4%)	5 (2.4%)
No	179 (98%)	203 (98%)	209 (96%)	194 (96%)	200 (98%)
Unknown	0 (0%)	1 (0.5%)	1 (0.5%)	0 (0%)	2 (1%)
Highest level of education					
No school qualifications	0 (0%)	2 (1.0%)	3 (1.4%)	2 (1.0%)	2 (1.0%)
School Certificate	10 (5.5%)	7 (3.3%)	18 (8.2%)	9 (4.4%)	16 (7.7%)
Higher school certificate	27 (15%)	33 (16%)	35 (16%)	32 (16%)	32 (15%)
Trade Apprenticeship	10 (5.5%)	11 (5.3%)	11 (5.0%)	11 (5.4%)	12 (5.8%)
Diploma or Certificate	38 (21%)	33 (16%)	35 (16%)	41 (20%)	34 (16%)
University degree	98 (54%)	123 (59%)	117 (53%)	108 (53%)	111 (54%)
Health literacy screener					
Adequate health literacy	142 (78%)	177 (85%)	184 (84%)	169 (83%)	181 (87%)
Limited health literacy	41 (22%)	32 (15%)	35 (16%)	34 (17%)	26 (13%)
eHEALS Score (8-40) [1]	31.00 (26.00,	32.00 (28.00,	31.00 (28.00,	32.00 (28.00,	31.00 (27.00,

Variable	Standard (original tool) N = 183	AI-enhanced N = 209	AI-enhanced + Multimedia N = 219	AI-enhanced + Numbered steps N = 203	AI-enhanced + Detailed info about AI N = 207
	33.00)	33.00)	33.00)	33.00)	32.00)
Number of chronic conditions					
0	108 (59%)	128 (61%)	118 (54%)	122 (60%)	112 (54%)
1	47 (26%)	57 (27%)	68 (31%)	55 (27%)	55 (27%)
2	17 (9.3%)	13 (6.2%)	25 (11%)	20 (9.9%)	25 (12%)
3+	11 (6.0%)	11 (5.3%)	8 (3.7%)	6 (3.0%)	15 (7.2%)
healthdirect use					
Helpline	30 (16%)	14 (6.7%)	31 (14%)	25 (12%)	28 (14%)
Symptom Checker	30 (16%)	24 (11%)	35 (16%)	22 (11%)	26 (13%)
Web-based Information	47 (26%)	39 (19%)	46 (21%)	46 (23%)	41 (20%)
Videos	10 (5.5%)	6 (2.9%)	10 (4.6%)	6 (3.0%)	10 (4.8%)
Service Finder	11 (6.0%)	10 (4.8%)	21 (9.6%)	14 (6.9%)	18 (8.7%)
Mobile App	12 (6.6%)	5 (2.4%)	16 (7.3%)	10 (4.9%)	7 (3.4%)
ChatGPT use for health information (in the last 6 months)					
A few times	37 (20%)	43 (21%)	53 (24%)	29 (24%)	42 (20%)
Once a month	12 (6.6%)	15 (7.2%)	11 (5.0%)	12 (5.9%)	17 (8.2%)
Once a week	8 (4.4%)	11 (5.3%)	10 (4.6%)	9 (4.4%)	8 (3.9%)
More than once a week	9 (4.9%)	5 (2.4%)	11 (5.0%)	6 (3.0%)	2 (1.4%)
Has not used ChatGPT for health	117 (64%)	135 (65%)	134 (61%)	127 (63%)	137 (66%)

Evaluation of Symptom Checker intervention formats

Descriptive statistics for outcomes are shown in Figure 3 and eTables 2 and 3.

Low acuity symptoms (self-care)

Intention

For the low acuity presentation, which advised to self-care at home, there was a statistically significant effect of intervention format on intention to see a GP in 24 hours ($H(4) = 13.42$, $p = 0.009$). The effect size was small ($\eta^2 = 0.009$; 95% CI:0.001-0.03). Post-hoc analysis showed that participants who viewed the AI enhanced format (median = 3.00) had significantly lower intentions to see a GP within 24 hours than those in the Standard (original) tool group (median = 4.00; adjusted $p = 0.003$). There was no significant effect of intervention format on intention to follow self-care advice for the low acuity symptoms ($H(4) = 1.41$, $p=0.843$).

Knowledge

For the low acuity symptoms there was a statistically significant effect of intervention format on both knowledge measures – managing current symptoms and managing changing symptoms ($H(4) = 108.16$, $p <0.001$; $H(4) = 67.38$, $p <0.001$, respectively). The effect size was moderate for both outcomes (current symptoms: $\eta^2 = 0.101$; 95% CI:0.07-0.14; managing symptoms: $\eta^2 = 0.061$; 95% CI:0.04-0.10). Post-hoc analyses demonstrated that mean ranks for all AI-enhanced groups were significantly higher than those of the Standard group (original tool) (all adjusted p's <0.001; Figure 3 and eTable 2).

Trust

Intervention format did not have a significant effect on any trust outcomes for the low acuity symptoms (eTable 2).

Acceptability

Intervention format did not have a significant effect on acceptability for the low acuity symptoms (eTable 2).

Moderate acuity symptoms (see a GP in 24 hours)

Intention

For the moderate acuity symptoms, there was no significant effect of intervention format on intention to see a GP in 24 hours (H(4) = 8.35, p=0.080) or to follow the self-care advice (H(4) = 2.03, p=0.730) (eTable 3).

Knowledge

For the moderate acuity symptoms there was a statistically significant effect of intervention format on both knowledge measures – managing current symptoms and managing changing symptoms (H(4) = 72.14, p <0.001; H(4) 52.93, p <0.001, respectively). The effect size was moderate for both outcomes (current symptoms: η^2 = 0.067; 95% CI:0.04-0.14; managing symptoms: η^2= 0.048; 95% CI(0.03-0.08)). Post-hoc analyses demonstrated that mean ranks for all AI-enhanced groups were significantly higher than those of the Standard (original tool) group (all adjusted p's <0.001; eTable 3).

Trust

Intervention format did not have a significant effect on any trust outcomes for the moderate acuity symptoms (eTable 3).

Acceptability

Intervention format did not have a significant effect on acceptability for the moderate acuity symptoms (eTable 3).

Knowledge at follow-up

Within each symptom acuity level, all formats reported similar levels of knowledge about current and changing symptoms at two-week follow-up (time point 2) (eTables 4-5).

For the low acuity symptoms, the changes between time point 1 and time point 2 represented a small but significantly larger decrease in knowledge of current symptoms for participants in the AI-enhanced group, compared to those in the Standard group (original tool) (MD= -14.3%, 95% CI= -18.3% to -12.3%, p=0.006). For the moderate acuity symptoms (see a GP in 24 hours), we observed that two of the AI-enhanced formats showed small but significantly larger decreases in knowledge of current symptoms relative to the Standard group (original tool)

(Detailed info about AI: MD= -8.0%, 95% CI= -11.1% to -4.9%, p=0.023); Multimedia: MD= -9.4%, 95% CI= -12.3% to -6.6%, p=0.003). There were no significant changes in knowledge about changing symptoms over time.

Exploratory analyses

Participants who correctly recalled at least one of the two symptoms in the health scenario (vomiting, fever), and no incorrect symptoms, were included in a sensitivity analysis to explore the impact of attention on the statistical findings. Some participants (9.3% of the total sample, n=191) did not demonstrate appropriate attention. Analyses following the removal of inattentive participants, yielded very similar results, demonstrating that study findings were robust to inattentive participants (eTables 6 and 7).

Stratified analyses reported some differences in the effects of the Standard (original tool) and pooled effects of the AI-enhanced formats, depending on participant characteristics. These differences were only observed for the low acuity symptoms, in which the symptom checker recommended that people self-care at home: participants who viewed any AI format reported lower intentions to see a GP in 24 hours than those in the Standard (original tool) group if they had a university education (p=0.012), had not used ChatGPT in the past 6 months (p=0.012), did not have a chronic condition (p=0.007), and were ≤60 years old (p=0.008). This pattern was not observed for people without a university education (p=0.12), those who had used ChatGPT at least a few times in the past 6 months (p=0.10), people who reported at least one chronic condition (p=0.2), or those who were more than 60 years old (p=0.2). No differences were observed for gender.

Characteristics of those who did not complete the survey were similar overall and within intervention groups, to those included in the analysis. The sensitivity analysis using multiple imputation of the missing primary outcomes showed comparable results (eTable 8).

DISCUSSION

This study evaluated the effects of personalised, AI-enhanced online symptom checker formats on intentions to seek care, knowledge about managing symptoms, and trust and acceptability of the tool. For low acuity symptoms, where advice was to self-care, tailored formats reduced intentions to seek potentially unnecessary professional care. For moderate acuity symptoms, where advice was to see a GP in 24 hours, there was no effect of format on intentions to follow this advice. Across both symptom acuity levels, there was no impact of format on intentions to self-care. More tailored formats improved knowledge for both symptom acuity levels compared to existing formats at immediate follow up, however, these gains were not sustained at two weeks. Formats that included explicit statements about AI did not significantly affect participants' high levels of trust in the advice or acceptability of the tool.

Although several studies have investigated users' intentions to follow symptom checker advice,[11 13-15] few have quantitatively compared the effects of different formats. Our findings suggest that for low acuity symptoms, more personalised information may reduce potentially unnecessary use of health services. Two key design features may have contributed to this effect: the use of actionable plain language advice for managing both current and changing symptoms; and a clearer description of the rationale for managing the symptoms at home. These features address well-documented issues for symptom checkers.[5 10 17 18 20 21]

This is also one of the first studies to explore public responses to the use of generative AI in online symptom checkers. Explicit statements about the use of AI did not significantly impact on trust or acceptability of the tool. This finding is consistent with the broader literature. People are increasingly turning to general purpose AI tools such as ChatGPT for health advice,[7-9] and as people become more competent and familiar with generative AI tools, they are likely to also become more trusting[39] (although there is also evidence that those most familiar are also some of the least trusting).[40] In this study the AI was also integrated into a recognisable national health service (healthdirect), with visual cues such as government logos and URLs, and a statement about clinician review of the content. Both features may have enhanced perceived reliability.[41 42]

The strengths of this study are the use of pre-registered randomised controlled-trial study design to enhance internal validity of the findings, and community involvement to improve acceptability of the tool to people in the community. The study also has several limitations. We observed high intentions across all groups, potentially limiting our capacity to detect differential effects of the formats. The primary outcome (intention) and knowledge were purpose-built items. Though highly relevant to the study, these assessments were not validated, and the items may not reliably and accurately reflect the underlying constructs. This is a common challenge, particularly for performance-based assessment of knowledge. Lastly, the symptom checker formats were presented as screenshots. Participants may have responded differently to interactive versions, particularly those with a stronger emphasis on visual design e.g. the multimedia format.

Further research is needed to improve the external validity and generalisability of the findings. For example, enabling participants to engage with interactive versions of the tool, rather than screenshots, would deliver stronger evidence about the impact of different visual design elements on the outcomes of interest. This study used a hypothetical scenario; allowing participants to respond to real concerns would provide more generalisable findings across diverse health contexts and levels of acuity.

Conclusion

Symptom checkers are widely available worldwide to support people to make informed decisions about when to seek advice from healthcare services and how to manage symptoms at home. This study found that people were better equipped with knowledge to manage symptoms when they viewed online symptom checker formats that gave more tailored advice. There was also some evidence that these more tailored formats may help reduce the potentially unnecessary use of primary care services for symptoms that can be managed safely at home. The explicit use of AI did not impact significantly on trust or acceptability of online symptom checker. Future research should investigate these formats using interactive prototypes across a wider variety of health contexts.

ACKNOWLEDGEMENTS

We would like to acknowledge Ms Momo Hudson-Barton for her support in preparing the figures used in this manuscript. We would also like to acknowledge the contributions of our workshop members Geoffrey Edlund, Ivan CK Ma, Trang Vu, Sharon Ng, Waren Nadesan, Gauri Kapoor, Cheryl Knight, and Lorna Butters. We would also like to acknowledge Rochelle Seneviratna support in managing the project.

FUNDING

Julie Ayre and Kirsten McCaffery are supported by National Health and Medical Research Council fellowships (APP2017278, APP2016719). This work was funded by a research contract awarded to the Sydney Health Literacy Lab, University of Sydney with healthdirect, a government funded entity that provides national evidence-based health information and advice in Australia (#CON230550-Symptom Checker Gen AI Project-healthdirect).

COMPETING INTERESTS

Karen Gallagher, Angelica Scott, Alison Woods, Christopher Ng, and Yuanee Wickramasinghe are employed by healthdirect.

DATA AVAILABILITY

De-identified data produced in the present study can be made available upon request.

REFERENCES

1. Churruca K, Ellis LA, Pope C, et al. The place of digital triage in a complex healthcare system: An interview study with key stakeholders in Australia's national provider. *DIGITAL HEALTH* 2023;9:20552076231181201. doi: 10.1177/20552076231181201
2. Australian Department of Health D, and Ageing. National Health Reform Agreement (NHRA) 2025 [Available from: https://www.health.gov.au/our-work/national-health-reform-agreement-nhra accessed 11 November 2025.
3. Nundy S, Cooper LA, Mate KS. The Quintuple Aim for Health Care Improvement: A New Imperative to Advance Health Equity. *JAMA* 2022;327(6):521-22. doi: 10.1001/jama.2021.25181
4. Chambers D, Cantrell AJ, Johnson M, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open* 2019;9(8):e027743. doi: 10.1136/bmjopen-2018-027743
5. Pairon A, Philips H, Verhoeven V. A scoping review on the use and usefulness of online symptom checkers and triage systems: How to proceed? *Frontiers in Medicine* 2023;Volume 9 - 2022 doi: 10.3389/fmed.2022.1040926
6. Shahsavar Y, Choudhury A. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. *JMIR Hum Factors* 2023;10:e47564. doi: 10.2196/47564 [published Online First: 17.5.2023]
7. Ayre J, Cvejic E, McCaffery KJ. Use of ChatGPT to obtain health information in Australia, 2024: insights from a nationally representative survey. *Medical Journal of Australia* 2025;222(4):210-12. doi: https://doi.org/10.5694/mja2.52598
8. Ayo-Ajibola O, Julien C, Lin ME, et al. Association of Primary Care Access with Health-Related ChatGPT Use: A National Cross-Sectional Survey. *Journal of General Internal Medicine* 2025 doi: 10.1007/s11606-025-09406-9
9. Yun HS, Bickmore T. Online Health Information–Seeking in the Era of Large Language Models: Cross-Sectional Web-Based Survey Study. *J Med Internet Res* 2025;27:e68560. doi: 10.2196/68560 [published Online First: 31.3.2025]

10. You Y, Ma R, Gui X. User Experience of Symptom Checkers: A Systematic Review. *AMIA Annu Symp Proc* 2022;2022:1198-207. [published Online First: 20230429]
11. El-Osta A, Riboli-Sasco E, Al Ammouri M, et al. Factors Influencing the Use of Online Symptom Checkers in the United Kingdom: Cross-Sectional Study. *JMIR Form Res* 2025;9:e65314. doi: 10.2196/65314 [published Online First: 15.9.2025]
12. Meyer AND, Giardina TD, Spitzmueller C, et al. Patient Perspectives on the Usefulness of an Artificial Intelligence–Assisted Symptom Checker: Cross-Sectional Survey Study. *J Med Internet Res* 2020;22(1):e14679. doi: 10.2196/14679 [published Online First: 30.1.2020]
13. Nijland N, Cranen K, Boer H, et al. Patient use and compliance with medical advice delivered by a web-based triage system in primary care. *Journal of Telemedicine and Telecare* 2010;16(1):8-11. doi: 10.1258/jtt.2009.001004
14. Miller S, Gilbert S, Virani V, et al. Patients' Utilization and Perception of an Artificial Intelligence–Based Symptom Assessment and Advice Technology in a British Primary Care Waiting Room: Exploratory Pilot Study. *JMIR Hum Factors* 2020;7(3):e19713. doi: 10.2196/19713 [published Online First: 10.7.2020]
15. Arellano Carmona K, Chittamuru D, Kravitz RL, et al. Health Information Seeking From an Intelligent Web-Based Symptom Checker: Cross-sectional Questionnaire Study. *J Med Internet Res* 2022;24(8):e36322. doi: 10.2196/36322
16. Morse KE, Ostberg NP, Jones VG, et al. Use Characteristics and Triage Acuity of a Digital Symptom Checker in a Large Integrated Health System: Population-Based Descriptive Study. *J Med Internet Res* 2020;22(11):e20549. doi: 10.2196/20549 [published Online First: 30.11.2020]
17. Savolainen K, Kujala S. Testing Two Online Symptom Checkers With Vulnerable Groups: Usability Study to Improve Cognitive Accessibility of eHealth Services. *JMIR Hum Factors* 2024;11:e45275. doi: 10.2196/45275 [published Online First: 8.3.2024]
18. Aboueid S, Meyer S, Wallace JR, et al. Young Adults' Perspectives on the Use of Symptom Checkers for Self-Triage and Self-Diagnosis: Qualitative Study. *JMIR Public Health Surveill* 2021;7(1):e22637. doi: 10.2196/22637

19. Chen X, Hay JL, Waters EA, et al. Health Literacy and Use and Trust in Health Information. *Journal of Health Communication* 2018;23(8):724-34. doi: 10.1080/10810730.2018.1511658
20. Müller R, Klemmt M, Koch R, et al. “That’s just Future Medicine” - a qualitative study on users’ experiences of symptom checker apps. *BMC Medical Ethics* 2024;25(1):17. doi: 10.1186/s12910-024-01011-5
21. You Y, Tsai C-H, Li Y, et al. Beyond Self-diagnosis: How a Chatbot-based Symptom Checker Should Respond. *ACM Trans Comput-Hum Interact* 2023;30(4):Article 64. doi: 10.1145/3589959
22. Matz SC, Teeny JD, Vaid SS, et al. The potential of generative AI for personalized persuasion at scale. *Scientific Reports* 2024;14(1):4692. doi: 10.1038/s41598-024-53755-0
23. Nutbeam D, Milat AJ. Artificial intelligence and public health: prospects, hype and challenges. *Public Health Research and Practice* 2025;35(1):-. doi: https://doi.org/10.1071/PU24001
24. Ugas M, Huynh J, Lenarcik-Packham A, et al. The utility of artificial intelligence in plain language writing: A scoping review. *Patient Education and Counseling* 2026;145:109483. doi: https://doi.org/10.1016/j.pec.2026.109483
25. Dunn AG, Shih I, Ayre J, et al. What generative AI means for trust in health communications. *J Commun Healthc* 2023;16(4):385-88. doi: 10.1080/17538068.2023.2277489 [published Online First: 2023/11/03]
26. Richardson S, Lawrence K, Schoenthaler AM, et al. A framework for digital health equity. *npj Digital Medicine* 2022;5(1):119. doi: 10.1038/s41746-022-00663-0
27. Kopka M, Schmieding ML, Rieger T, et al. Determinants of Laypersons’ Trust in Medical Decision Aids: Randomized Controlled Trial. *JMIR Hum Factors* 2022;9(2):e35219. doi: 10.2196/35219 [published Online First: 3.5.2022]
28. Woodcock C, Mittelstadt B, Busbridge D, et al. The Impact of Explanations on Layperson Trust in Artificial Intelligence–Driven Symptom Checker Apps: Experimental Study. *J Med Internet Res* 2021;23(11):e29386. doi: 10.2196/29386 [published Online First: 3.11.2021]
29. Ng KKY, Matsuba I, Zhang PC. RAG in Health Care: A Novel Framework for Improving Communication and Decision-Making by Addressing LLM Limitations. *Nejm Ai* 2025;2(1):AIra2400380. doi: 10.1056/AIra2400380

30. Alkhalaf M, Yu P, Yin M, et al. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of Biomedical Informatics* 2024;156:104662. doi: https://doi.org/10.1016/j.jbi.2024.104662
31. Wang D, Liang J, Ye J, et al. Enhancement of the Performance of Large Language Models in Diabetes Education through Retrieval-Augmented Generation: Comparative Study. *J Med Internet Res* 2024;26:e58041. doi: 10.2196/58041 [published Online First: 8.11.2024]
32. Wallace LS, Rogers ES, Roskos SE, et al. Brief report: screening items to identify patients with limited health literacy skills. *J Gen Intern Med* 2006;21(8):874-7. doi: 10.1111/j.1525-1497.2006.00532.x [published Online First: 2006/08/03]
33. Norman CD, Skinner HA. eHealth literacy: essential skills for consumer health in a networked world. *Journal of medical Internet research* 2006;8(2):e9. doi: 10.2196/jmir.8.2.e9
34. Mac O, Ayre J, McCaffery K, et al. The Readability Study: A Randomised Trial of Health Information Written at Different Grade Reading Levels. *Journal of General Internal Medicine* 2024 doi: 10.1007/s11606-024-09200-z
35. Komiak SY, Benbasat I. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS quarterly* 2006;30(4):941-60.
36. Jin E, Ryoo Y, Kim W, et al. Bridging the health literacy gap through AI chatbot design: the impact of gender and doctor cues on chatbot trust and acceptance. *Internet Research* 2025;35(3):1299-329. doi: 10.1108/INTR-08-2023-0702
37. Dolan RJ. Emotion, cognition, and behavior. *Science* 2002;298(5596):1191-4. doi: 10.1126/science.1076358
38. Maher JM, Markey JC, Ebert-May D. The other half of the story: effect size analysis in quantitative research. *CBE—Life Sciences Education* 2013;12(3):345-51.
39. Lāma G, Lastovska A. AI competence and sentiment: a mixed-methods study of attitudes and open-ended reflections. *Frontiers in Artificial Intelligence* 2025;Volume 8 - 2025 doi: 10.3389/frai.2025.1658791
40. Kauttonen J, Rousi R, Alamäki A. Trust and Acceptance Challenges in the Adoption of AI Applications in Health Care: Quantitative Survey Analysis. *J*

Med Internet Res 2025;27:e65567. doi: 10.2196/65567 [published Online First: 21.3.2025]

41. Sun Y, Zhang Y, Gwizdka J, et al. Consumer Evaluation of the Quality of Online Health Information: Systematic Literature Review of Relevant Criteria and Indicators. *J Med Internet Res* 2019;21(5):e12522. doi: 10.2196/12522 [published Online First: 02.05.2019]
42. Ryan K, Hogg J, Kasun M, et al. Users' Perceptions and Trust in AI in Direct-to-Consumer mHealth: Qualitative Interview Study. *JMIR Mhealth Uhealth* 2025;13:e64715. doi: 10.2196/64715 [published Online First: 20.5.2025]

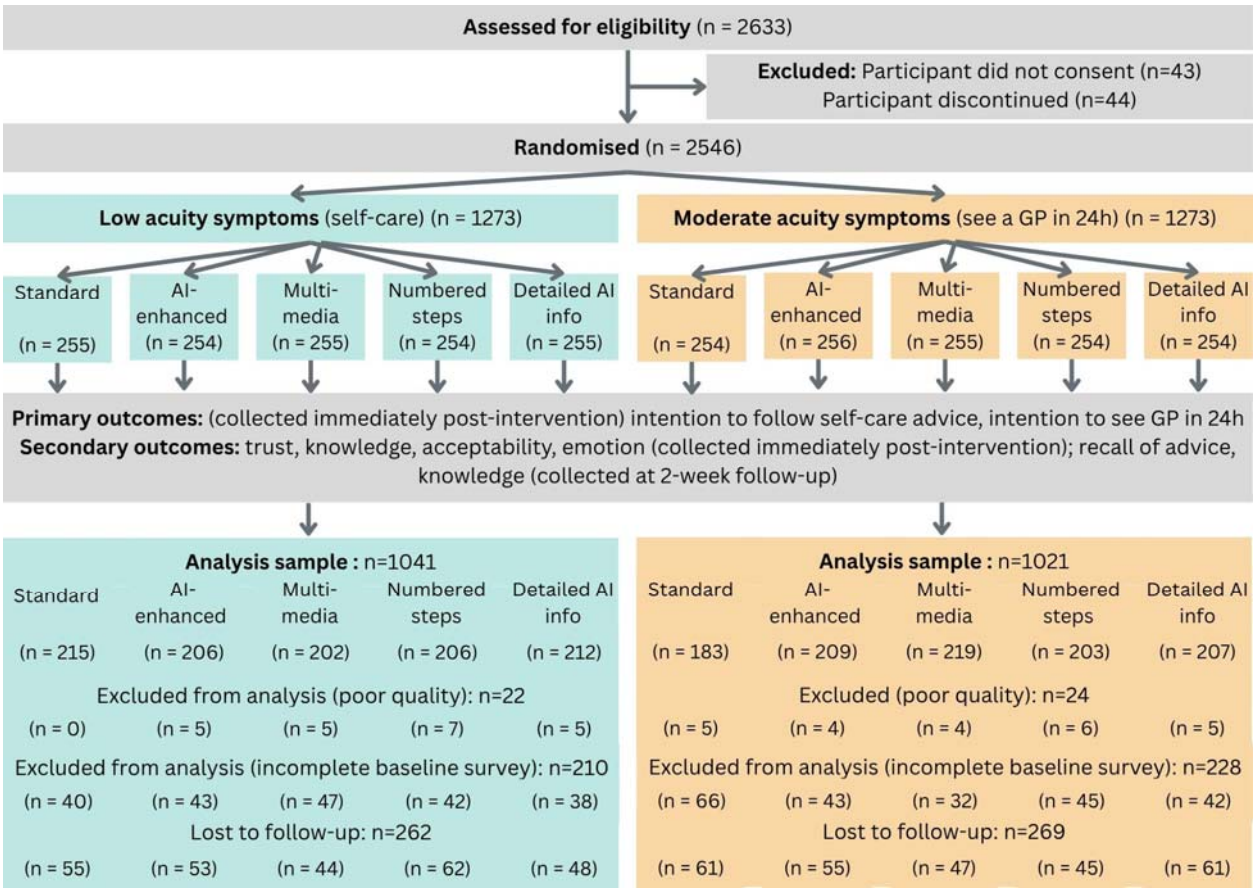

Figure 1. Study flow diagram

Note. Participants were excluded from the analysis if they were considered ‘speeders’ i.e. completed the survey in less than 5 minutes.

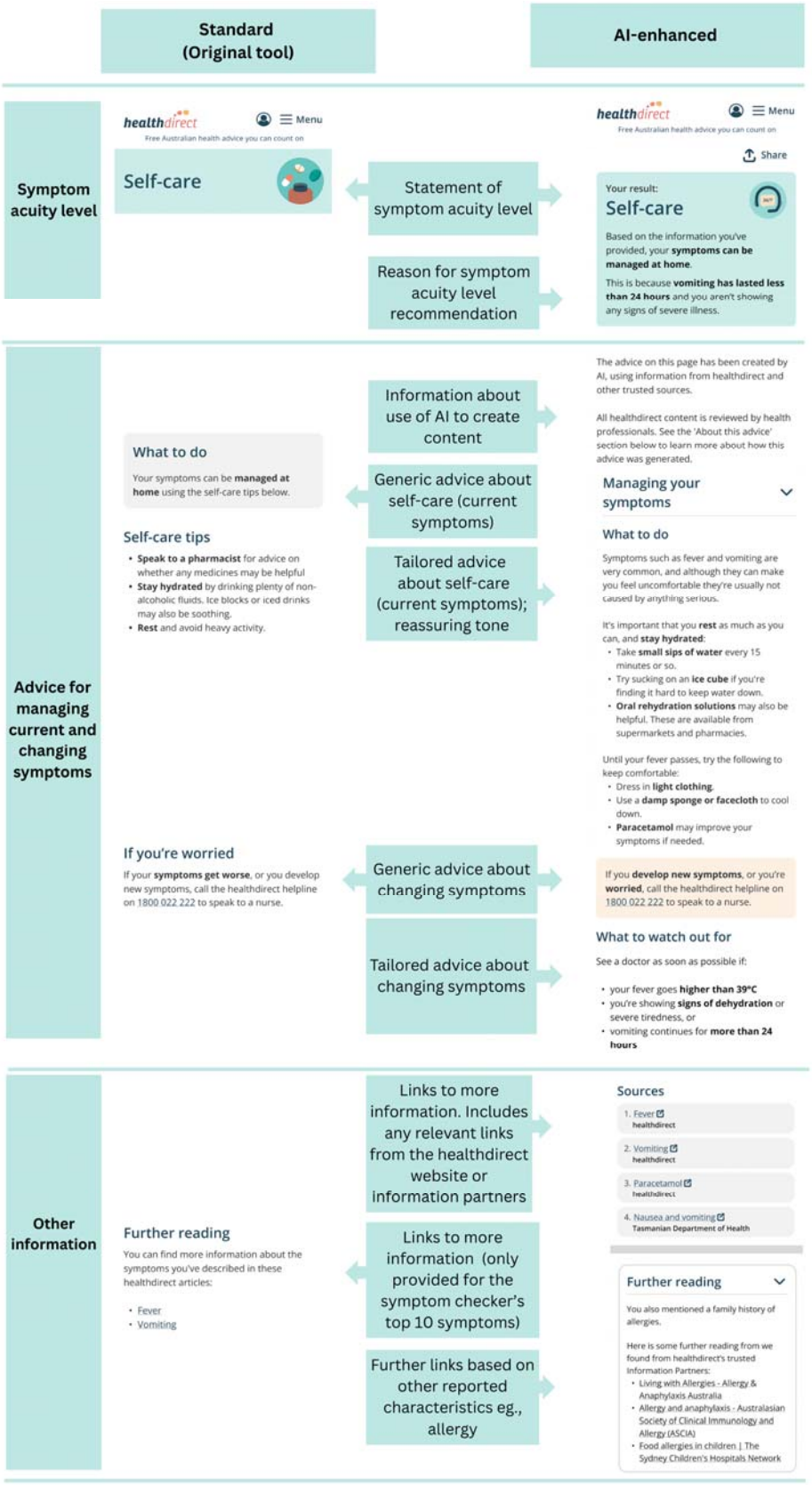

Figure 2. Comparison of standard and AI-enhanced groups, low symptom acuity level (self-care)

Panel A

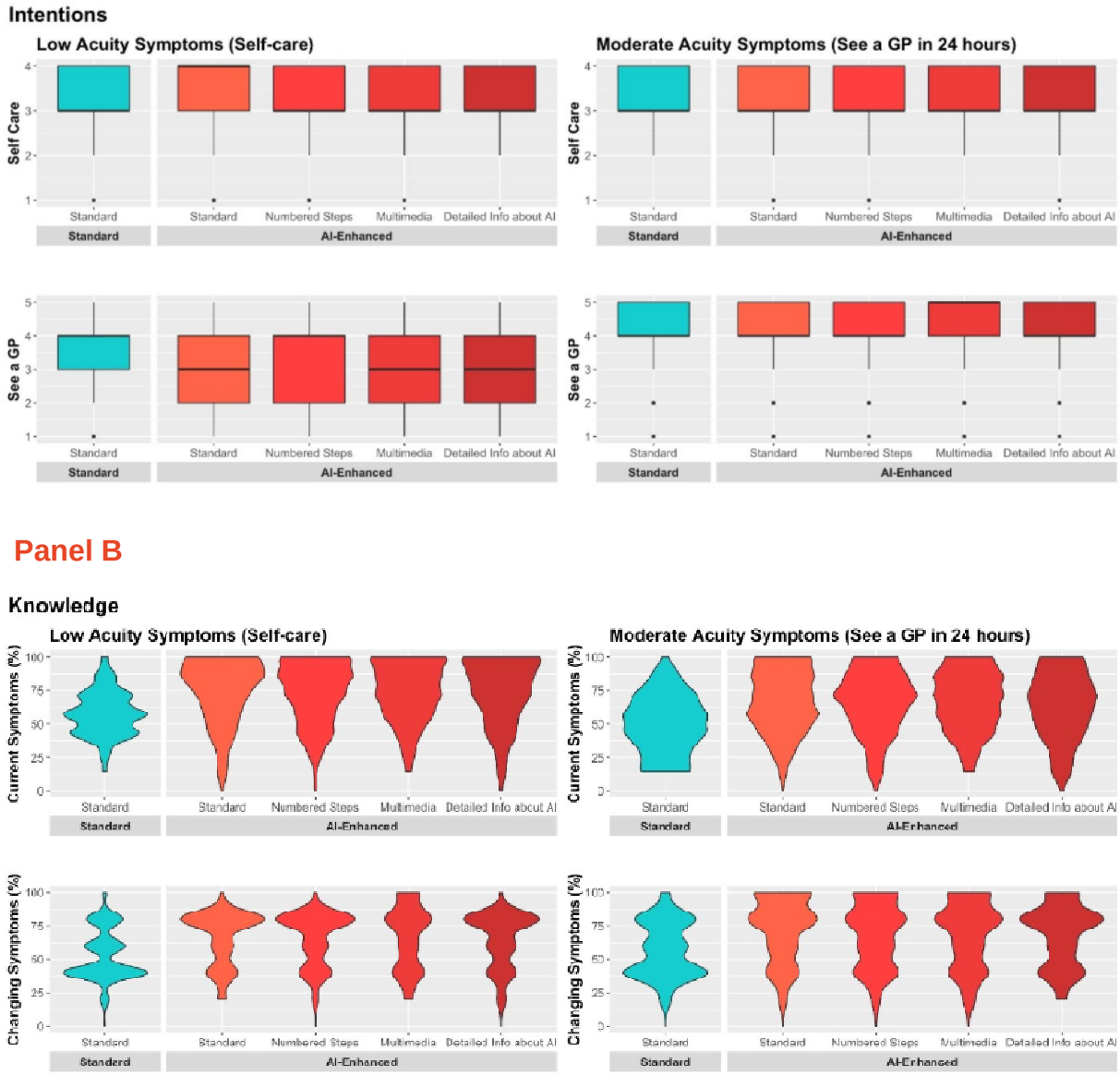

Figure 3. Distribution of intention and knowledge scores (immediately post-intervention), by intervention group and symptom acuity level.