

# **Manage Your Life Online a Logic-Based Chatbot: Case-Series Examining Effectiveness and User Experience After the Release of Chat-GPT**

Aimee-Rose Wrightson-Hester, Gee Anderson, Joel Dunstan, Peter M McEvoy, Christopher J Sutton, Bronwyn Myers, Sarah Egan, Sara Tai, Melanie Johnston-Hollitt, Wai Chen, Tom Gedeon, Isabeau K Tindall, Joanna C Moullin, Warren Mansell

Submitted to: Journal of Medical Internet Research  
on: December 11, 2024

**Disclaimer:** © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 44

    Figures ..... 45

        Figure 1..... 46

        Figure 2..... 47

    Multimedia Appendixes ..... 48

        Multimedia Appendix 1..... 49

# Manage Your Life Online a Logic-Based Chatbot: Case-Series Examining Effectiveness and User Experience After the Release of Chat-GPT

Aimee-Rose Wrightson-Hester<sup>1,2</sup> PhD; Gee Anderson<sup>1</sup>; Joel Dunstan<sup>3</sup> MIT; Peter M McEvoy<sup>1,2,4</sup> PhD; Christopher J Sutton<sup>5</sup> PhD; Bronwyn Myers<sup>1,6,6</sup> PhD; Sarah Egan<sup>1,2</sup> PhD; Sara Tai<sup>7</sup> DClinPsy; Melanie Johnston-Hollitt<sup>3</sup> PhD; Wai Chen<sup>1,6,6,6</sup> PhD; Tom Gedeon<sup>8</sup> PhD; Isabeau K Tindall<sup>1,2</sup> PhD; Joanna C Moullin<sup>1,9</sup> PhD; Warren Mansell<sup>1,2,7</sup> DClinPsy

<sup>1</sup>Curtin enAble Institute, Faculty of Health Sciences, Curtin University Perth AU

<sup>2</sup>Discipline of Psychology, School of Population Health, Curtin University Perth AU

<sup>3</sup>Curtin Institute for Data Science, Curtin University Perth AU

<sup>4</sup>Centre for Clinical Interventions, North Metropolitan Health Service Nedlands AU

<sup>5</sup>Centre for Biostatistics, School of Health Sciences, The University of Manchester Manchester GB

<sup>6</sup>Department of Clinical Psychology, School of Health Sciences, The University of Manchester Manchester GB

<sup>7</sup>School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University Perth AU

<sup>8</sup>School of Pharmacy, Faculty of Health Sciences, Curtin University Perth AU

## Corresponding Author:

Aimee-Rose Wrightson-Hester PhD

Curtin enAble Institute, Faculty of Health Sciences, Curtin University

Kent Street

Perth

AU

## Abstract

**Background:** There is a shortage of services available to address the growing demand for mental health support in Australia and worldwide. Digital interventions, including conversational agents, can overcome barriers to accessing mental health support. The recent advances in large language models (LLMs) have led to an improvement in the perceived human-like naturalness of chatbot conversations, but there is little research on the experience and efficacy of chatbots to support mental health. Manage Your Life Online (MYLO) is a rule-based chatbot that was co-designed with young people that uses questions to help users explore their problems. In a case series conducted prior to the release of ChatGPT, users rated a new smartphone interface for MYLO acceptable and results demonstrated a reduction in problem-related distress.

**Objective:** To assess a later version of MYLO's impact on target outcomes over a four-week period in a sample of young people with lived experience of anxiety and/or depression. We hypothesized that these results would replicate the previous case-series for problem-related distress, anxiety, psychiatric impairment and goal conflict reorganization. We also anticipated the longer usage time would lead to improvements on general health, depression and self-efficacy. We also aimed to compare the user experience of MYLO in this case-series to the previous version that was completed in November 2022.

**Methods:** We replicated and extended the previous two-week case-series, conducted in September to November 2022, by testing four-week usage of MYLO, in October to December 2023, with a larger sample. To do this we recruited 24 young people living in Western Australia who self-described as having lived experience of anxiety and/or depression. Participants had access to and used MYLO over a four-week period while completing online weekly surveys that included a range of health and psychological questionnaires. After the four-week testing phase participants were invited to attend either an interview or focus group to provide feedback on their experience of using MYLO.

**Results:** We found improvements in problem-related distress ( $d = -1.07$ ), anxiety ( $d = -0.41$ ) and psychiatric impairment ( $d = 0.60$ ) and some evidence of reliable improvement in clinical outcomes. However, satisfaction with MYLO conversations was rated more poorly compared to the previous study. In qualitative interviews, participants spoke about their experiences with ChatGPT (released in November 2022 after the previous case-series concluded) and other generative AI tools, stating that they had expected MYLO to possess similar functionality.

**Conclusions:** These findings have implications for mental health chatbots in the age of ChatGPT and highlight a need for

researchers to engage with new technologies to improve user experience, while maintaining necessary safety and ethical standards.

(JMIR Preprints 11/12/2024:69841)

DOI: <https://doi.org/10.2196/preprints.69841>

## Preprint Settings

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

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

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

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org>

## Original Manuscript

# Manage Your Life Online a Logic-Based Chatbot: Case-Series Examining Effectiveness and User Experience After the Release of Chat-GPT

## Introduction

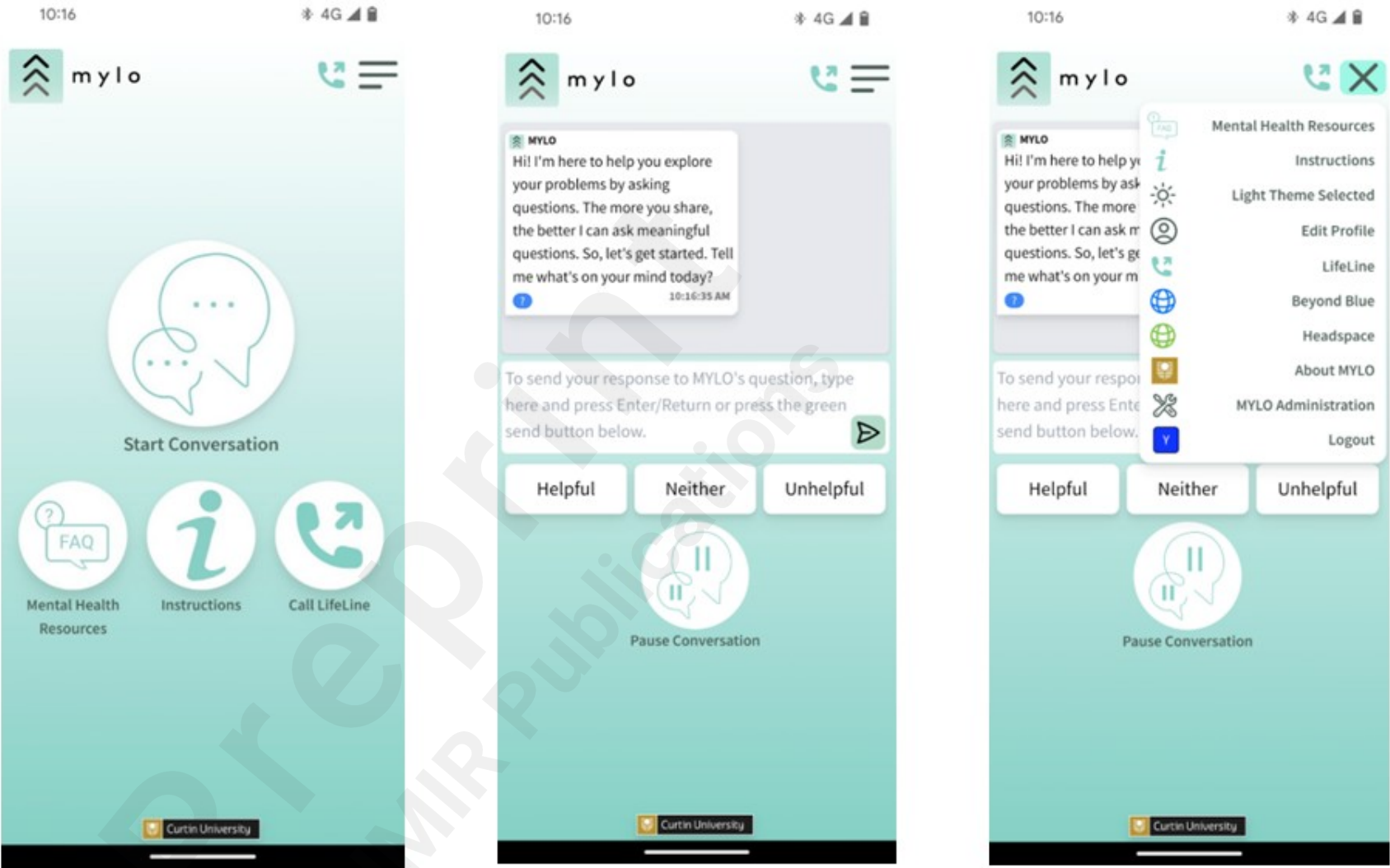
To address the growing need for mental health care [1], researchers and policymakers are investigating the use of digital and artificial intelligence solutions to improve scalability of interventions [2]. Digital interventions might be particularly helpful for young people who have grown up with the internet and who experience barriers to accessing appropriate mental health care [3]. Conversational agents or chatbots are one technology that people are using to support their mental health. Mental health conversational agents are specifically designed to support user's mental health by emulating various forms of traditional psychotherapy [4]. Studies have found high user satisfaction with chatbots [5,6], with users enjoying the interactive approach and building a relationship with the chatbots like that of a human therapist or friend [7,8]. Although feedback is generally positive, many studies report negative feedback from users about mental health chatbots, such as repetitive content, not feeling understood [9], inappropriate responses from the bot [7] and technical issues [10]. In late 2022, highly naturalistic conversations with chatbots became widely accessible with the release of generative artificial intelligent agents such as ChatGPT [11]. Since ChatGPT's release, the use of AI technology has seen substantial growth driven by improvements in natural language processing and the integration of AI technologies into day-to-day personal and business applications [12].

This paper describes the assessment of Manage Your Life Online (MYLO), a progressive web-application conversational agent that uses rule-based artificial intelligence to exchange messages with the user. MYLO aims to emulate a conversation like that of a Method of Levels therapist [13]. Method of Levels therapists help clients explore internal conflicts between their current state and their ideal state [14]. It is these conflicts that lead to psychological distress according to Perceptual Control Theory which underpins Method of Levels therapy [6,13-15].

MYLO responds to user input in real-time with curious questions tailored to the user's current problem. These questions aim to help the user sustain their awareness of the conflicts they are experiencing and explore the root cause of their distress [16].

The current MYLO interface (see Figure 1) was co-designed with young people to create a conversational agent that is accessible and appealing to use [17]. The acceptability of the current interface was evaluated by young people in a case-series in which users tested MYLO for two weeks and answered weekly online surveys [17]. Participants rated MYLO as acceptable on the System Usability Scale (mean = 73.57) and reported the interface was easy to use and that they liked its functionality in interviews and focus groups. Six of the eight participants who attended an interview or focus group said they would recommend MYLO to their friends. Further, reliable change scores [18] were calculated for each participant that completed baseline and post-testing (n=11) and seven participants showed reliable improvements in their problem-related distress or other clinical outcome (such as depression, anxiety and psychiatric impairment) scores. Due to the modest sample size and short testing phase (2 weeks), no formal statistical testing was performed.

Following the previous case-series, the MYLO database underwent modest updates based on user feedback and guided by a new youth advisory panel. Specifically, new questions, themes and terms were added to assist users with a wider range of problems. Two themes to identify risk of suicide or non-suicidal self-harm were also added, that when triggered responded to users by reminding them of the local mental health crisis hotline accessible through the MYLO interface. During this development phase, the ability to decrypt user conversations was also added. Analyzing user conversations will allow researchers to identify the helpful and unhelpful elements of MYLOs conversations, building on the work of Gaffney et al. [6,16]. While members of the youth advisory panel shared sections of their conversations to discuss with the research team, it was unclear whether research participants would consent to share their conversations during a research trial.





In the current study, we, therefore, aimed to extend the previous case-series by testing MYLO for a longer period of time, to examine the impact of MYLO on users' wellbeing and to test the changes made to the methodology (see Method section for full details) prior to conducting a full-scale randomized controlled trial. We also assessed whether any aversive events during the study period could be attributed to MYLO, again to establish an indication of safety ahead of a trial. Specifically, we hypothesized similar to the previous case-series:

H1. After using MYLO for four weeks, participants' problem-related distress, anxiety and psychiatric impairment will decrease.

H2. After using MYLO for four weeks, participants' ability to engage in goal conflict reorganization will improve.

We also hoped to observe changes in participants' clinical outcomes that did not change during the shorter case-series, i.e., general health, depression and self-efficacy. Specifically, we hypothesized that:

H3. After using MYLO for four weeks, participants' general health, depression and self-efficacy would improve.

We also assessed the feasibility of conversation decryption, and our methodology for gaining consent for this procedure. We did not have a target number of conversations or participants. We will report the method used and then the number of participants and conversations obtained. Finally, we explored the user experience of MYLO using the Session Impact Scale, the System Usability Scale, and qualitative interviews and focus groups. We compare our findings to the previous case-series through the magnitude of the effect sizes rather than inferential statistics.

## Methods

### Design

The case-series protocol was similar to Wrightson-Hester et al. [17] which was

conducted from September to November 2022. In the current study the intervention duration was extended to four weeks, and participants completed this between October and December 2023, rather than the two weeks in the previous study. Therefore, a brief method will be described here but for full details, see Wrightson-Hester et al. [17]. One major difference was the procedure used to gain consent for the decryption and analysis of conversations which will be described in full in the procedure. To measure any adverse impacts of using MYLO we also added an adverse events survey post-testing.

## **Ethics Approval**

Approval for the research was obtained from the Curtin University Human Research Ethics Committee (HREC2022-0466).

## **Participants**

The study was advertised on social media via Metaverse, X, and LinkedIn and shared by members of the research group through their existing networks and personal social media pages. The inclusion criteria were (1) aged 16 to 24years, (2) currently living in Western Australia, (3) a self-reported experience of anxiety or depression (either current or past), (4) access to a smartphone and the internet, and (5) ability to read and type in English. The exclusion criteria were (1) scored greater than 20 on the PHQ-9, the threshold for severe depressive symptoms, or (2) reported experiencing frequent suicidal thoughts (by scoring 2 or more on the PHQ-9 item 9) [19]. Like Wrightson-Hester et al., [17] we also aimed to recruit a diverse range of young people including a range of genders, sexualities, cultures and mix of people who lived in metropolitan and rural Western Australia. Participants expressed their interest in participating through an online survey via Qualtrics (Qualtrics International Inc) and completed questions to assess their eligibility and demographic information and gave informed consent to participate along with their contact details.

## Materials

### *Web-Based Survey*

Participants completed anonymous online surveys hosted by Qualtrics at baseline, then after each week of testing for a total of four weeks, between October and December 2023. A subject-generated identification code [20] was used to link participants' surveys across time points. The self-report questionnaires included in the online survey are presented in Table 1.

**Table 1.** The questionnaires used in the case series.

Questionnaire	Measures	Scoring
Patient health questionnaire-9 [19]	9 items; depression	0-4: minimal depression, 5-9: mild depression, 10-14: moderate depression, 15-19: moderately severe depression, and 20-27: severe depression.
Generalized anxiety disorder assessment-7 [21]	7 items; anxiety	0-4: minimal anxiety, 5-9: mild anxiety, 10-14: moderate anxiety, and 15-21: severe anxiety.
General health questionnaire-12 [22]	12 items; psychiatric impairment	Traditional (acute) scoring method used. Scores range from 0 to 12, and higher scores indicate a greater possibility of psychological distress.
Short Form-6D version 2 [23]	6 items; general health	Scores range from -0.685 to 1, with 1 indicating perfect health. Australian weights were used for this sample.
Psychological outcome profiles [24]	4 items used for scoring; change in problem-related distress over the course of therapy	Scores range from 0 to 20. Decreases in score between pretherapy and posttherapy indicate that a positive change has occurred.
Reorganization of conflict scale [25]	10-item subscale; goal conflict awareness and the proposed mechanism of change in the method of levels therapy	Each item is scored from 0 (I do not believe this at all) to 100 (I believe this completely). The mean of the 10 items is used as the outcome.
General self-efficacy scale [26]	10 items; self-efficacy	Scores range from 10 to 40. Higher scores indicate higher perceived general self-efficacy.
Session impact scale <sup>a</sup> [27]	17 items; session (therapeutic) satisfaction	Each item is scored from 1 (not at all) to 5 (very much). We calculated the mean scores for the unwanted thoughts, relationship impacts, hindering impacts, understanding, and problem-solving subscales. Item 17 measures “other impacts,” an optional item that is not used in scoring.
System usability scale <sup>a</sup> [28]	10 items; user experience of digital systems	Outcome is a percentile ranking from 0 to 100, with scores >68 considered above average.
Adverse event survey <sup>b</sup>	26 items; adverse events	The items have been adapted from the FOCUS clinical trial (Morrison et al., 2018). Each item is scored from 0 (Not at all) to 4 (Very much), participants can also select “Not applicable”.

<sup>a</sup>Denotes questionnaires that were only used after the baseline survey.

<sup>b</sup>This survey was included in the post-testing (week 4) survey and also sent to non-completers

separately. Please note this table is adapted from Wrightson-Hester et al. [17].

### ***Manage Your Life Online (MYLO)***

Participants accessed MYLO as a progressive web application through a link provided by the research team. MYLO's interface allows users to engage in a text-to-text conversation, access other mental health resources (including a local suicide crisis center hotline), pause and resume conversations, and alter the colors used in MYLO and their own avatars (colored squares with the user's initial). MYLO asks users curious questions in response to free-text provided by the user, for example if a user states they feel nervous MYLO might ask "What do you do when you feel nervous?". The question that MYLO presents is determined by an algorithm that picks the best question based on key terms identified in the user's text and previous ratings of the question and term pairings. The questions aim to sustain the user's awareness of their problems, explore them in new ways, and ultimately, resolve them through shifts in their thinking or actions to resolve them.

### **Procedure**

Participants accessed the EOI survey via links or QR codes on study adverts. Participants completed questions to access their eligibility, completed the PHQ-9 and provided their contact details. Eligible participants from the EOI survey were then contacted by the research team and invited to complete the baseline survey via email or text. Those who completed the baseline survey were then provided with a MYLO account (via Auth0) and sent the link and instructions to download MYLO. Participants then had access to and used MYLO for four weeks and completed weekly online surveys, between October and December 2023. After the testing phase, participants were invited to attend an online focus group or an individual interview to provide further feedback on their time using MYLO. All focus groups and interviews were conducted in December 2023. The same topic guide was used to guide the interviews as was used in the previous case-series (see [17]). Participants were reimbursed for their time completing the online surveys, as well as the focus group or

interview with a digital gift card at a rate of \$20 AUD per hour with a maximum commitment of five hours if all surveys and an interview or focus group was completed (total of \$100 AUD).

At the end of the interview or focus group, participants were informed of the research team's intent to ask for their consent to decrypt and analyze the conversations they had with MYLO during the testing phase. After the interview or focus group, case-series participants were emailed an information sheet and consent form for the decryption of their conversations. Participants were able to ask any questions, and willing participants returned their consent forms via email. Emails with the information sheet and consent form were also sent to all other participants who did not attend an interview or focus group or complete all the surveys when they were sent their gift card for taking part in the study. In the same email non-completers were also provided a link to complete the optional adverse events survey.

## Data Analysis

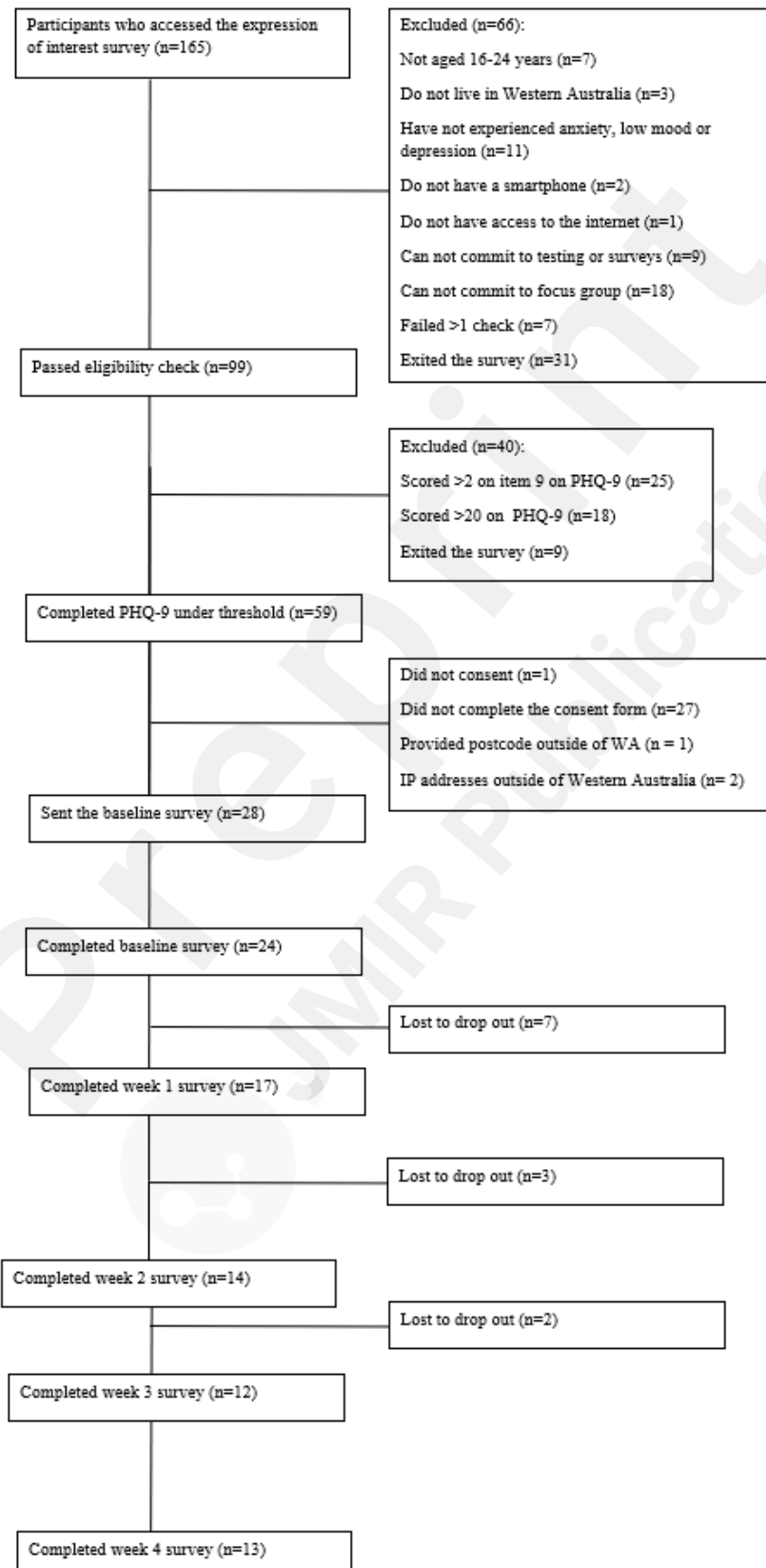
Due to the small number of participants, statistical testing was not conducted, and instead descriptive statistics including (standardized) effect size estimates, with 95% confidence intervals are presented to assess MYLO's impact on user wellbeing and experience. We also calculated whether any participants achieved a reliable improvement or deterioration [18] during the testing phase on each of the quantitative outcome measures described previously. To further assess users' experience of MYLO, interviews and focus groups were conducted, transcribed and analyzed by the first author. Analysis was conducted according to Vears and Gillam's [29] inductive content analysis procedure.

## Results

### Participants

Figure 2 shows how many participants accessed the EOI survey and how many were lost at each stage of the survey. This process left 28 eligible participants who were contacted

to participate in the study. Of these, 24 completed the baseline survey. The demographics of the participants are summarized in Table 2 and met our diversity targets. Of the 24 participants, 17 completed the week 1 survey, 14 completed week 2, 12 completed week 3 and 13 (54.2%) completed week 4. The 13<sup>th</sup> respondent in week 4 was unable to complete the week 3 survey. Focus groups and interviews were conducted with 13 participants. Eight (seven completers, one non-completer) participants completed and consented to the analysis of the adverse events survey.

**Figure 2.** Participant flowchart through the study from EOI to post-testing survey.



**Table 2.** Participant demographics at baseline.

	<b>Total Sample (n=24)</b>	<b>Completers (n=13)</b>
<b>Age</b>	<b>n (%)</b>	<b>n (%)</b>
16-17	12 (50%)	9 (69%)
18-19	5 (21%)	1 (8%)
20-21	2 (8%)	0 (0%)
22-24	5 (21%)	3 (23%)
<b>Gender</b>		
Woman	13 (54%)	8 (62%)
Man	7 (29%)	3 (23%)
Non-binary or self-described <sup>a</sup>	4 (17%)	2 (15%)
<b>Sexuality</b>		
Heterosexual	9 (37%)	3 (23%)
Non-heterosexual	11 (46%)	6 (46%)
Did not disclose	4 (17%)	4 (31%)
<b>Culture</b>		
Australian	14 (58%)	9 (69%)
Australian and other <sup>b</sup>	6 (25%)	3 (23%)
Other <sup>c</sup>	4 (17%)	1 (8%)
<b>Metropolitan or Regional area</b>		
Metropolitan	19 (79%)	10 (77%)
Regional or Remote	5 (21%)	3 (23%)

<sup>a</sup>Three participants described themselves as non-binary and one as a transgender male.

<sup>b</sup>Australian and other (self-described): Chinese, Dutch and Sri Lankan, English, Italian, Hispanic, and Maori- New Zealander.

<sup>c</sup>Other (self-described): Afrikaans South African, Burmese, and Iranian.

These are only examples of possible headings. Please feel free to use different headings to best describe your results.

## Usage Data

Table 3 presents the MYLO usage data. Participants usage data was collected in the MYLO app while their survey responses were collected anonymously via an external host (Qualtrics). Therefore, it was not possible to link participant outcomes to their MYLO usage in this phase, and the usage data reported is for all conversations had by all participants who downloaded MYLO and had at least one conversation during the trial period (n = 22). In addition to the results in Table 3, when users rated their MYLO conversations, the mean

number of messages they sent in conversations rated as helpful ( $M = 15.6$  user messages), and neither helpful or unhelpful ( $M = 15$  user messages) were higher than in conversations rated as unhelpful ( $M = 9.5$  user messages).

**Table 3. MYLO Usage data**

	<b>n (%)</b>
Number of users	22
Total Conversations	90
Mean conversations per user	4.1
<b>Number of messages from user per conversation</b>	
Mean	11.54
SD	11.10
Median	8.5
Range	1-61
<5 messages	25 (28)
5 –10 messages	26 (29)
11-20 messages	29 (32)
21+ messages	10 (11)
<b>Conversation ratings</b>	
Helpful	14 (16%)
Neither	16 (18%)
Unhelpful	37 (41%)
Unrated	23 (26%)

## Conversation Decryption and Consent Acceptability

All 24 participants who completed the baseline survey were sent an information sheet and consent form via email to allow the research team to decrypt and analyze the conversations they had with MYLO during the project. Of the 24 participants, four consented to their conversations being analyzed, one responded saying they did not want their conversations decrypted and analyzed, and 18 participants did not respond to the email. One participant let the research team know they would complete their consent form via email but then did not contact the team again. From these participants we were able to collect 14 conversations which will be analyzed and presented in a future paper.

## Adverse Events

Eight participants completed the adverse events survey. All items were scored 0 (Not

at all), 1 (Very little), 2 (A little), 3 (Quite a lot) or 4 (Very much), “Not Applicable” responses were not scored or included in calculations. A total of 25 items in the survey describe adverse events and an average score of each was calculated for each participant. These scores ranged from 0.04 to 2.48,  $M = 0.81$   $SD = 0.78$ . Only two participants achieved a mean of over 1 on the scale. Nine items had mean scores between 1 and 2. Three of these items related to the time and effort required to participate. One item scored 2.38 and that was “Taking part hasn't helped me with my problems”. Item 26 is a positive item and asks participants agreement with the statement “My problems have improved to the point whereby I no longer feel I need help” and this had a mean score of 2. Participants were also able to provide qualitative feedback on the adverse events survey, and four participants chose to do so. Three participants left positive comments about MYLO, one of whom chose to explain their responses on the adverse event survey:

“I had a very easy time figuring out how MYLO worked, but midway had a bout of depression that was unrelated to using MYLO, which made it very difficult for me to find the motivation to use the app.”

“Doing the MYLO trial helped me have a better understanding of mental health and how it can affect you and the people around you. It definitely made me have a larger understanding of how much impact I can have on all the daily activities you do and your thoughts and mindset. I think that MYLO was a very well-created system, and I hope that our community can also enjoy and learn from it.”

“I think the MYLO experience was overall good. I can't wait to see the following phases of MYLO as I can see this AI has the basic structure to become something that is going to be very helpful towards a large population of people.”

One participant left a negative response:

“During times of mild anxiety/dissociation/depression, etc., I tried to use MYLO to

help support me instead of my usual support system. MYLO was very unhelpful and made the problems worse due to its inability to provide any advice, practical help, or even have a proper conversation about the topic of mental health issues. Because of this I found that I needed to use other support systems to not only help with the original problem but counterbalance the lack of support that I felt from MYLO.”

## Wellbeing Outcomes

Table 4 presents the mean scores for each self-report questionnaire at each time point as well as the change from baseline to week 4. Table 5 presents how many participants achieved reliable improvement or deterioration from baseline to post-testing for all outcomes.

- H1: After using MYLO for four weeks participants’ problem-related distress, anxiety and psychiatric impairment will decrease.

We found a large effect size for problem-related distress and a mean change of 3.77 from baseline to week 4 ( $d = -1.07$ ). Anxiety scores showed a mean reduction of 2.08 ( $d = -0.41$ ) indicating a small-to-moderate reduction from baseline to week 4. A medium effect size was found for psychiatric impairment, with the mean score post-testing dropping below the case threshold (i.e., clinical threshold for psychiatric impairment) of 3 [30] to 2.46 ( $d = 0.60$ ). Of the 13 participants who completed both the baseline and post-testing (week 4), 11 individuals scored 3 or more at baseline and six scored 3 or more post-testing (see supplementary materials). The reliable change scores show that, on the measures of problem-related distress, six participants achieved a reliable improvement, and none had a reliable deterioration. For anxiety, two participants achieved a reliable improvement, and none had a reliable deterioration. For psychiatric impairment, six achieved a reliable improvement and one had a reliable deterioration.

- H2: After using MYLO for four weeks, participants' ability to engage in goal conflict reorganization would improve.

We found no effect on goal conflict reorganization with a mean change of 1.00 from baseline to week 4 ( $d = -0.06$ ). The reliable change scores show that two participants achieved a reliable improvement in their ability to reorganize goal conflicts, and three reliably deteriorated.

- H3: After using MYLO for four weeks, participants' general health, depression and self-efficacy would improve.

Participants showed a small-to-moderate effect size improvement in general health between baseline and post-testing ( $d = 0.44$ , mean change = 0.04). No participants experienced a reliable change in general health. We found a small-to-moderate effect size on participants' depression scores between baseline and post-testing ( $d = 0.47$ , mean change = 1.85). We found no effect on self-efficacy ( $d = 0.07$ , mean change = -0.31). The reliable change scores show that no participants experienced a reliable change in their general health, two participants achieved reliable improvements in their depression scores, one participant reliably deteriorated, three participants achieved reliable improvements in their self-efficacy scores, and one reliably deteriorated between baseline and post-testing.

**Table 4.** Mean scores on clinical outcomes at baseline, week 1, week 2, week 3, and week.

	Baseline	Week 1	Week 2	Week 3	Week 4	Change <sup>a</sup> , mean (SD)	95% CI	$d^b$
	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>			
N	24	17	14	12	13			
General health	0.56 (0.25)	0.69 (0.19)	0.71 (0.17)	0.65 (0.36)	0.71 (0.22)	0.04 (0.09)	-0.02 to 0.09	0.44
Depression	12.54 (5.45)	10.35 (4.18)	8.71 (4.34)	9.00 (5.31)	9.23 (5.89)	-1.85 (3.96)	-4.24 to 0.54	-0.47
Anxiety	9.58 (4.68)	8.82 (4.67)	8.00 (3.72)	8.58 (4.32)	7.08 (4.03)	-2.08 (5.02)	-5.11 to 0.96	-0.41
Psychiatric impairment	5.63 (2.58)	3.06 (2.73)	2.57 (2.85)	2.92 (2.81)	2.46 (2.96)	-2.31 (3.86)	-4.64 to 0.03	-0.60
Goal conflict	66.81	63.53	62.21	64.08	64.92	-1.00 (16.50)	-10.97 to	-0.06

reorganization	(13.12)	(10.40)	(13.38)	(15.20)	(15.06)		8.97	
Self-efficacy	28.71 (4.75)	28.53 (4.16)	28.93 (5.37)	28.17 (4.39)	29.46 (4.39)	0.31 (4.13)	-2.19 to 2.80	0.07
Problem-related distress	13.96(2. 71)	10.53 (2.21)	10.14 (2.96)	10.75 (2.99)	9.00 (3.46)	-3.77 (3.52)	-5.89 to - 1.65	-1.07

<sup>a</sup>The change column presents mean change between baseline and week 4; only scores of those who completed the week 4 survey are included (n=13).

<sup>b</sup> Cohen's *d* was calculated using Jamovi for each outcome post-testing (end of week 4) relative to baseline [31].

**Table 5.** Participants’ change scores from baseline to post-testing for all target outcomes.

			Participant												
	$\alpha^d$	Index <sup>c</sup>	1	2	3	4	5	6	7	8	9	10	11	12	13
General Health	0.82	0.26	-0.010	0.111	0.053	0.038	-0.126	0.000	0.029	-0.037	-0.005	0.165	0.251	0.014	0.054
Depression	0.86	5.78	-1	-4	-7 <sup>a</sup>	-1	3	-5	2	6 <sup>b</sup>	-1	-1	-8 <sup>a</sup>	-3	-4
Anxiety	0.85	5.72	1	-2	-4	-2	1	-4	4	3	-1	-8 <sup>a</sup>	-15 <sup>a</sup>	1	-1
Psychiatric Impairment	0.78	3.44	-2	-1	-4 <sup>a</sup>	3	4 <sup>b</sup>	-7 <sup>a</sup>	3	0	-5 <sup>a</sup>	-5 <sup>a</sup>	-7 <sup>a</sup>	-3	-6 <sup>a</sup>
Goal Conflict Reorganization	0.81	13.90	-14 <sup>b</sup>	17 <sup>a</sup>	30 <sup>a</sup>	11	-33 <sup>b</sup>	-5	-5	5	-12	2	6	5	-20 <sup>b</sup>
Self-Efficacy	0.86	4.57	-1	1	7 <sup>a</sup>	-1	-8 <sup>b</sup>	-2	-2	5 <sup>a</sup>	1	6 <sup>a</sup>	2	0	-4
Problem-Related Distress	0.72	3.62	-4 <sup>a</sup>	0	-12 <sup>a</sup>	1	-7 <sup>a</sup>	-1	-2	-2	-3	-2	-6 <sup>a</sup>	-7 <sup>a</sup>	-4 <sup>a</sup>

<sup>a</sup> Denotes participant who achieved reliable improvement.  
<sup>b</sup> Denotes participant who achieved reliable deterioration.  
<sup>c</sup> Index = Reliable Change Index [18].  
<sup>d</sup> Cronbach’s  $\alpha$ .

## User Experience

In the current and previous case-series [17], participants rated MYLO's usability using the System Usability Scale (SUS) [28] every week during the testing phase (over 4 weeks and 2 weeks, respectively). In the previous case-series, MYLO's SUS ratings ranged from 37.5-97.5, with a mean rating of 73.57 ( $SD = 16.02$ ). In the current case-series, the ratings ranged from 40 – 97.5, with a mean rating of 70.67 ( $SD = 14.59$ ). For both case-series, the mean ratings indicate MYLO is “acceptable” to most users, and scores were above average [32]. This suggests users' perceptions of MYLO's usability did not change between the two case-series, and the system is performing similarly to the previous case-series.

## Accessibility of MYLO Interface

These results were echoed in the qualitative data as many of the participants were generally positive about MYLO “I think it's rad. I think the idea is great.” (16-year-old). All the participants described the MYLO interface as easy to use “overall for me, I found it really easy to use.” (17-year-old). Many participants liked the accessibility of being able to talk about their problems anywhere “I could just click on the app when I wanted to use it, it would just start right away...it was very convenient.” (17-year-old). Please note that quotes throughout have been edited lightly for clarity.

Some participants described times where they would prefer to use MYLO instead of talking to a person, “I found it useful in that respect...instead of feeling like I had to... sit down and talk to someone about it.” (23-year-old). These participants were in two groups, those who would use MYLO when the people they would usually talk to were not available, “I do a lot of thinking in the nighttime...around those times when I'd be sorting



problems out. And it's good to have MYLO...especially at those times.” (16-year-old), and those who were reluctant to talk to people about their problems, “it was those points that I sort of reached for MYLO. I don't tend to be someone that usually reaches out and talks to someone about this stuff. I'm more private about it.” (17-year-old). One participant felt they were able to share more authentically than if they had been talking to a person,

“Having someone right there in front of you and looking you dead in the eyes, that's really confrontational. For me personally, I'll get nervous...now I don't feel comfortable talking about this. So it [MYLO] definitely felt comfortable. Just to be able to say it as it was.” (23-year-old).

Some participants felt the familiar chat format was helpful to support problem-processing,

“I prefer that [texting], I wouldn't mind talking on the phone, but I think it's nicer to be able to [text]. When I'm typing, I can process it better in my mind and it helps me come to terms with what I'm upset about.” (17-year-old).

and made them feel comfortable,

“I like the chat format setting because, I guess it almost feels like you're chatting with a friend in a way, because it's like a text message. So that provides almost a sense of comfort.” (17-year-old).

## Satisfaction With Therapy Sessions

We compared participants' satisfaction with MYLO's therapeutic conversations in the current case-series to the previous case-series. Table 6 presents the mean therapy satisfaction scores measured by the Session Impact Scale [27] across all weeks

(excluding baseline where the scale was not included) for all participants. We found a large difference in (standardized) effect sizes between the previous and current case series on the hindering impact scale, with current case-series participants rating MYLO higher on the subscale (indicating a larger hindering impact) than the previous case-series' participants. We also found a moderate difference in (standardized) effect sizes between the previous and current case series on unwanted thoughts, the higher rating in the current case-series indicates participants experienced more unwanted thoughts than the previous case-series participants.

**Table 6.** Session impact subscale scores for Manage Your Life Online (MYLO) in the current case-series and the previous case-series [17].

Session impact subscale	Current case-series (n=17)	Previous case-series (n=11)	Mean difference (95% CI)	Cohen's <i>d</i>
<b>Understanding</b>				
Mean (SD)	2.11 (0.61)	2.36 (0.99)	0.26 (-0.45 to 0.97)	0.33
<b>Problem-solving</b>				
Mean (SD)	1.91 (0.82)	2.09 (1.04)	0.18 (-0.60 to 0.96)	0.20
<b>Relationship</b>				
Mean (SD)	2.02 (0.83)	2.22 (0.92)	0.20 (-0.52 to 0.91)	0.23
<b>Hindering</b>				
Mean (SD)	2.49 (0.87)	1.76 (0.52)	-0.74 (-1.28 to -0.20)	-0.98
<b>Unwanted thoughts</b>				
Mean (SD)	1.84 (0.70)	1.50 (0.50)	-0.34 (-0.81 to 0.12)	-0.54

<sup>a</sup>Session impact subscale score: 1=not at all, 2=slightly, 3=somewhat, 4=very much, and 5=very much. n = number of responses across all weeks of each case-series. The Mean and SD were calculated by first calculating the mean for each participant in each case-series for each sub-scale, the table then displays the mean of these means. The Mean Difference is the Mean from the Previous case-series minus the Mean from the Current Case Series and so represents the reduction in mean session impact score from the previous to the current case-series. The SD to perform the standardization for *Cohen's d* is the pooled SD across the two case-series.

## Helpfulness of MYLO Conversations

Some participants found the conversations they had with MYLO useful, and that MYLO achieved its intended goal, "Sometimes it genuinely did make me reflect on my

thoughts and stuff like that. Like with those questions.” (16-year-old), “You know one of the defining characteristics for rumination, it's unproductive and persistent. So, the biggest thing for me about MYLO was that it felt productive because I think it helped me to think about things in a different way” (23-year-old). One participant felt at times, that MYLO could do a better job of helping the participant than a person,

“And I almost felt like it's therapeutic...I can just type it [a problem] out and then MYLO would just pick out all the important bits for me. And I don't have to focus. Whereas I feel like sometimes when you... [are] talking to people about it [a problem] ...you sometimes [have to] give them like all the context, and they'll get confused in the small details.” (23-year-old).

## Difficulties with MYLO Conversations

Despite some positive feedback, the SIS scores suggest that participants experienced hindering effects when using MYLO. Participants described four main problems: MYLO's questions were generic, MYLO was repetitive, MYLO had a limited purpose, MYLO was not “as good” as some freely available generative-AI chatbots.

### *Repetitive*

Like the previous case-series, the largest reported problem participants had with MYLO was repetition or looping. Although MYLO is programmed not to repeat the same question for at least 20 exchanges, participants felt questions were sometimes too like each other, “It just seems to be quite repetitive.” (16-year-old), or that MYLO asked questions about a topic they had already discussed, “Like sometimes I felt like I was going in circles. Like I'd say something. Then it'll give me a question and I'll be like, I just told that” (17-year-old).

## ***Generic***

Participants, also, felt that sometimes MYLO asked generic questions, “The only thing that I sort of felt was on the downside is that sometimes, especially if you're talking about a problem that's quite specific, it would only give general answers.” (17-years-old). One participant was conflicted as they sometimes found the similar or general questions useful, but at other times they caused frustration, “[MYLO] asking you to write more, sometimes was a bit frustrating when I didn't have anything to add. Sometimes it did help me to expand on what I was thinking and realize that maybe it's more complicated.” (17-year-old).

When either of these issues (i.e., repetitive or generic questioning) occur in a conversation they disrupt the flow of conversation and cause the participant to disengage, “It sort of jolted me backwards out of the conversation. I didn't expect [that response], that's not a response I would ever expect from a person that I was talking to.” (19-year-old). “If you'd kind of forgotten for a bit, you were reminded that this is just a computer.” (23-year-old).

## ***Limited Purpose***

Participants also felt that MYLO had a limited purpose and functionality at times. MYLO is programmed to only ask curious questions about a problem, and therefore does not engage in other forms of dialogue. Some participants wanted MYLO to engage in more casual conversation with them and suggested MYLO ask questions about their everyday life as well, “for people who have like a good day...or even just a bit lonely [but] you're still OK...I just want to have a bit of conversation” (17-year-old). This could improve MYLOs scores on the SIS as some participants did not currently feel understood by MYLO, “It doesn't feel like it really knows you.” (17-year-old).

### *Comparisons to Generative AI Chatbots*

A novel critique that participants described (6 out of 13), that did not appear in the previous case-series [17], was how MYLO compared to freely available generative AI chatbots (such as ChatGPT). Participants had experience of generative AI chatbots, “ChatGPT started blowing up this year as well. The top two examples [for] me personally I would take #1 being Snapchat and second being ChatGPT” (16-year-old). Although not all participants were as positive about generative AI or digital therapies,

“Knowing that it's a computer sometimes I don't take it super seriously or don't think that it's gonna be as helpful and that in itself can be a bit of a barrier. I'm not maybe trying or engaging with it as well as I could just because I'm not expecting it to be helpful.” (17-year-old).

Many had expectations that MYLO would look, “I guess something like the interfaces of AI chatbots, like for example like ChatGPT. [Were] what I was kind of expecting...it kind of like subverted my expectations.” (16 years old), and behave like other generative AIs (i.e., be able to answer user questions and have considerable natural language processing skills). Some were disappointed when these expectations were not met,

“I've used ChatGPT a bit...that [ChatGPT] kind of framed it for me that I was expecting [a similar experience] ...but that's where I got to a bit of a dead end with it [MYLO]. I wanted to ask it things like give me...some time management strategies” (23-year-old).

In the last case-series most participants would recommend MYLO to their friends, whereas the current participants appear to have a higher expectation for natural language programs given the inception of ChatGPT and other generative AIs,

“Definitely could be something I would recommend to other people, but maybe in this stage kind of what he (23-year-old) was saying before when you compare it to things like ChatGPT, which obviously have so much more money and like people behind it, it does kind of fall short” (17-year-old).

Participants had some specific issues and recommendations informed by their knowledge and experience of generative AI chatbots. For example, ChatGPT has a context window that allows it to “remember” what a user types within each conversation [33]. MYLO on the other hand, uses the immediate user response to generate its next question rather than all responses within a conversation, aside from rules included to avoid repetition of the exact same question within a certain number of messages. This is by design, so that MYLO is better able to emulate an MOL therapist and respond to current statements made by the user. Participants, however, wanted MYLO to recall everything they said, in line with their perception of how ChatGPT operates,

“I think that would be really, really useful, because when you look at ChatGPT like I've definitely used it and had asked it a question and then asked a question that was related to what it said and the previous question like following on and it's being able to respond and adjust. So I think that sort of capability in MYLO would definitely, really improve the conversation and would also probably really improve the chance of me recommending it.” (19-year-old).

Other participants enjoyed the free text responding they got from generative AI chatbots and ability to have a general conversation with them, compared to MYLO's pre-authored and structured responses, “I feel like MYLO doesn't really do that because of the fact that

it doesn't have the free speech. As I mentioned before, like for example ChatGPT, ...you can actually have a conversation with them, for example, the Snapchat AI.” (16-year-old). Despite these comments, participants did recognize the dangers of generative AI, “it's a bit more risky like you have to make sure where the information is drawing from” (17-year-old), and the therapeutic benefit of MYLO’s style of conversation,

“It's good that MYLO was just always asking me questions. In any sort of therapy setting, I guess it's better that the client would do more of the talking and the therapist would do more listening and asking questions.” (23-year-old).

## Discussion

The purpose of this study was to evaluate MYLO, a conversational agent’s ability to support the wellbeing of young people over four weeks. We aimed to build on a previous shorter case-series [17] by extending the time that users had access to MYLO, and by increasing the types of data collected; specifically, we added an adverse events survey and the potential for participants to share the conversations they had with MYLO with the research team. The addition of the adverse events survey allowed us to further assess the safety of MYLO as an intervention, and the results suggest that no adverse events were attributed to MYLO, although one individual found its lack of advice made them feel worse and caused them to seek advice from their usual supports. Although MYLOs instructions state that MYLO only asks questions and will not give advice, some users might not read the instructions. To avoid other users experiencing this problem in future, MYLO’s purpose and functionality could be made clearer and unavoidable either upon sign in or when opening a conversation.

Only a small number of participants (4 out of 24) consented to the research team decrypting and analyzing their conversations in this case-series. It is possible participants are uncomfortable sharing their conversations with MYLO due to the sensitive nature of the conversations. It is also possible that the method of informing and collecting consent contributed to this low participation. In future studies we aim to address both issues by implementing a method for participants to provide consent within the MYLO interface for each conversation they would be willing to have analyzed. This method will increase participants autonomy over their data and provide an easy and accessible way for participants to provide consent.

To evaluate MYLO's ability to support the wellbeing of young people we tested three hypotheses:

- H1 After using MYLO for four weeks participants' problem-related distress, anxiety and psychiatric impairment will decrease.
- H2 After using MYLO for four weeks participants' ability to engage in goal conflict reorganization would improve.
- H3 After using MYLO for four weeks participants' general health, depression and self-efficacy would improve.

The results supported hypothesis one and indicated that participants' problem-related distress, anxiety and psychiatric impairment showed substantial effect sizes for improvement after testing MYLO for four weeks. These findings aligned with the previous case-series [17]. The small sample size, large attrition rate, and lack of a control group mean these findings should be interpreted with caution, but they are promising and suggest MYLO could be an effective intervention for reducing problem-related distress,



anxiety and psychiatric impairment.

Hypothesis two was not supported as no change was found in participants' ability to engage in goal conflict reorganization. Inspection of the reliable change scores for all participants showed two participants achieved reliable improvement, while three reliably deteriorated. It is not clear why these results were different to the previous case-series [17], which showed a medium effect size increase in goal conflict reorganization. Goal conflict reorganization is the proposed mechanism of change for MOL and MYLO, by increasing clients' and users' ability to resolve the conflicts they experience, their problem-related distress should improve and subsequently their depression and anxiety symptoms should reduce [16]. Several MOL studies have demonstrated this increase in ROC scores [34-36]. Further research is needed to establish the impact of using MYLO on goal conflict resolution with a fully powered sample.

In relation to Hypothesis three, we found an effect on depression scores, and the reliable change score of individuals suggests some participants experienced an improvement in their self-efficacy scores. However, more research in a fully powered randomised trial is needed to validate these findings. The null findings related to general health could be due to the items in the SF-6Dv2, which focus on six domains of health including mental health. MYLO's focus on mental health and wellbeing limits its ability to improve the other domains in the SF-6Dv2, unless a user is seeking to explore a problem related to one of the other domains.

Regarding the user experience of MYLO, it scored similarly on the session impact scale to the previous case series, except for achieving a higher score on the hindering impacts sub-scale. As the system usability scale scores were similar, suggesting

participants found MYLO similarly acceptable to the participants in the last case-series, we examined the qualitative data to find a possible explanation for this difference. Much of the positive qualitative feedback was similar to the previous case-series, and other mental health chatbots [5,6]; participants praised the accessibility of the interface. Participants also described incidents where MYLO had helped them to achieve a new perspective on their problems. Participants also experienced similar difficulties regarding repetition of questions, and instances when MYLO did not understand them. These are difficulties that many chatbots, especially logic or rule-based chatbots, experience [9].

One difference that emerged was comparisons between MYLO and widely available generative AI models such as Chat-GPT and Gemini. The widespread availability of these chatbots appears to have altered users' expectations of what interacting with a chatbot is like. Our previous participants had not used these models, as they were not yet available, and while they experienced similar issues and scored MYLO similarly on all other SIS sub-scales (i.e., understanding, problem-solving and relationship), they were much more likely to recommend MYLO in its current form. In contrast, the current participants were less likely to recommend MYLO, despite similar reductions in problem-related distress and other mental health symptoms. Specifically, our users pointed to the natural language processing and generative AI chatbot's ability to adapt and perform many tasks (from writing computer code to answering complex questions on many topics) as superior to MYLO's specific question-answer structure.

Many mental health chatbots lack the functionality of these large generative AI models for many reasons (e.g., funding), but often by design. For example, as one participant noted, it is often beneficial for the client to do most of the talking during a

therapeutic conversation, whereas current generative AI models produce large amounts of text in response to minimal user input. Another drawback from generative AI recognized by our participants was the lack of transparency regarding where information or advice came from, whereas MYLO and other mental health chatbots generate text previously authored by mental health professionals, or generate text constrained by very specific rules created by mental health professionals [37]. Despite this, as technology advances, these interventions might no longer meet users' standards for how naturalistic a digital conversation should be. This could exacerbate the already difficult issue of retention in digital interventions and ultimately, limit the effectiveness of MYLO and other mental health chatbots.

The concerns regarding generative AI have been expressed more widely by researchers [38-41]. These include the significant cost and environmental impact of maintaining open-access models, which may become prohibitive for widespread applications as costs shift to consumers [42]. Additionally, maintaining conversational context is resource-intensive, a limitation noted across systems like ChatGPT and MYLO, though OpenAI's retention of previous discussions illustrates attempts to address this [43]. GenAI struggles with handling ambiguity, often generating poor responses to unclear inputs, and lacks true reasoning capabilities, relying instead on probabilistic predictions based on training data [43]. Emotional intelligence is another limitation, as these systems, including MYLO, cannot currently detect or adapt to users' emotional states. Furthermore, ethical concerns arise from biases in training datasets, particularly in sensitive areas like mental health [40-41, 44]. Unlike commercial GenAI tools, MYLO's design potentially offers a more controlled and transparent approach to mitigating these

risks, and therefore it may benefit from a judicious hybridization of the two systems.

Currently, the logic-based system behind MYLO uses human-selected therapeutic questions about specific themes and it prioritizes certain questions over others based upon ranking of topics. For example, terms indicating goal conflict such as ‘dilemma’ or ‘in two minds’ are ranked higher than terms about activities, such as ‘going to the gym’. In addition, questions that are consistently rated by past users as unhelpful are less likely to be selected. These elements will remain, yet they could be supplemented by large language models to address some of the recognized issues. For example, integrating better natural language processing could improve MYLO’s ability to: build rapport and user confidence in MYLO, detect a wider range of terms, identify themes that are repeated across and within conversations, detect user statements that are not directed at problem exploration to answer them separately, and redirect the user to the purpose of MYLO. The current research has provided a good basis to justify and begin to guide such improvements through the next co-design phase.

## Limitations

We used a case-series design to evaluate MYLO and assess the acceptability of new data collection methods. The lack of a control group limits our ability to attribute the treatment effects to usage of MYLO, and they should be interpreted with caution. The findings of this study will be used to inform the design of a randomized controlled trial that compares MYLO users to a waitlist group to provide a more robust and valid assessment of MYLO’s efficacy. We also experienced high levels of dropout between baseline and the first week of testing when participants were required to download and log in to MYLO. This step was added to this case-series. It was not necessary in the

previous MYLO projects [17] where participants were able to use MYLO without an account through a weblink. The current process requires the participant to inform the lead researcher they had completed the baseline survey; the lead researcher then creates an account for the user and emails or texts them their log in details as well as instructions on how to download and use MYLO. This process was a necessity for this stage given the authentication platform being used (Auth0) and MYLOs current stage of development. However, previous consumer panel members have indicated a preference for password-less access, stating that passwords and usernames are a barrier for some young people [17]. We aim to address this issue in the next research and development phase by allowing users to create their own accounts and let them use their saved passwords, biometrics or social media log ins to authenticate their identity. Ultimately, the goal is to create a smartphone application version of MYLO that can be placed on all available application stores. Doing so will greatly increase MYLOs accessibility and availability to a broader sample of users.

## Conclusions

This study aimed to extend our previous work on MYLO by providing the application for a longer period of time, 4 weeks, and examining its impact on MYLO users' wellbeing. We also assessed the feasibility of changes to the trial methodology, i.e., adding the option for participants to consent to decryption and analysis of their conversations prior to conducting a full-scale randomized controlled trial. All changes will be retained moving forward but we are hoping to integrate consent to decrypt conversations into the application interface to increase the number of participants opting in, and to allow participants to opt in for individual conversations. Regarding the

quantitative results, they show promising evidence that MYLO can effectively reduce problem-related distress, anxiety, and psychiatric impairment, aligning with previous research on MYLO. However, the lack of change in participants' goal conflict reorganization and the mixed findings on self-efficacy and general health highlight the need for further investigation with a fully powered randomised controlled trial to better understand the intervention's impact on participants' wellbeing and mental health, as well as the mechanisms driving these changes.

The study also revealed key insights into the user experience of mental health conversational agents, particularly in comparison to widely available generative AI models like Chat-GPT and Gemini. While MYLO scored similarly on user experience and usability metrics, the changing expectations of users in response to advances in AI technologies suggest that integrating more sophisticated natural language processing features might enhance its usability, rapport-building, and therapeutic outcomes. Further, addressing technological barriers, such as streamlining user authentication and developing a smartphone application, will enhance MYLO's accessibility and engagement. Future iterations of MYLO, incorporating co-designed improvements and hybrid approaches leveraging large language models, hold the potential to better meet users' evolving expectations and improve retention in digital mental health interventions.

## Acknowledgements

All authors contributed to study conception and design. AW-H, GA, and JD contributed to data collection. AW and WM contributed to the analysis and interpretation of the results. AW prepared the first draft. All authors reviewed the results, provided feedback on drafts and approved the final version of the manuscript. The authors would like to thank Jason Wright, the youth advisory committees, and the study participants for their contributions to this study. Their expertise, feedback, and time were invaluable in making Manage Your Life Online what it is today.

## Conflicts of Interest

None declared

## Abbreviations

MOL: method of levels

MYLO: Manage Your Life Online

PCT: perceptual control theory

RCT: randomized controlled trial

SF-6Dv2: Short Form-6D version 2

SIS: session impact scale

## References

1. Racine, N, McArthur, B A, Cooke, J E, Eirich, R, Zhu, J, & Madigan, S. Global prevalence of depressive and anxiety symptoms in children and adolescents during COVID-19: a meta-analysis. *JAMA Pediatrics* 2021;175(11):1142-1150. Doi: 10.1001/jamapediatrics.2021.2482
2. Minerva F, Giubilini A. Is AI the Future of Mental Healthcare? *Topoi (Dordr)*. 2023 May 31;42(3):1-9. doi: 10.1007/s11245-023-09932-3. Epub ahead of print. PMID: 37361723; PMCID: PMC10230127.
3. Lehtimäki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on Digital Mental Health Interventions for Adolescents and Young People: Systematic Overview. *JMIR Ment Health*. 2021 Apr 29;8(4):e25847. doi: 10.2196/25847. PMID: 33913817; PMCID: PMC8120421.
4. Koh, J, Tng, GYQ, & Hartanto, A. Potential and pitfalls of mobile mental health apps in traditional treatment: an umbrella review. *Journal of Personalized Medicine* 2022;12 (1376):1 -27. Doi: 10.3390/jpm12091376.
5. Boucher, EM, Harake, NR, Ward, HE, Stoeckl, SE, Vargas, J, Minkel, J, Parks, AC, & Zilca, R. Artificially intelligent chatbots in digital mental health interventions: a review. *Expert Review of Medical Devices* 2021;18(sup1):37-49. Doi: 10.1080/17434440.2021.2013200
6. Gaffney, H, Mansell, W, & Tai, S. Agents of change: Understanding the therapeutic processes associated with the helpfulness of therapy for mental health problems with relational agent MYLO. *Digital Health* 2020;6:1-19. Doi: 10.1177/2055207620911580
7. Gaffney, H, Mansell, W, & Tai, S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. *JMIR Mental Health* 2019;6(10):e14166. Doi: 10.2196/14166
8. Vaidyam, AN, Wisniewski, H, Halamka, JD, Kashavan, MS, & Torous, JB. Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry* 2019;64(7):456-464. Doi: 10.1177/0706743719828977
9. Egan, SJ, Johnson, C, Wade, TD, Calbring, P, Raghav, S, & Shafran, R. A pilot study of the perceptions and acceptability of guidance using artificial intelligence in internet cognitive behaviour therapy for perfectionism in young people. *Internet Interventions* 2024;35:1-7. Doi: 10.1016/j.invent.2024.100711.
10. Haque, MDR & Rubya, S. "For an app supposed to make its users feel better, it sure is a joke" – An analysis of user reviews of mobile mental health applications. *Proc. ACM Hum. – Comput. Interact.* 2022;6;No. CSCW2;Article 421:1-29. Doi: 10.1145/3555146.



11. Ma, X, & Huo, Y. Are users willing to embrace ChatGPT? Exploring the factors on the acceptance of chatbots from the perspective of AIDUA framework. *Technology in Society* 2023; 75:1-13. Doi: 10.1016/j.techsoc.2023.102362.
12. Singla, A, Sukharevsky, A, Yee, L, Chui, M, & Hall, B. The state of AI in early 2024: Gen AI adoption spikes and starts to generate value. *Quantum Black AI by McKinsey* 2024. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai#/>
13. Carey, TA. *The Method of Levels: How to do psychotherapy without getting in the way*. Hayward, CA: Living Control Systems Publishing; 2006. ISBN: 0974015547
14. Mansell, W, & Goldstein, D. *Methods of Levels Therapy. The Interdisciplinary Handbook of Perceptual Control Theory*. Academic Press; 2020. ISBN: 9780728189481
15. Powers, WT. *Behavior: The control of perception*. New York: Aldine; 1973.
16. Gaffney, H, Mansell, W, Edwards, R, & Wright, J. Manage your life online (MYLO): a pilot trial of a conversational computer-based intervention for problem solving in a student sample. *Behavioural and Cognitive Psychotherapy* 2014;42(6):731-746. Doi: 10.1017/s135246581300060x
17. Wrightson-Hester A, Anderson G, Dunstan J, McEvoy P, Sutton C, Myers B, Egan S, Tai S, Johnston-Hollitt M, Chen W, Gedeon T, & Mansell W. An Artificial Therapist (Manage Your Life Online) to Support the Mental Health of Youth: Co-Design and Case Series. *JMIR Hum Factors* 2023;10:e46849 URL: <https://humanfactors.jmir.org/2023/1/e46849>. Doi: 10.2196/46849
18. Evans, C, Margison, F, & Barkham, M. The contribution of reliable and clinically significant change methods to evidence-based mental health. *Evidence-Based Mental Health* 1998;1(3):70-100. Doi: 10.1136/ebmh.1.3.70
19. Kroenke, K, Spitzer RL, Williams, JBW. The PHQ-9, validity of a brief depression severity measure. *Journal of General Internal Medicine* 2001;16:606-613. Doi: 10.1046/j.1525-1497.2001.016009606.x
20. Audette, LM, Hammond, MS, & Rochester, NK. Methodological issues with coding participants in anonymous psychological longitudinal studies. *Educational and Psychological Measurement* 2020;80(1):163-185. Doi: 10.1177/0013164419843576
21. Spitzer RL, Kroenke, K, Williams, JBW, Löwe, B. A brief measure for assessing generalised anxiety disorder: the GAD-7. *Archives of Internal Medicine* 2006;166(10):1092-1097. Doi: 10.1001/archinte.166.10.1092
22. Goldberg, D, & Williams, P. *A user's guide to the General Health Questionnaire*. Windsor, UK: NFER-Nelson; 1988.
23. Brazier, JE, Mulhern, BJ, Bjorner, JB, Gandek, B, Rowen, D, Alonso, J, Vilagut, G, &

- Ware, JE. Developing a new version of the SF-6D health state classification system from the SF-36v2: SF-6Dv2. *Medical Care* 2020;58(6):557-565. Doi: 10.1097/mlr.0000000000001325
24. Ashworth, M, Shepherd, M, Christey, J, Matthews, V, Wright, K, Parmentier, H, Robinson, S, & Godfrey, E. A client-generated psychometric instrument: The development of 'PSYCHLOPS' ('Psychological Outcome Profiles'). *Counselling and Psychotherapy Research* 2004;4:27-31. Doi: 10.1080/14733140412331383913
25. Bird, T. An Investigation of transdiagnostic processes and interventions in clinical and non-clinical settings (Doctoral thesis). University of Manchester, United Kingdom; 2013.
26. Schwarzer, R, & Jerusalem, M. Generalized self-efficacy scale. In J. Weinman, S. Wright, & M. Johnston, *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35-37). United Kingdom: NFER-NELSON; 1995.
27. Elliot, R, & Wexler, MM. Measuring the impact of sessions in process-experiential therapy of depression: The session impacts scale. *Journal of Counselling Psychology* 1994;41(2):166-174.
28. Brooke, J. SUS: A "quick and dirty" usability scale. In PW. Jordan, B Thomas, BA. Weerdmeester, & IL McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). London: Taylor & Francis; 1996. ISBN: 0748404600
29. Vears, DF, & Gillam, L. Inductive content analysis: a guide for beginning qualitative researchers. *Focus on Health Professional Education* 2022;23(1):111-127. Doi: 10.11157/fohpe.v23i1.544
30. Anjara, SG, Bonetto, C, Van Bortel, T, & Brayne, C. Using the GHQ-12 to screen for mental health problems among primary care patients: psychometrics and practical considerations. *International Journal of Mental Health Systems* 2020; 14(62): 1-13. Doi: 10.1186/s13033-020-00397-0
31. Navarro, DJ, & Foxcroft, DR. Comparing two means: Effect size. [https://lsj.readthedocs.io/ko/latest/Ch11/Ch11\\_Test\\_07.html](https://lsj.readthedocs.io/ko/latest/Ch11/Ch11_Test_07.html)
32. Sauro, J. 5 Ways to Interpret a SUS Score. *Measuring U*; 2018. <https://measuringu.com/interpret-sus-score/>
33. Pena, D. Understanding how Chat-GPT maintains context. Sitepoint: <https://www.sitepoint.com/understanding-how-chatgpt-maintains-context/>
34. [Gluckman, N. An evaluation of online method of levels therapy with young people \(Doctoral Thesis\). University of East London 2022. https://repository.uel.ac.uk/item/8v668](https://repository.uel.ac.uk/item/8v668)
35. Churchman, A, Mansell, W, & Tai, S. A process-focused case-series of a school-based intervention aimed at giving young people choice and control over their attendance and their goals in therapy. *British Journal of Guidance & Counselling*

- 2021; 49(4): 565-586. Doi: [10.1080/0369885.2020.1815650](https://doi.org/10.1080/0369885.2020.1815650)
36. Griffiths, R, Mansell, W, Carey, TA, Edge, D, Emsley, R, Tai, SJ. Method of levels therapy for first-episode psychosis: The feasibility randomized controlled Next Level trial. *Journal of Clinical Psychology* 2019; 75(1): 1756-1769. Doi: 10.1002/jclp.22820
37. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics* 2019; 132:103978. Doi: [10.1016/j.ijmedinf.2019.103978](https://doi.org/10.1016/j.ijmedinf.2019.103978)
38. Pandya, A, Lodha, P, & Ganatra, A. Is ChatGPT ready to change mental healthcare? Challenges and considerations: a reality-check. *Frontiers in Human Dynamics* 2024; 5. Doi: 10.3389/fhumd.2023.1289255.
39. Alanezi, F. Assessing the effectiveness of ChatGPT in delivering mental health support: A qualitative study. *Journal of Multidisciplinary Healthcare* 2024; 17:461-471. Doi: 10.2147/JMDH.S447368
40. Kalam, KT, Rahman, JM, Islam, MR, & Dewan, SMR. ChatGPT and mental health: Friends or foes? *Health Science Reports* 2024; 7(2): e1912. Doi: 10.1002/hsr2.1912
41. Raile, P. The usefulness of ChatGPT for psychotherapists and patients. *Humanit Soc Sci Commun* 2024; 11:47. Doi: 10.1057/s41599-023-02567-0
42. Crawford, K. Generative AI's environmental costs are soaring – and mostly secret. *Nature* 2024; 626, 693. Doi: 10.1038/d41586-024-00478-x
43. OpenAI. Memory and new controls for ChatGPT. OpenAI; 2024 <https://openai.com/index/memory-and-new-controls-for-chatgpt/>
44. Yan L, Greiff S, Teuber Z, Gašević D. Promises and challenges of generative artificial intelligence for human learning. *Nat Hum Behav.* 2024 Oct;8(10):1839-1850. doi: 10.1038/s41562-024-02004-5. Epub 2024 Oct 22. PMID: 39438686.

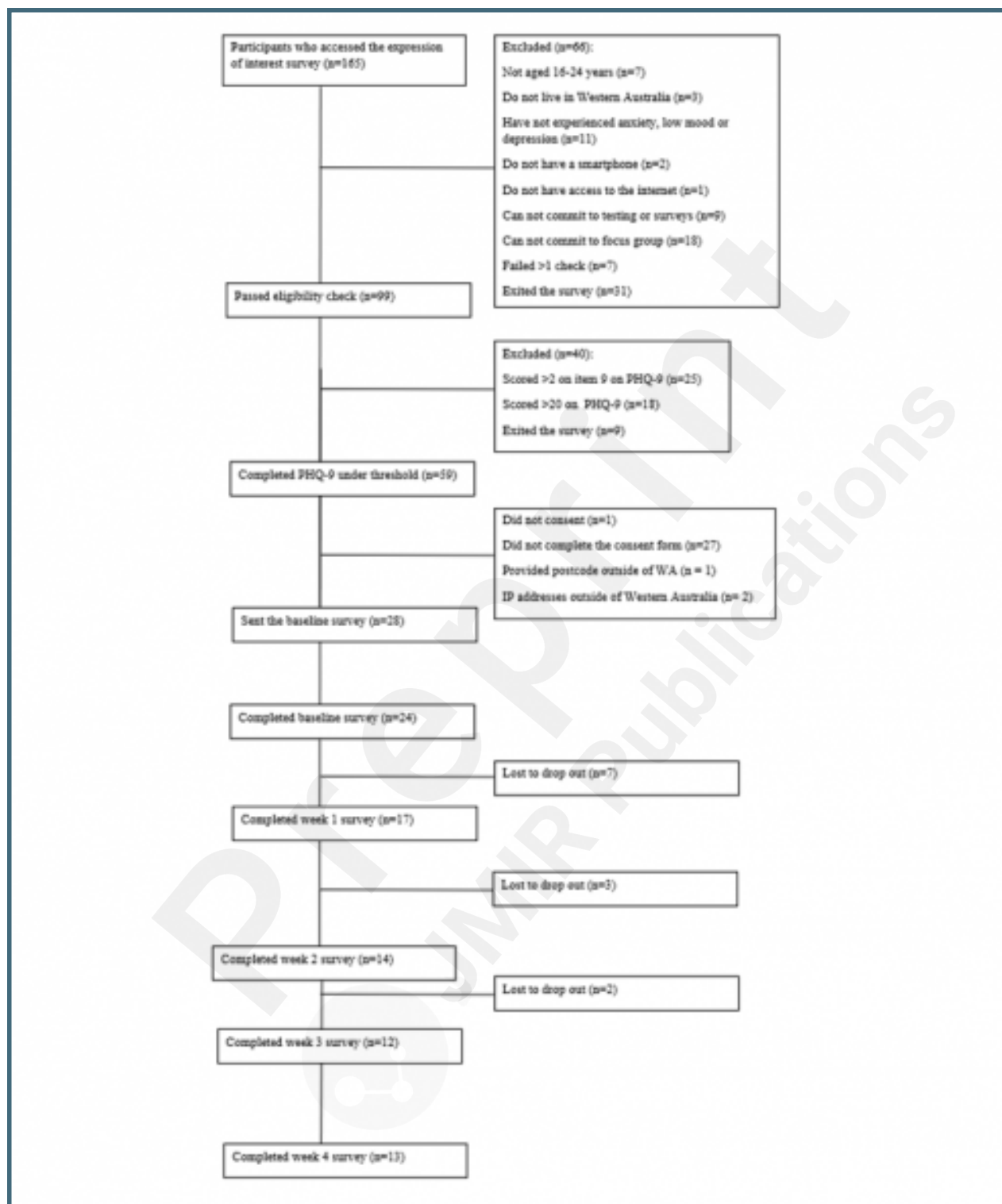
## Supplementary Files

## Figures

Manage Your Life Online interface.



Participant flowchart through the study from EOI to post-testing survey.



## Multimedia Appendixes



Spreadsheet presenting the baseline and 4 week (post-testing) scores on all clinical outcomes for completers.

URL: <http://asset.jmir.pub/assets/d57a663d197064e89e392c0826b59cb9.xlsx>

