A pilot randomised controlled trial exploring the feasibility and efficacy of a human-AI sleep coaching model for improving sleep among university students

Jintana Liu[1] iD, Sakura Ito[1] iD, Tra My Ngo[1], Ashwini Lawate[1],
Qi Chwen Ong[1] iD, Tatiana Erlikh Fox[1], Si Yuan Chang[1], Duy Phung[1],
Elizabeth Nair[2], Malar Palaiyan[3], Shafiq Joty[4,5], John Abisheganaden[6],
Chuen Peng Lee[6] iD, May Oo Lwin[7], Yin Leng Theng[7], Moon-Ho Ringo Ho[8],
Michael Chia[9] iD, Iva Bojic[1] and Josip Car[1,10]

Abstract

Objective: Sleep quality is a crucial concern, particularly among youth. The integration of health coaching with question-answering (QA) systems presents the potential to foster behavioural changes and enhance health outcomes. This study proposes a novel human-AI sleep coaching model, combining health coaching by peers and a QA system, and assesses its feasibility and efficacy in improving university students' sleep quality.

Methods: In a four-week unblinded pilot randomised controlled trial, 59 university students (mean age: 21.9; 64% males) were randomly assigned to the intervention (health coaching and QA system; $n = 30$) or the control conditions (QA system; $n = 29$). Outcomes included efficacy of the intervention on sleep quality (Pittsburgh Sleep Quality Index; PSQI), objective and self-reported sleep measures (obtained from Fitbit and sleep diaries) and feasibility of the study procedures and the intervention.

Results: Analysis revealed no significant differences in sleep quality (PSQI) between intervention and control groups (adjusted mean difference $= -0.51$, 95% CI: $[-1.55-0.77]$, $p = 0.40$). The intervention group demonstrated significant improvements in Fitbit measures of total sleep time (adjusted mean difference $= 32.5$, 95% CI: $[5.9-59.1]$, $p = 0.02$) and time in bed (adjusted mean difference $= 32.3$, 95% CI: $[2.7-61.9]$, $p = 0.03$) compared to the control group, although other sleep measures were insignificant. Adherence was high, with the majority of the intervention group attending all health coaching sessions. Most participants completed baseline and post-intervention self-report measures, all diary entries, and consistently wore Fitbits during sleep.

Conclusions: The proposed model showed improvements in specific sleep measures for university students and the feasibility of the study procedures and intervention. Future research may extend the intervention period to see substantive sleep quality improvements.

[1]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore
[2]Work & Health Psychologists, Singapore, Singapore
[3]University Counselling Centre, Nanyang Technological University, Singapore, Singapore
[4]Salesforce AI Research, San Francisco, CA, USA
[5]School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore
[6]Department of Respiratory and Critical Care Medicine, Tan Tock Seng Hospital, Singapore, Singapore
[7]Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore
[8]School of Social Sciences, Nanyang Technological University, Singapore, Singapore
[9]Physical Education and Sports Science, National Institute of Education, Nanyang Technological University, Singapore, Singapore
[10]School of Life Course & Population Sciences, King's College London, London, UK

Corresponding author:
Iva Bojic, Lee Kong Chian School of Medicine, Nanyang Technological University, 11 Mandalay Road, Singapore 308207, Singapore.
Email: iva.bojic@ntu.edu.sg

Introduction

Sleep is a daily-living activity that plays a vital role in maintaining the health and well-being of individuals.[1] However, university students often encounter various stressors that can disrupt quality sleeping patterns. This includes emotional and academic demands,[2] new living arrangements, financial stress[3] and issues such as alcohol use[4] or stimulant abuse.[5] Despite compelling evidence highlighting the benefits of adequate sleep,[6] poor sleep quality is highly prevalent among university students. Many are at high risk for sleep disorders,[7] with over 60% of university students experiencing sleep disturbances.[2] Moreover, daytime sleepiness affects over half of university students compared to 36% of adolescents and adults.[8] The majority of students report having less than the recommended eight hours of sleep,[2] despite the National Sleep Foundation recommendations for young adults to sleep between seven and nine hours (although individual variability exists).[9] Young adults also often display poor sleep quality characteristics such as high levels of wake after sleep onset (WASO).[10]

Poor sleep quality has been linked to various negative psychological consequences including depression,[11] anxiety[12] and loneliness.[13] It is also commonly associated with numerous chronic health problems such as heart disease,[14] high blood pressure,[15] stroke[16] and obesity.[17] These outcomes lead to further impacts on common sleep quality indicators, such as sleep efficiency (SE),[18] night-time sleep awakening frequency[19] and sleep latency (SL).[20] Furthermore, additional contributors to poor sleep behaviours include a lack of sleep education and sleep hygiene knowledge among university students.[21–23] Given these prevalence rates and the adverse effects of poor sleep, there is a cogent need to develop effective sleep interventions in university students that increase behaviours that improve sleep quality among university students.

Achieving behavioural change for healthy sleep quality is complex and challenging, with lack of knowledge often being a key contributor.[23] Understanding the importance of sleep health is the initial step, but the abundance of information can be overwhelming. To bridge this gap, question-answering (QA) systems have emerged as a tool to enhance knowledge and understanding on numerous topics by providing short and precise answers to questions posed in natural language.[24] This is achieved through natural language processing (NLP), a branch of artificial intelligence (AI) with rapid developments and vast applications using large language models (LLM) for QA. Question-answering systems possess an abundance of domain knowledge, where biomedical QA systems can be trained on evidence-based medical information to increase the accessibility of expert opinions. This mimics direct access to an expert by providing timely and accurate responses to user's queries, allowing them to access evidence-based information in real-time. These QA systems have been applied for use in clinical decision support,[25,26] medical examinations,[27,28] consumer health questions[29] and to improve numerous health outcomes,[24,25] including sleep outcomes in university settings.[30] However, despite the QA system's abundance of knowledge, providing information to patients alone in this form is unlikely to be sufficient to promote behavioural change.[31]

Hence, health coaching, a person-centred approach, offers valuable human interaction that maximises an individual's potential, increases health knowledge and awareness and improves health outcomes in a sustainable and positive manner.[32,33] This can empower patients for long-term behavioural change[34] through active learning, involvement in social support and problem-solving.[35] Health coaches are adept at developing personalised plans with tailored behavioural change techniques for the client's needs, with abilities to synthesise evidence-based information to create actionable insights and goals. This has displayed promise in intervening in various health conditions, leading to improvements in diet,[36] mental well-being,[37] physical activity[38] and sleep.[39] In university settings, health coaching implemented by peers has proven effective in promoting healthy lifestyles among students.[40–42] By employing peer health coaches to intervene on health outcomes like sleep, it serves as an effective, scalable and sustainable approach to intervention delivery.

Although QA systems can effectively provide information, behavioural interventions rely heavily on psychological principles for their successes. Hence, to enhance these interventions, incorporating human health coaches who can incorporate evidence-based data obtained from the QA system and personalise the information to each participant's individual needs is crucial. Thus, AI solutions aiming to change individual behaviours are best complemented and overseen by human involvement. Question-answering systems have additional potential to

augment peer coaching by providing extra information and insight for decision-making.[43,44] Furthermore, they can assist in overcoming any knowledge gaps that may arise from coaches not being domain experts. By integrating QA systems, health coaches can access evidence-based and medically reviewed information for effective, accurate and comprehensive coaching. Therefore, we propose a human-AI symbiosis approach to combine the benefits of human health coaching (with weekly peer coaching interactions), with the advantages of AI (via the QA system) to facilitate improvements in sleep quality.

To our knowledge, there is no human-AI sleep coaching approach designed specifically to address sleep quality among university students. This study aimed to assess the feasibility and efficacy of the human-AI sleep coaching model to improve sleep quality among university students. The primary outcome assessed is change in sleep quality as measured by the Pittsburgh Sleep Quality Index (PSQI). Secondary outcomes include: (1) changes in objective and subjective measures of sleep as measured by Fitbit and sleep diaries; and (2) feasibility of study procedures to inform the design of a full-scale randomised controlled trial (RCT).

Methods

Study design

The study was a two-arm pilot RCT investigating the human-AI sleep coaching model to improve sleep quality among university students across four weeks. The study was conducted at a local university in Singapore, and ethical approval was obtained from the Institutional Review Board of Nanyang Technological University (IRB-2021-739). This study was not pre-registered. This study followed the reporting guidelines for pilot studies set by Consolidated Standards of Reporting Trials and a corresponding checklist is available in Supplemental Appendix 1.

Sleep QA system

The extractive Sleep QA system[45,46] developed by the study team, utilises NLP techniques to provide answers on sleep-related queries presented in a factoid manner (i.e., questions that elicit answers that can be succinctly expressed in short texts).[47] The system was built and trained on an expert-annotated dataset comprising over 7000 medically reviewed online articles related to sleep health to curate an extensive and evidence-based dataset. A data-centric approach was adopted in the development of the QA system to improve its accuracy by refining the identification of negative passages in retrieval fine-tuning, and question reformulation (e.g., paraphrasing, back translation).

To use the system, users can enter their questions into the text box, and the system generates an immediate response. An online web interface was created to provide easy access to the QA system via computer, tablet or smartphone. It is important to note that the system is currently intended for research purposes only and should not be considered as a substitute for professional healthcare or medical advice. It was designed as a supplementary tool for care and all users were required to sign and acknowledge a medical disclaimer acknowledging this prior to accessing the system.

Intervention protocol

Control (QA system). Participants in this group were advised to engage directly with the QA system (see Figure 1) on a weekly basis and were informed to access the interface on their preferred platform (e.g., phone computer or tablet). They were given guidance to ask factoid questions relating to aspects of sleep (e.g., sleep quality, sleep hygiene).

Intervention (QA system and health coaching). Participants in the intervention group engaged in a weekly 30 min health coaching session across the study period conducted via an online synchronous text-based platform (Microsoft Teams) on their computer, tablet or smartphone. During the sessions, health coaches would initiate the weekly interaction with their assigned peers and ask them questions on various sleep aspects (e.g., duration, patterns, quality) while incorporating behaviour change techniques (e.g., motivational interviewing, goal setting) to adjust to their respective needs. Participants were also instructed to seek additional sleep advice by asking their coaches questions that pertain to sleep. When participants posed factoid questions, health coaches could retrieve answers from the QA system for additional evidence-based information to aid their response. If the answers and explanations provided by the system were accurate, they were advised to incorporate the information obtained into their coaching conversations. However, if the system's response was deemed incorrect or unhelpful, coaches were instructed to use their expertise and reliable sources (e.g., clinical guidelines) to answer participants' questions and provide tailored coaching. Health coaches were undergraduate and postgraduate current students from various backgrounds (e.g., psychology, medicine, business) who had completed and passed the 8-week university health coaching course: 'Health Coaching: Introduction to Being Cool & Sleep as a Power', which covered how to apply coaching skills in areas such as sleep, mental health and exercise. Health coaches also received training to provide an overview of how to use the QA system in their coaching. Additionally, they attended weekly group supervision with professional health coaches to communicate concerns and receive advice for their respective coaching sessions.
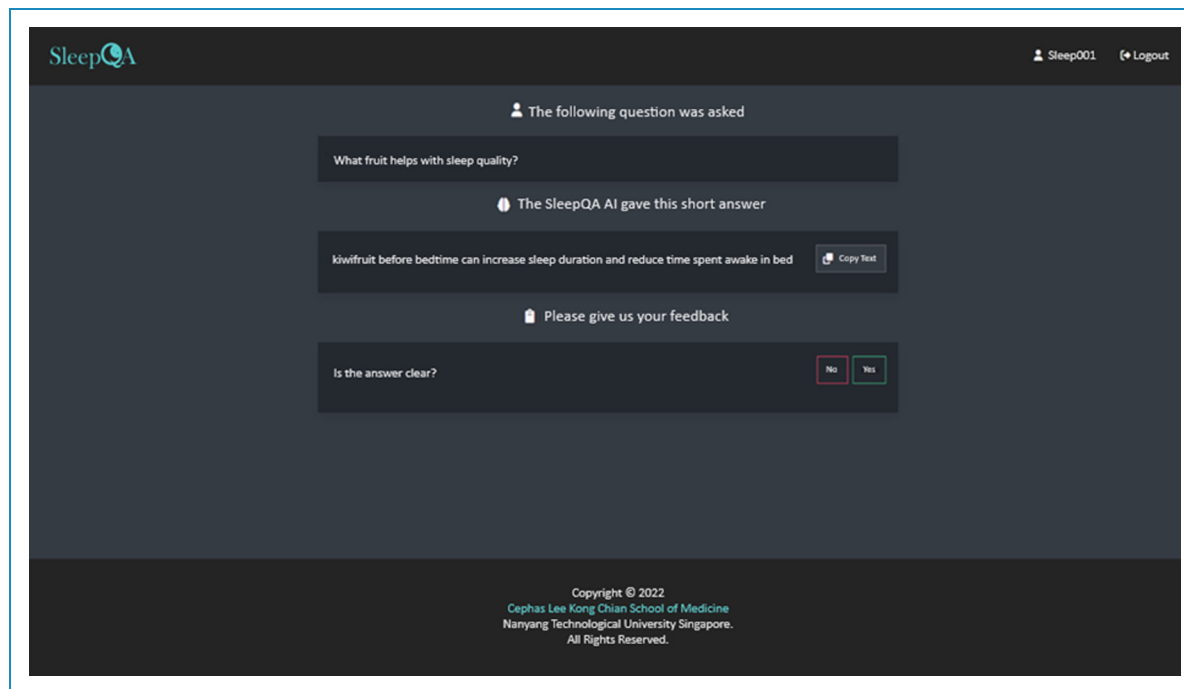
Figure 1. Screenshot of the developed question-answering (QA) system.

Lastly, the health coaches had access to the study team's dedicated sleep specialists, providing them with a valuable resource for addressing any specific questions or concerns related to sleep.

Participant recruitment

Participants were recruited by sending university-wide emails and by approaching student organisations. Through a link provided in the email, participants were directed to a form on Qualtrics (an online survey platform), where they were provided with information including the study aim, objectives and participation requirements. Students who were interested to participate were able to register their interest and were subsequently screened for eligibility. Written informed consent was obtained from participants before the study commenced.

Participant eligibility

Participants were deemed eligible to participate if they met the following inclusion criteria: (a) were at least 21 years old (in compliance with Singapore's legal age of consent); (b) did not display depressive symptoms (as assessed by having a score below 10 on the Patient Health Questionnaire-9 (PHQ-9); (c) were a current student at this university; and (d) were not undergoing any treatment for sleep and/or mental disorders and not under the care of a psychologist or psychiatrist as our intervention was not tailored for severe sleep disorders (e.g.,

narcolepsy, parasomnias) which require more intensive treatments. Individuals at screening who: (a) indicated depressive symptoms above PHQ-9 > 10; (b) indicated suicidal ideation on the PHQ-9 via Item 9; (c) were not a current student at this university; or (d) were undergoing any treatment for sleep and/or mental disorders, or under the care of a psychologist or psychiatrist were deemed ineligible to participate. In the case that the participants demonstrated depressive symptoms or suicidal ideation, they were referred to professional support services at the university's counselling centre.

Sample size

As a pilot study, this study aimed to inform the development of a future larger trial. Therefore, a formal sample size calculation was not conducted. Based on best practices for sample sizes of pilot studies, 30 participants per group is recommended.[48] As such, we planned to recruit 60 participants (30 per group).

Trial procedures

Participants meeting the eligibility criteria for study participation remotely provided written electronic informed consent and completed the baseline survey. Participants then scheduled for an onboarding session for study briefing and were loaned a Fitbit Inspire 2 for data collection during the study period. Attendance of the onboarding session was mandatory for participants to proceed with the study

procedure. Eligible participants were randomly assigned to the intervention group or the control group in a 1:1 ratio using a random number generator. Participants were not blinded to their allocated group and were notified during the onboarding session of their group assignment.

During the intervention period, all participants were asked to complete a daily sleep diary and weekly sync their Fitbits for data collection. Participants received weekly reminders via email to upload sleep diaries and to sync their Fitbit data by accessing the Fitbit mobile app. After the intervention period, participants were asked to complete the post-intervention questionnaire. They were also asked to schedule and attend an off-boarding session for study debriefing and to return the Fitbit. Participants were reimbursed with vouchers for participant burden.

Safety net procedure

A safety net procedure was put into place throughout this trial. Health coaches were trained to identify signs of mental distress, sleep disorders or at-risk behaviours (e.g., suicide ideation, harm to self and others). While participants were not explicitly asked about their mental distress or sleep disorders, health coaches were instructed to ask their clients about their sleep health and well-being and monitor any changes in sleep behaviours. In the case that participants demonstrated any signs of mental distress or sleep disorders (e.g., irregular sleep patterns, difficulty falling asleep and severe mood swings), health coaches were instructed to inform the study team. They were told to not proceed with coaching and to consult with the professional health coach on stand-by. A referral would be made for the participant to the university's counselling centre (in the case of severe mental distress) or to sleep physicians (in the case of sleep disorder) if necessary. If participants demonstrated at-risk behaviours, health coaches were instructed to follow a university-wide protocol for at-risk students.

Primary outcome

Sleep quality. The PSQI was administered at baseline and post-intervention. The self-report questionnaire assesses subjective quality of sleep through 19 questions with scores ranging from 0 to 21. Total scores between 0 and 4 represent good sleep quality, while scores 5 and above are the threshold indicating poor sleep quality. The PSQI assesses seven components, namely subjective sleep quality, SL, sleep duration, SE, sleep disturbances, use of sleep medication and daytime dysfunction.[49] A clinically significant change in PSQI is a change of three points or more.[50]

Secondary outcomes

Fitbit sleep measures. The Fitbit Inspire 2 was used to measure sleep metrics objectively. Sleep efficiency, total sleep time (TST), time in bed (TIB), SL, number of awakenings during the night (NWAK) and WASO were derived from the Fitbit data for the purposes of this study. These variables are commonly reported in the literature to assess sleep objectively among university students.[51,52] By definition, SE is a percentage and could take values from 0 to 100%. Total sleep time, TIB, SL and WASO are measured in minutes and take non-negative values.

Self-reported sleep. The Sleep Foundation Sleep Diary[53] was adopted as a daily self-monitoring tool to record sleep habits and sleep hygiene. A sleep diary is considered the 'gold standard' in terms of subjective sleep assessment[54] and in non-laboratory conditions.[54,55] Data collected included subjective sleep quality, sleep disturbances, caffeine and medication consumption, daily exercise and bedtime routine. Additional data collected included bedtime, wake-up time, number and duration of daytime naps and sleep disruptions. For the purposes of this study, analyses on sleep diary data were only conducted on the following metrics: SE, TST, SL, NWAK and WASO, in alignment with the Fitbit data collected. TIB was not calculated for sleep diary data due to the absence of questions relating to 'What time did you get out of bed for the day?'.

Feasibility. The feasibility of the study procedure and the intervention was measured based on existing feasibility indicators.[56,57] The feasibility of the study procedure measured the following: (a) recruitment rate (as calculated by the percentage of people who agree to participate in the study among those who are approached and are eligible to participate); (b) retention rate (as calculated by the percentage of participants who completed the post-intervention questionnaire); and (c) adherence to sleep diary and Fitbit data collection. The feasibility of the intervention (defined as intervention adherence and engagement) was measured by the following: (a) participants' weekly attendance to health coaching sessions; and (b) time spent per coaching session.

Statistical analyses

To assess treatment effect for the primary sleep quality outcome PSQI, linear regression models with post-treatment measurement as dependent variable and pre-treatment measurement as independent variable were fitted. Interaction between pre-treatment measurement and interventions was not included to allow for better power. For sensitivity analysis, a proportional odds model for PSQI was also fitted.

For the Fitbit and self-reported sleep outcomes, we fitted mixed regression models with baseline measurement, time (number of days from baseline), interaction of treatment and baseline as well as interaction of treatment and time

as independent variables, with all models having a random intercept to account for the repeated measurements across subjects. For these outcomes, interaction terms were included since the analyses were exploratory by nature and thus statistical power is less of a concern. Fitbit and self-reported SE, TST, WASO as well as Fitbit measured TIB and self-reported SL, were fitted with linear mixed models. Sleep latency from Fitbit was fitted with mixed Tweedie regression. The NWAK measures were fitted with mixed Poisson model. The significance level for hypothesis testing was chosen at 5%, where all *p*-values are reported unadjusted. Treatment effects were tested by comparing the difference between the groups at Day 26 (D26) using least square means (with Kenward–Roger degree of freedom for linear models).

In the model building stage for the secondary outcomes, before we arrived at the aforementioned models, linear mixed regression was first evaluated for both Fitbit and sleep-reported TIB, TST, WASO, SL and SE measures since these variables are non-negative and continuous, and Poisson regression for NWAK measures, since these outcomes are count variables. The models for TIB, TST and NWAK fit the data rather well, which to some degree could be expected from their Gaussian- or Poisson-like distributions. However, linear mixed regression did not yield a good fit for WASO, SL and SE measures. We thus proceeded to evaluate whether (a) Tweedie regression would be a more appropriate choice for WASO and SL, since their support is $[0;\infty)$, and (b) beta mixed regression with Smithson and Verkuilen's outcome correction[58] would be a more appropriate choice for SE, since their support is $[0,1]$. For this evaluation, only Tweedie regression fits Fitbit-measured SL data well, and we thus used linear regression as the finalised choice for SE measures, WASO measures and self-reported SL for simplicity. Model fits were assessed with residual QQ-plots and plots of residuals versus predicted values, where DHARMa residuals[59] were used for non-Gaussian models.

For daily diary data, to rectify diary entry errors for SL and WASO, the highest 1% data points were removed for each of these variables. Analysis sets were obtained separately for each of the following categories of analysis: (a) PSQI; (b) Fitbit sleep outcomes; and (c) self-reported sleep outcomes, where only subjects with complete baseline observations across all measures of the respective category were retained.

Results

Baseline characteristics

Table 1 provides a detailed overview of participant baseline characteristics. The majority of the participants were males (63%), of Chinese ethnicity (92%), and undergraduate students (98%). Prior to the intervention, sleep quality was

Table 1. Demographic and baseline characteristics of participants.

Characteristic	Total (n = 59)	Intervention group (n = 30)	Control group (n = 29)
Age (years) (M, SD)	21.9 (3.20)	21.6 (4.26)	22.4 (1.31)
Gender (n, %)			
Female	21 (36%)	12 (40%)	9 (31%)
Male	38 (64%)	18 (60%)	20 (69%)
Race (n, %)			
Chinese	55 (93%)	28 (93%)	27 (93%)
Indian	4 (7%)	2 (7%)	2 (7%)
Year of study (n, %)			
Undergraduate	58 (98%)	30 (100%)	28 (97%)
Postgraduate	1 (2%)	0 (0%)	1 (3%)
Major (n, %)			
Medicine	24 (41%)	14 (47%)	10 (34%)
Engineering	11 (19%)	6 (20%)	5 (17%)
Social science	9 (15%)	4 (13%)	5 (17%)
Business	7 (12%)	3 (10%)	4 (14%)
Computer science	2 (3%)	0 (0%)	2 (7%)
Science	2 (3%)	0 (0%)	2 (7%)
Art, design, media	2 (3%)	2 (7%)	0 (0%)
Exercise and sports studies	2 (3%)	1 (3%)	1 (4%)
Living condition (n, %)			
Living at home	27 (46%)	17 (57%)	10 (34%)
Living at student accommodation	30 (51%)	12 (40%)	18 (62%)
Other	2 (2%)	1 (3%)	1 (3%)
PSQI (M, SD)	5.68 (2.78)	5.23 (2.19)	6.14 (3.26)

PSQI: Pittsburgh Sleep Quality Index.

poor among university students, with an average PSQI score of 5.68 (SD = 2.78). Notably, 40 participants (68%) reported sleeping less than 7 h a day, falling below the

minimum recommendation for individuals in this age group.[60] On average, participants slept for 6.05 (SD = 1.83) hours a day.

Participant flow

Study recruitment for the participants occurred between 31 January and 7 February 2023. A total of 89 participants were screened for eligibility, of which 12 were excluded as they did not meet the eligibility criteria (>10 on PHQ-9 or currently seeking mental health treatment) and 17 did not register for onboarding. Consequently, 60 participants consented to participate and entered the study procedure. 30 participants were randomly allocated to the intervention group and 30 participants to the control group. However, one participant from the control group did not attend the onboarding session; hence 59 participants received their allocated intervention. All participants completed the baseline questionnaire, and 29 of the 30 participants (97%) in the intervention group and 29 participants (100%) in the control group completed the post-intervention questionnaire. The full recruitment flow is shown in Figure 2.

Intervention efficacy

Table 2 presents baseline and post-intervention values for PSQI, Fitbit and self-reported sleep measures, as well as analyses conducted on the aforementioned measures. For the primary outcome, PSQI, linear regression revealed no statistical significance between the intervention and control group. The mean for PSQI was 5.32 (SD = 2.18) for the intervention group and 6.22 (SD = 2.65) for control group, and the difference was not statistically significant (adjusted mean difference = −0.51, 95% CI: [−1.55, 0.77], $p = 0.40$). Similar results were observed for the proportional odds model for PSQI (odds ratio of intervention group versus control group = 0.71, 95% CI: [0.28, 1.81], $p = 0.47$).

For TST and TIB measured using Fitbit, we observed significant differences between the intervention and control group, where the adjusted mean difference assessed at mean baseline value was 32.5 [95% CI: 5.9, 59.1, $p = 0.02$] and 32.3 [95% CI: 2.7, 61.9, $p = 0.03$], respectively. Additionally, on average, the longer a person spent in bed and asleep at baseline, the more pronounced effect our intervention would have on their TST (Supplemental Appendix 2) and TIB (Supplemental Appendix 3) at D26 compared to if they were assigned to the control group. Our analyses suggest that the differences between the two groups for TST and TIB resulted from an increasing trend in the intervention group and a decreasing trend in the control group over time – the estimated slope of time trend equals 0.39, 95% CI: −0.55, 1.34 for TST of intervention group; −0.53, 95% CI: −1.43, 0.37 for TST of control group (group B); 0.36, 95% CI: −0.72, 1.43 for TIB of intervention group; −0.342, 95% CI: −1.37, 0.68 for TIB of control group). For the remaining sleep variables (e.g., WASO, SE, SL), no significant differences were observed. Similarly, for sleep variables assessed by sleep diary, no significant differences were observed for variables although TST was approaching statistical significance ($p = 0.06$).

Data collection adherence

Overall, the adherence to daily sleep diary completion and wearing the Fitbits was high throughout the study. Out of 59 participants, 50 (85%) consistently filled the sleep diary every day during the study duration, with an average of 27.5 (SD = 2.43) entries completed per participant. Among the remaining nine participants who did not complete the sleep diary daily, six individuals missed only one diary entry, two individuals missed three entries and one participant had 16 entries unrecorded. This highlights the overall robustness of data collected, as only 28 days (1.69%) of sleep entries were missed of the total 1652 entries expected. For Fitbit, each participant recorded an average of 24.0 (SD = 5.27) nights of sleep. However, only 11 participants (19%) wore the device consistently to sleep every day during the study duration.

Intervention adherence

Adherence to the intervention was high among participants. The average number of coaching sessions attended was 3.83 (SD = 0.53), with 27 of the 30 participants (90%) having attended all four sessions with their assigned health coach. For the remaining participants, one participant (3.33%) attended three sessions, and two participants (6.67%) attended only two sessions. Reasons for not attending the coaching session include participants failing to remember their scheduled session time and being unable to identify a suitable alternative appointment time within the week. The average duration of the coaching sessions across four weeks was 37.0 (SD = 13.3) minutes which was above the instructed 30 min.

Attrition rate

The attrition rate in the study was low. Only one participant (3.45%) from the control condition did not receive the allocated intervention for not attending the mandatory onboarding session. As such, 59 participants received their allocated intervention (30 in the intervention group and 29 in the control group).

Discussion

This pilot study assessed the feasibility and efficacy of a novel human-AI coaching model, which synergises and integrates human health coaching with a QA system,
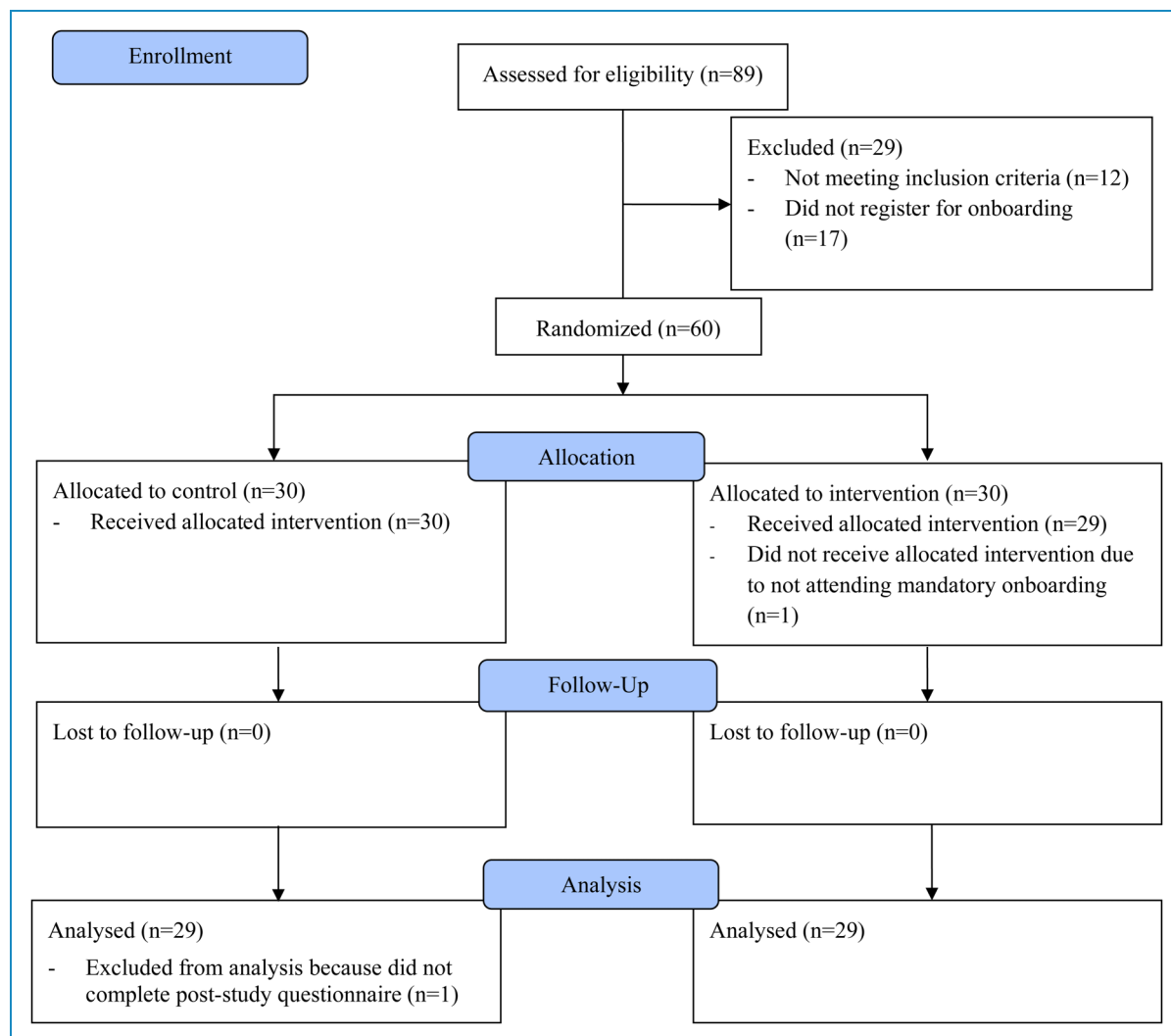
Figure 2. Consolidated Standards of Reporting Trial (CONSORT) flow diagram of recruitment flow of participants.

leveraging both psychological principles and evidence-based data, to improve sleep quality among university students. Our findings illustrate that health coaching, when combined with the QA system, did not achieve improvements in sleep quality as measured by the PSQI (adjusted mean difference $= -0.51$, 95% CI: $[-1.55, 0.77]$, $p = 0.40$). However, improvements were observed in objective measures as measured by Fitbit for TST and TIB, which demonstrated better results in the intervention group than the control group across the study duration ($p < 0.05$). Other Fitbit measures of sleep did not show statistical improvements among university students. Self-reported sleep assessed by the sleep diary also showed no significant difference between the groups. However, it is worth noting that self-reported TST was approaching significance in favour of the intervention group ($p = 0.06$). Notably, adherence to the intervention and study protocol was high in both groups. Participants exhibited a strong commitment to the

intervention suggesting its feasibility. Although the sleep outcomes in PSQI did not align with expected changes, the changes observed in Fitbit measures of sleep, TST and TIB, and the engagement observed between participants and health coaches highlights the potential value of incorporating human support with AI information interventions.

Overall, we observed similar university student characteristics of sleep quality through the PSQI as well as Fitbit and self-reported sleep measures. First, we observed an overall poor sleep quality among university students, indicated by an average baseline PSQI score of 5.68 (SD $= 2.78$). These findings align with previous literature consistently demonstrating that university students' sleep quality tends to be suboptimal.[2,61] Furthermore, the Fitbit and self-reported measures of sleep included in this study are consistent with existing sleep interventions among university students.[52,61] The average TST reported in this study

Table 2. Pre- and post-treatment analyses of all sleep measures between groups.

	Pretreatment		Posttreatment		Comparison at D26 between intervention and control group[a]		
	Intervention group	Control group	Intervention group	Control group	Estimate	95% CI	p
	M (SD)	M (SD)	M (SD)	M (SD)			
	N = 28	N = 27	N = 28	N = 27			
PSQI	5.25 (2.27)	6.26 (3.28)	5.32 (2.18)	6.22 (2.65)	−0.51	−1.55, 0.77	0.40
Fitbit	N = 27	N = 28	N = 24	N = 28			
SE	86.5 (2.9)	87.4 (4.8)	87.3 (3.2)	86.9 (3.0)	1.22	−0.29, 2.72	0.11
TST	381 (93.)	368 (86)	403 (89)	366 (110)	32.5	5.9, 59.1	0.02
TIB	440 (102)	422 (95)	462 (105)	422 (126)	32.3	2.7, 61.9	0.03
SL	4.67 (5.14)	4.55 (4.10)	5.25 (5.08)	6.20 (8.63)	1.14	0.85, 1.52	0.71
NWAK	1.89 (1.42)	1.79 (1.23)	2.08 (1.47)	1.61 (1.57)	1.10	0.90, 1.35	0.53
WASO	44.7 (20.5)	41.1 (21.1)	46.7 (24.5)	42.4 (22.2)	4.79	−1.25, 10.8	0.12
Sleep diary	N = 30	N = 28	N = 28	N = 25			
SE	85.2 (11.6)	86.3 (12.6)	88.6 (8.41)	90.0 (10.5)	0.41	−3.18, 4.01	0.82
TST	385 (90)	391 (76)	441 (106)	374 (141)	27.5	−0.52, 55.5	0.06
SL	17.8 (32.0)	16.6 (24.0)	7.9 (7.3)	9.2 (18.8)	−0.60	−6.38, 5.1	0.84
NWAK	1.87 (2.47)	1.50 (2.89)	1.68 (1.72)	0.36 (0.76)	0.40	−0.14, 0.93	0.15
WASO	16.1 (27.7)	11.1 (17.9)	22.2 (32.8)	4.8 (13.0)	4.24	−2.4, 10.9	0.21

PSQI: Pittsburgh Sleep Quality Index; SE: sleep efficiency; TST: total sleep time; TIB: time in bed; SL: sleep latency; NWAK: number of awakenings during the night; WASO: wake after sleep onset.
[a]Estimate, calculated based on the fitted mixed regression models, refers to: (a) the expected ratio of the intervention group compared to the control group for objective SL and NWAK; (b) the expected mean difference in log scale for subjective NWAK; and (c) to the expected mean difference for the remaining variables.

is also similar to that reported in other studies utilising Fitbit devices for university students.[62,63] TIB, on the other hand, was slightly lower in our study compared to a similar study.[51] Nonetheless, our findings are inconsistent with our initial hypothesis and diverge from existing literature in this field as we did not observe improvements in participant sleep quality after the assigned intervention.

Several factors may explain the lack of improvements in sleep outcomes observed in this study. Firstly, it is important to consider the duration of the health coaching intervention. Previous studies implementing sleep interventions for university students have used intervention periods ranging from 1 week to 10 weeks,[64] while health coaching interventions in the literature range from 3 weeks to 18 months[65]

with varying significance. The intervention duration within this study was four weeks and within the range of previously shown to be effective interventions, albeit on the shorter end of the spectrum. However, behavioural change is a complex process, and changes in sleep may not have immediate, detectable outcomes.[66,67] Therefore, four weeks may have been insufficient to observe significant improvements in sleep quality, and a more long-term intervention period may be necessary.

Another crucial factor to consider is the expertise of the health coaches involved. While health coaching provided by peers has shown promise in improving lifestyle behaviours among university students,[40,68] it is important to acknowledge that they may have relatively less training

and coaching experience compared to professional health coaches.[40,68] The level of training and experience undertaken by these health coaches within this study may have impacted their ability to effectively address sleep-related concerns. For these coaches, this was their first experience applying their coaching skills with a client after attending and passing the 8-week health coaching course, which consisted of online lectures, problem-based learning and role-play. Alongside, it was the first time for the health coaches to use a QA system, and subsequently they had to also learn how to integrate the system into their health coaching practices. To uphold the quality of the health coaching delivered, we arranged weekly health coaching supervision sessions facilitated by professional health coaches to provide their support and expertise.

Nevertheless, it is commonly reported that both QA systems[69,70] and health coaching[30,68] are often effective in improving health outcomes, including sleep. Question-answering systems have been shown to provide accessible treatment support and management for individuals,[69,70] whereas health coaching provided by peers has demonstrated effectiveness in promoting behaviour change.[30,68] Moreover, previous research has emphasised the benefits of incorporating a humanistic aspect into digital interventions for increased intervention adherence and effectiveness.[71] Consequently, we proposed a human-AI symbiosis model that merges the QA system for quick retrieval of evidence-based information with health coaching to foster behavioural change in individuals' sleep quality. Despite efforts to promote human-AI symbiosis within the sleep human-AI coaching model, we did not observe the anticipated improvements in sleep outcomes.

However, it is imperative to note that the limitations observed in our study may be attributed to the nature of the QA system utilised – an information retrieval-based system constrained by the text corpus it was trained on. Additionally, within the current landscape, field of NLP lacks models that are inherently interpretable and explainable.[72,73] Introducing models with interpretability and transparency in AI decision-making processes hold promise for addressing the limitations observed in this study and bridge the gap between questions extracted by our QA system and changes in sleep outcomes observed in students.

Limitations

There are several limitations of the study. As a pilot RCT, a power calculation was not conducted. Hence, the sample size may have been underpowered to detect any significant effect. As a result, caution should be exercised when interpreting these findings. Moreover, the study participants were predominantly of Chinese ethnicity and undergraduate students. This limits the generalisability of our findings across individuals of various ethnicities or across diverse

university student populations. Further research should aim to recruit more diverse groups to enhance the validity of the findings. In this present study, Fitbit devices were employed to evaluate objective sleep measures; however, it is important to note that these measurements are influenced by Fitbit's proprietary machine-learned algorithms which may change without user knowledge over time. Consequently, the sleep metrics obtained from Fitbit might deviate from the true values derived from polysomnography, the gold standard for sleep testing, affecting its validity. In this study, we opted not to use polysomnography, as our aim was to assess student sleep under real-life, free-living conditions. Lastly, while wearable devices like Fitbit offer valuable sleep metrics, their reliability should be approached with caution as many studies have attempted to validate Fitbit's sleep measures by comparing them with polysomnography.[74–77] Nonetheless, in the case of the Fitbit Inspire 2, the model used in this study, recent comparisons with polysomnography revealed no significant differences in SE, TIB and WASO, although it significantly overestimated TST.[74]

Conclusions

With the increasing global interest in medical QA, and its potential applications to promote and improve numerous health outcomes, this study explored the implementation of a human-AI symbiosis model to intervene on sleep quality. This study has shown preliminary feasibility for the model, specifically for sleep health coaching in university settings, as evidenced by a high adherence rate to the intervention. However, despite this, our proposed human-AI sleep coaching model did not demonstrate significant changes in PSQI or self-reported sleep. Nonetheless, TST and TIB as assessed by Fitbit demonstrated significant changes in favour of the intervention group. Consequently, future research should consider extending the intervention duration of the proposed model as promising preliminary improvements were observed in Fitbit sleep within the current timeframe. Moreover, further research will focus on extending the effective integration of QA systems as an information and knowledge tool into health coaching practices for sleep and other conditions, while continuing to enhance the reliability and accuracy of the QA system. The potential for QA systems to be a valuable tool that equips health coaches when providing guidance to their clients is magnified with recent advancements and abilities of LLM QA systems such as ChatGPT and MedGPT. With increased accessibility to quick and accurate information, the enhanced efficiency extends the potential for human-AI symbiosis and harnesses the benefits for QA systems to enhance effective health coaching for behavioural change. This synergy of health coaching and QA systems continues to hold

revolutionary promise in ultimately transforming healthcare delivery among university students.

ORCID iDs: Jintana Liu https://orcid.org/0009-0000-5907-7334
Sakura Ito https://orcid.org/0000-0002-3671-3761
Qi Chwen Ong https://orcid.org/0009-0009-5621-5071
Chuen Peng Lee https://orcid.org/0000-0002-9229-6448
Michael Chia https://orcid.org/0000-0002-1289-0552

Supplemental material: Supplemental material for this article is available online.

References

1. Altun I, Cınar N and Dede C. The contributing factors to poor sleep experiences in according to the university students: a cross-sectional study. *J Res Med Sci* 2012; 17: 557–561.
2. Lund HG, Reider BD, Whiting AB, et al. Sleep patterns and predictors of disturbed sleep in a large population of college students. *J Adolesc Health* 2010; 46: 124–132.
3. Brougham R, Zail C, Mendoza C, et al. Stress, sex differences, and coping strategies among college students. *Curr Psychol* 2009; 28: 85–97.
4. O'Malley PM and Johnston LD. Epidemiology of alcohol and other drug use among American college students. *J Stud Alcohol Suppl* 2002; 14: 23–39.
5. Arria AM. Compromised sleep quality and low GPA among college students who use prescription stimulants nonmedically. *Sleep Med* 2011; 12: 536–537.
6. Sexton-Radek K and Pichler-Mowry R. Daily activities and sleep quality in young adults. *Percept Mot Skills* 2011; 112: 426–428.
7. Gaultney JF. The prevalence of sleep disorders in college students: impact on academic performance. *J Am Coll Health* 2010; 59: 91–97.
8. Oginska H and Pokorski J. Fatigue and mood correlates of sleep length in three age-social groups: school children, students, and employees. *Chronobiol Int* 2006; 23: 1317–1328.
9. Hirshkowitz M, Whiton K, Albert SM, et al. National sleep foundation's sleep time duration recommendations: methodology and results summary. *Sleep Health* 2015; 1: 40–43.
10. Cellini N, Menghini L, Mercurio M, et al. Sleep quality and quantity in Italian university students: an actigraphic study. *Chronobiol Int* 2020; 37: 1538–1551.
11. Harvey AG. Sleep and circadian functioning: critical mechanisms in the mood disorders? *Annu Rev Clin Psychol* 2011; 7: 297–319.
12. Kalmbach DA, Abelson JL, Arnedt JT, et al. Insomnia symptoms and short sleep predict anxiety and worry in response to stress exposure: a prospective cohort study of medical interns. *Sleep Med* 2019; 55: 40–47.
13. Griffin SC, Williams AB, Ravyts SG, et al. Loneliness and sleep: a systematic review and meta-analysis. *Health Psychol Open* 2020; 7: 2055102920913235.
14. Covassin N and Singh P. Sleep duration and cardiovascular disease risk: epidemiologic and experimental evidence. *Sleep Med Clin* 2016; 11: 81–89.
15. Calhoun DA and Harding SM. Sleep and hypertension. *Chest* 2010; 138: 434–443.
16. Koo DL, Nam H, Thomas RJ, et al. Sleep disturbances as a risk factor for stroke. *J Stroke* 2018; 20: 12–32.
17. Beccuti G and Pannain S. Sleep and obesity. *Curr Opin Clin Nutr Metab Care* 2011; 14: 402–412.
18. Armand MA, Biassoni F and Corrias A. Sleep, well-being and academic performance: a study in a Singapore residential college. *Front Psychol* 2021; 12: 672238.
19. Ho GW, Yang Z, Xing L, et al. Nighttime sleep awakening frequency and its consistency predict future academic performance in college students. *Int J Environ Res Public Health* 2022; 19: 2933.
20. Joshi SC. Sleep latency and sleep disturbances mediates the association between nighttime cell phone use and psychological well-being in college students. *Sleep Biol Rhythms* 2022; 20: 431–443.
21. Dietrich SK, Francis-Jimenez CM, Knibbs MD, et al. Effectiveness of sleep education programs to improve sleep hygiene and/or sleep quality in college students: a systematic review. *JBI Database System Rev Implement Rep* 2016; 14: 108–134.
22. Al-Kandari S, Alsalem A, Al-Mutairi S, et al. Association between sleep hygiene awareness and practice with sleep quality among Kuwait university students. *Sleep Health* 2017; 3: 342–347.
23. Brown FC, Buboltz WC Jr and Soper B. Relationship of sleep hygiene awareness, sleep hygiene practices, and sleep quality in university students. *Behav Med* 2002; 28: 33–38.

24. Budler LC, Gosak L and Stiglic G. Review of artificial intelligence-based question-answering systems in healthcare. *WIRES Data Mining Knowl Discov* 2023; 13: e1487.

25. Wu H, Wang M, Wu J, et al. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *NPJ Digit Med* 2022; 5: 186.

26. Goodwin TR and Harabagiu SM. Medical question answering for clinical decision support. *Proc ACM Int Conf Inf Knowl Manag* 2016; 2016: 297–306.

27. Mihalache A, Popovic MM and Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA Ophthalmol* 2023; 141: 589–597.

28. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023; 620: 172–180.

29. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023; 183: 589–596.

30. Oh YJ, Zhang J, Ji X, et al. Efficacy of a chatbot-based sleep intervention on sleep quality improvement among young adults. *Sleep* 2022; 45: A42-A.

31. World Health Organization. Adherence to long-term therapies: Evidence for action, https://apps.who.int/iris/bitstream/handle/10665/42682/9241545992.pdf?sequence=1&isAllowed=y (2003).

32. Interactional Coaching Federation. All things coaching, https://coachingfederation.org/about

33. National Board of Health and Wellness. Code of ethics.

34. Werbrouck A, Swinnen E, Kerckhofs E, et al. How to empower patients? A systematic review and meta-analysis. *Transl Behav Med* 2018; 8: 660–674.

35. Aujoulat I, d'Hoore W and Deccache A. Patient empowerment in theory and practice: polysemy or cacophony? *Patient Educ Couns* 2007; 66: 13–20.

36. Lin CL, Huang LC, Chang YT, et al. Effectiveness of health coaching in diabetes control and lifestyle improvement: a randomized-controlled trial. *Nutrients* 2021; 13: 3878.

37. Aboalshamat K, Al-Zaidi D, Jawa D, et al. The effect of life coaching on psychological distress among dental students: interventional study. *BMC Psychol* 2020; 8: 106.

38. Suminski RR, Leonard T, Obrusnikova I, et al. The impact of health coaching on weight and physical activity in obese adults: a randomized control trial. *Am J Lifestyle Med* 2022; 8: 233–242.

39. Topiwala A, Braniste A, Patki S, et al. How a personalised health coaching intervention can improve sleep in junior doctors working night shifts. *Eur Respiratory Soc* 2023; 9: 120.

40. DeShaw KJ, Lansing JE, Perez ML, et al. Effects of a peer health coaching program on college student lifestyle behaviors. *J Am Coll Health* 2023: 1–8.

41. Duren-Winfield V, Kimnya, Onsomu E, et al. Champions for outreach and advocacy for campus and community health: a college-based peer health coach program. *J Commun Engage Higher Educ* 2011; 3: 1–11.

42. Quintiliani LM and Whiteley JA. Results of a nutrition and physical activity peer counseling intervention among nontraditional college students. *J Cancer Educ* 2016; 31: 366–374.

43. Fadhil A, Wang Y and Reiterer H. Assistive conversational agent for health coaching: a validation study. *Methods Inf Med* 2019; 58: 9–23.

44. Fitzpatrick KK, Darcy A and Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Ment Health* 2017; 4: e19.

45. Bojic I, Ong QC, Thakkar M, et al.. SleepQA: a health coaching dataset on sleep for extractive question answering. *Proceedings of the 2nd machine learning for health symposium: proceedings of machine learning research* 2022: 199–217.

46. Bojic I, Halim J, Suharman V, et al. A Data-centric Framework for Improving Domain-specific Machine Reading Comprehension Datasets. *arXiv preprint arXiv:230400483* 2023.

47. Jurafsky D. *Speech & language processing*. Pearson Education India, 2000.

48. Browne RH. On the use of a pilot sample for sample size determination. *Stat Med* 1995; 14: 1933–1940.

49. Buysse DJ, Reynolds CF 3rd, Monk TH, et al. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry Res* 1989; 28: 193–213.

50. Buysse DJ, Germain A, Moul DE, et al. Efficacy of brief behavioral treatment for chronic insomnia in older adults. *Arch Intern Med* 2011; 171: 887–895.

51. Schlarb AA, Friedrich A and Claßen M. Sleep problems in university students – an intervention. *Neuropsychiatr Dis Treat* 2017; 13: 1989–2001.

52. Barber LK and Cucalon MS. Modifying the sleep treatment education program for students to include technology use (STEPS-TECH): intervention effects on objective and subjective sleep outcomes. *Stress Health* 2017; 33: 684–690.

53. Sleep Foundation. Sleep diary. https://www.sleepfoundation.org/wp-content/uploads/2021/02/SF-23-127_Sleep_Diary_Interactive.pdf.

54. Carney CE, Buysse DJ, Ancoli-Israel S, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. *Sleep* 2012; 35: 287–302.

55. Klier K and Wagner M. Agreement of sleep measures – A comparison between a sleep diary and three consumer wearable devices. *Sensors* 2022; 22: 6189.

56. Teresi JA, Yu X, Stewart AL, et al. Guidelines for designing and evaluating feasibility pilot studies. *Med Care* 2022; 60: 95–103.

57. Orsmond GI and Cohn ES. The distinctive features of a feasibility study: objectives and guiding questions. *OTJR (Thorofare N J)* 2015; 35: 169–177.

58. Smithson M and Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol Methods* 2006; 11: 54–71.

59. Hartig F. DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. *R package version 03* 2020; 3.

60. Centers for Disease Control and Prevention. Do you get enough sleep? https://www.cdc.gov/chronicdisease/resources/infographic/sleep.htm (2021).

61. Taylor DJ, Zimmerman MR, Gardner CE, et al. A pilot randomized controlled trial of the effects of cognitive-behavioral

therapy for insomnia on sleep and daytime functioning in college students. *Behav Ther* 2014; 45: 376–389.

62. Wang C, Lizardo O and Hachen DS. A longitudinal study of Fitbit usage behavior among college students. *Cyberpsychol Behav Soc Netw* 2022; 25: 181–188.

63. Garcia L, Ferguson SE, Facio L, et al. Assessment of well-being using Fitbit technology in college students, faculty and staff completing breathing meditation during COVID-19: a pilot study. *Ment Health Prev* 2023; 30: 200280.

64. Saruhanjan K, Zarski AC, Bauer T, et al. Psychological interventions to improve sleep in college students: a meta-analysis of randomized controlled trials. *J Sleep Res* 2021; 30: e13097.

65. Kivelä K, Elo S, Kyngäs H, et al. The effects of health coaching on adult patients with chronic diseases: a systematic review. *Patient Educ Couns* 2014; 97: 147–157.

66. Bjørnnes AK, Torbjørnsen A, Valeberg BT, et al. What is known about students and sleep: systematic review and evidence map. *SAGE Open* 2021; 11: 21582440211032162.

67. Hale L, Troxel W and Buysse DJ. Sleep health: an opportunity for public health to address health equity. *Annu Rev Public Health* 2020; 41: 81–99.

68. Yan Z, Peacock J, Cohen JF, et al. An 8-week peer health coaching intervention among college students: a pilot randomized study. *Nutrients* 2023; 15: 1284.

69. Milne-Ives M, de Cock C, Lim E, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *J Med Internet Res* 2020; 22: e20346.

70. Dhinagaran DA, Sathish T, Soong A, et al. Conversational agent for healthy lifestyle behavior change: web-based feasibility study. *JMIR Form Res* 2021; 5: e27956.

71. Baumeister H, Reichler L, Munzinger M, et al. The impact of guidance on internet-based mental health interventions – A systematic review. *Internet Interv* 2014; 1: 205–215.

72. Gurrapu S, Kulkarni A, Huang L, et al. Rationalization for explainable NLP: a survey. *Front Artif Intell* 2023; 6: 1225093.

73. Minh D, Wang HX, Li YF, et al. Explainable artificial intelligence: a comprehensive review. *Artif Intell Rev* 2022; 55: 3503–3568.

74. Lim SE, Kim HS, Lee SW, et al. Validation of Fitbit Inspire 2(TM) against polysomnography in adults considering adaptation for use. *Nat Sci Sleep* 2023; 15: 59–67.

75. de Zambotti M, Goldstone A, Claudatos S, et al. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiol Int* 2018; 35: 465–476.

76. Hakim M, Miller R, Hakim M, et al. Comparison of the Fitbit® charge and polysomnography for measuring sleep quality in children with sleep disordered breathing. *Minerva Pediatr (Torino)* 2022; 74: 259–263.

77. Miller DJ, Sargent C and Roach GD. A validation of six wearable devices for estimating sleep, heart rate and heart rate variability in healthy adults. *Sensors (Basel)* 2022; 22: 6317.